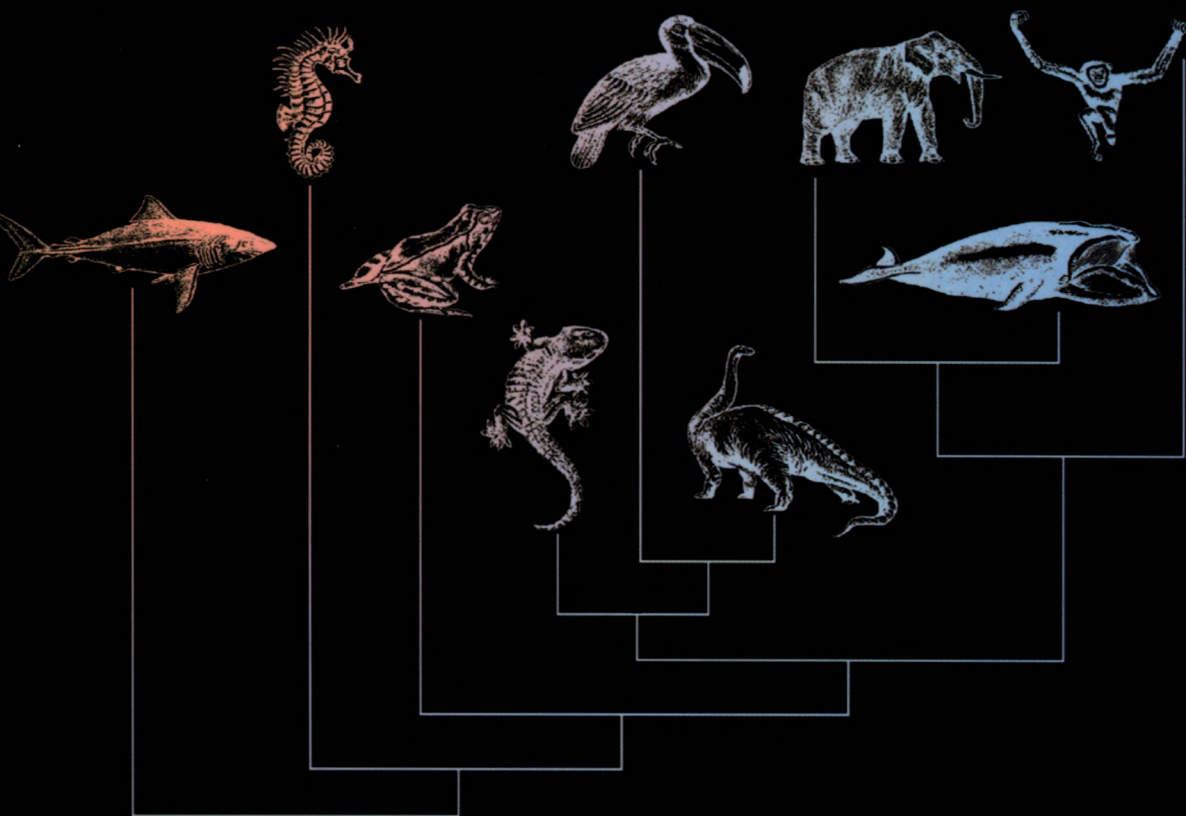


Andrzej
Falniowski

Metody numeryczne w taksonomii



Metody
numeryczne
w taksonomii

Andrzej
Falniowski

Metody
numeryczne
w taksonomii



Wydawnictwo Uniwersytetu Jagiellońskiego

© Copyright by Andrzej Falniowski & Uniwersytet Jagielloński

Wydanie I, Kraków 2003

All rights reserved

RECENZENCI

Prof. dr hab. Zbigniew Dąbrowski

Prof. dr hab. Krzysztof Kaczanowski

Dr hab. Andrzej Kozik

PROJEKT OKŁADKI

Jolanta Olszowska

REDAKTOR

Jerzy Hrycyk

KOREKTOR

Szczepan Catek

Publikacja dofinansowana ze środków centralnej rezerwy na badania własne Uniwersytetu Jagiellońskiego oraz Instytut Zoologii Uniwersytetu Jagiellońskiego

ISBN 83-233-1745-3

www.wuj.pl

Wydawnictwo Uniwersytetu Jagiellońskiego

Dystrybucja: ul. Bydgoska 19 C, 30-056 Kraków

tel. (012) 638-77-83, 636-80-00 w. 2022, fax (012) 423-31-60, 636-80-00 w. 2023

tel. kom. 0506-006-674, e-mail: wydaw@if.uj.edu.pl

Konto: BPH PRK SA IV/O Kraków nr 10601389-320000478769

Spis treści

Wstęp	7
1. Drzewa	9
1.1. Elementy i symbolika drzewa, politomie; sieć.....	9
1.2. Kladogramy, drzewa addytywne i ultrametryczne.....	14
1.3. Drzewa ukorzenione i nieukorzenione, liczba możliwych topologii.....	15
1.4. Filogeneza genów i organizmów, wewnątrz i powyżej gatunku.....	18
1.5. Rekonstrukcja ewolucji cech, grupowanie kladystyczne i fenetyczne.....	20
2. Dane	25
2.1. Cechy „kladystyczne” i „fenetyczne”.....	25
2.2. Homologie i ich testowanie.....	27
2.3. Cechy jakościowe: polaryzacja, serie transformacyjne, wagi.....	29
2.4. Kodowanie cech ilościowych ciągłych jako dyskretnych.....	32
2.5. Kodowanie cech jakościowych i ilościowych dyskretnych, brakujące dane, polimorfizm.....	34
2.6. Wstępna analiza cech ilościowych; ujęcie opisowe i stochastyczne.....	37
2.7. Transformacja i standaryzacja danych ilościowych.....	40
2.8. Rozkłady nieznanne, techniki Monte Carlo, liczby losowe, <i>jackknife</i> , <i>bootstrap</i> ; test Mantela.....	42
2.9. Odległości dla cech ilościowych ciągłych.....	46
2.10. Odległości dla cech jakościowych wielostanowych i binarnych.....	50
2.11. Częstości elektromorf i odległości genetyczne.....	52
2.12. Odległości dla sekwencji kwasów nukleinowych i białek.....	58
3. Analiza fenetyczna	69
3.1. Statystyczna analiza wielowymiarowa – wprowadzenie, macierze, rozkłady, jednorodność.....	69
3.2. Analiza głównych składowych.....	76
3.3. Analiza głównych współrzędnych.....	84
3.4. Nieliniowe skalowanie wielowymiarowe.....	85
3.5. Analiza odpowiadania.....	88
3.6. Najkrótsze drzewo połączeń.....	92
3.7. Analiza skupisk, odległości kofenetyczne.....	94
3.8. Wielowymiarowa analiza wariancji.....	101
3.9. Analiza dyskryminacyjna.....	103

4. Analiza filogenetyczna	111
4.1. Zakres metod filogenetycznych, algorytm a model, kryteria optymalizacji	111
4.2. Metody algorytmiczne oparte na odległościach	114
4.3. Problem znalezienia najlepszego drzewa	118
4.4. Metody stosujące kryterium optymalizacji oparte na odległościach	126
4.5. Metoda kladystyczna (redukjonistyczna)	132
4.6. Metody oparte na maksymalizacji wiarygodności	156
4.7. Analiza spektralna	167
4.8. Specjalne odmiany metody kladystycznej	175
4.9. Próbkowanie numeryczne i drzewa losowe	183
4.10. Wnioskowanie na podstawie więcej niż jednego drzewa i zestawu danych	185
4.11. Błędy losowe i systematyczne, wiarygodność znalezionych drzew	197
5. Programy komputerowe	215
6. Wybrana literatura	217

Wstęp

Metody numeryczne znajdują obecnie szerokie zastosowanie w taksonomii biologicznej. W wielu przypadkach są pomocne, w innych – jak przede wszystkim dla cech molekularnych – bez ich wykorzystania jakkolwiek analiza pokrewieństw czy choćby podobieństw zupełnie nie byłaby możliwa. Użyteczność tych technik wykracza poza samą taksonomię, przykładem rekonstrukcja historii mutacji wirusa HIV, pozwalające na śledzenie jego ewolucji, a więc dostarczające ważnych wskazówek dotyczących biologii i sposobów zwalczania tego wirusa. Wydaje się, że współczesny taksonom czy biolog ewolucyjny powinien znać przynajmniej podstawy tych technik. Taki właśnie zakres podstawowych wiadomości starałem się przedstawić w tej książce. Omawiane są tu metody analizy danych zarówno morfologicznych, jak i molekularnych, choć punkt ciężkości przesunięty jest na te ostatnie. Jest tak dlatego, że właśnie analiza danych molekularnych rozwija się ostatnio najszybciej, a szereg technik tylko tu znajduje zastosowanie.

Na wstępie czuję się zobowiązany do przedstawienia kilku uwag. Zakres tej książki obejmuje – właśnie i wyłącznie – metody numeryczne: zaczyna od danych, kończy na drzewach, obliczonych na podstawie danych wyjściowych. Choć uważny Czytelnik zapewne dostrzeże, że autor skłania się do zupełnie nieortodoksyjnego kładyzmu, książka nie zajmuje się teorią klasyfikacji, odsyłając do stosownej literatury. Powinna być użyteczna zarówno dla taksonomów filogenetycznych, jak i ewolucyjnych czy fenetyków. Podobnie w książce nie zajmuję się interpretacją uzyskanych drzew. Uznałem też za celowe włączenie podstawowych technik fenetycznych, bowiem często nie ma możliwości użycia – w sposób uprawniony – technik taksonomii filogenetycznej, a ponadto techniki fenetyczne doskonale nadają się do wstępnej analizy danych – czyli właśnie badania różnorodności.

Dla wielu Czytelników kontrowersyjne będzie traktowanie strony matematycznej omawianych metod. Pragnę jednak podkreślić, że takie podejście głęboko przemyślałem. Metod numerycznych uczę studentów Uniwersytetu Jagiellońskiego od ponad dziesięciu lat. Czuję się więc uprawniony do stwierdzenia, że studenci biologii matematykę znają bardzo słabo, nie lubią jej, a nawet wręcz się jej boją. To nastawienie pozostaje też często i później. Są oczywiście wyjątki – zgodnie z moimi doświadczeniami nieliczne – ci jednak poradzą sobie i bez tej książki, korzystając z bogatej specjalistycznej literatury. Pozostali natomiast albo starają się całkowicie unikać zastosowania technik numerycznych, albo też stosują je zupełnie bez zrozumienia. Pożądana więc dla nich będzie umiejętność prawidłowego wyboru metody, poprawnego jej uży-

cia i właściwej interpretacji wyników, nawet jeżeli aparat matematyczny pozostanie dla nich nie całkiem jasny. Niewątpliwie to lepsze rozwiązanie niż mechaniczne i nieodpowiednie użycie któregoś z licznych dostępnych programów komputerowych, czego przykładów w literaturze znajdujemy dostatek. Stąd też stronę matematyczną starałem się ograniczyć do minimum, kładąc nacisk na intuicyjne zrozumienie przedstawianych technik. Warto także pamiętać, że w wielu przypadkach różni badacze różnie oceniają wady i zalety kolejnych technik, nieraz w sposób skrajny. Choć przedstawiając kolejne metody, starałem się, na ile to możliwe, zachować bezstronność i obiektywizm, to muszę zaznaczyć, że pozostawiłem sobie swobodę wyboru i interpretacji i z tej swobody szeroko korzystałem.

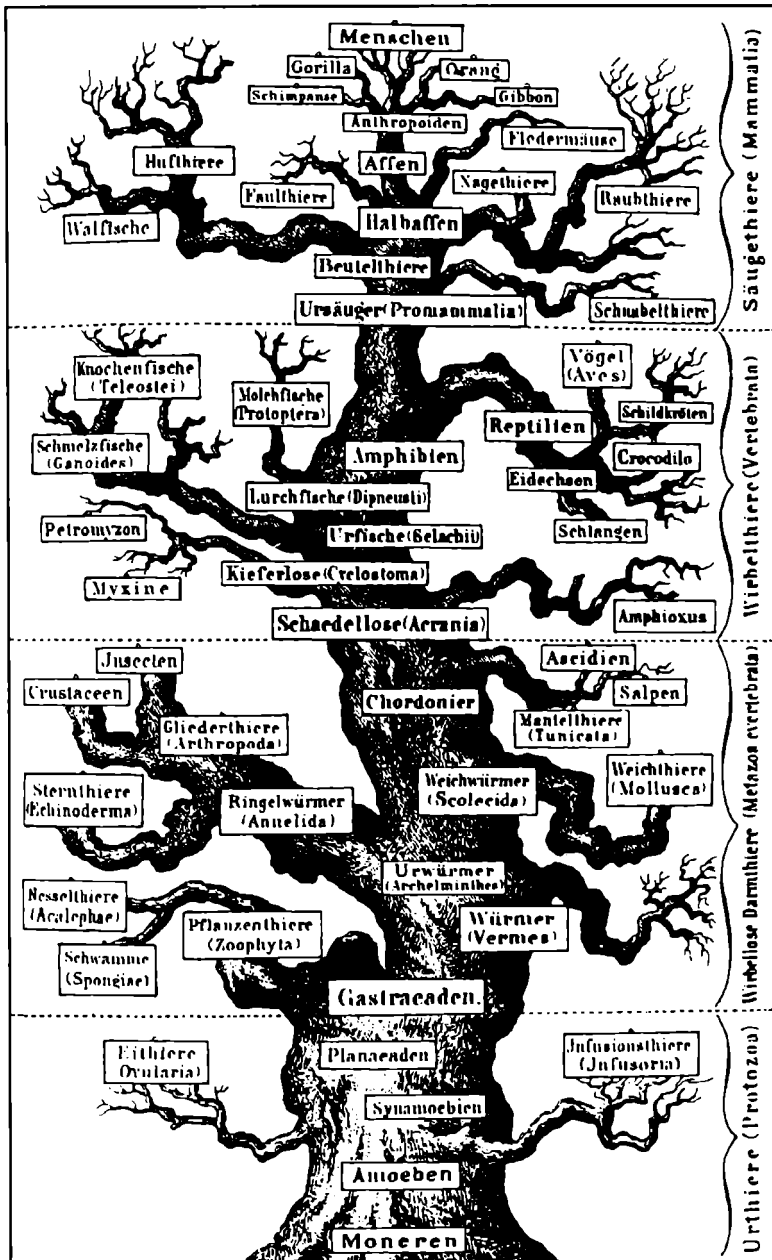
Metody numeryczne stosowane w taksonomii, zwłaszcza dla analizy sekwencji kwasów nukleinowych, rozwijają się niezwykle szybko. Uniemożliwia to oczywiście napisanie podręcznika kompletnego, omawiającego wszystkie, także najnowsze techniki. Tym bardziej nie jest to możliwe w książce, która z założenia nie może być zbyt obszerna. W dodatku spotyka się opinie, że szybki postęp w tej dziedzinie powoduje błyskawiczne starzenie się takich opracowań – niektórzy twierdzą wręcz, że podręczniki omawiające te metody można, podobnie jak opisujące procesory czy programy komputerowe, po roku lub dwóch wyrzucić, jako zupełnie już nieaktualne. Nie podzielam tej opinii. Oczywiście istnieje szereg nowych, właśnie wprowadzanych technik, często o zaledwie zarysowujących się własnościach, a inne metody przestają być stosowane. Uważam jednak, że obok nich istnieje podstawowy „rdzeń” modeli, założeń i metod, które jeszcze długo się nie zmienią i których zrozumienie jest warunkiem zarówno prawidłowego korzystania z zakresu tego „rdzenia”, jak i rozumienia najnowszych, wciąż zmieniających się technik. Ten właśnie rdzeń staram się przedstawić w tej książce.

1. Drzewa

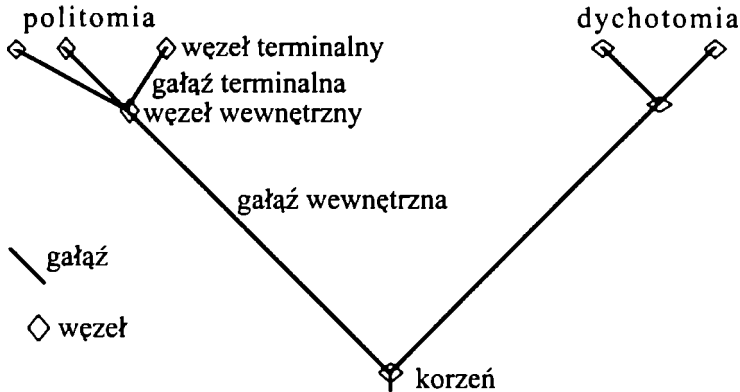
1.1. Elementy i symbolika drzewa, politomie; sieć

Jeżeli założymy – zgodnie ze współczesną wiedzą biologiczną – że wszystkie organizmy żywe powstały na drodze stopniowej ewolucji od jednego, wspólnego przodka, to przebieg tej ewolucji – genealogiczną historię życia – przedstawić można za pomocą drzewa, a jej fragmenty za pomocą odpowiednich, fragmentarycznych drzew. Drzewa genealogiczne, przedstawiające pokrewieństwa w obrębie rodzin, najczęściej rodów panujących, znane są od stuleci; podobne drzewa wykorzystywano do zilustrowania pokrewieństw w obrębie świata żywego już w XIX stuleciu, szeroko znane są np. drzewa Ernesta Haeckla (Ryc. 1.1). Pierwotnie ozdobne i stylizowane rzeczywiście na drzewa, powoli stały się symbolicznym przedstawieniem pokrewieństw, przypominając drzewa jedynie pokrojem, z rozgałęziającymi się – najczęściej dychotomicznie – gałęziami. Prawidłowo zrekonstruowane drzewa są ostatecznym rezultatem większości przedstawionych dalej technik, ale zaczniemy od omówienia samych drzew, bowiem w ten sposób najlepiej wprowadzić szereg pojęć, później użytecznych. Pewne zagadnienia, zarysowane w tej części książki, omówione zostaną szerzej w częściach następnych.

Drzewem (*tree*) nazywamy zestaw połączonych linii łamanych, rozwidlających się, lecz nie tworzących łamanych zamkniętych (Ryc. 1.2), w sposób symboliczny przedstawiający **historię ewolucyjną**, czyli **filogenezę** grupy, rozumianą jako historia genealogicznego następstwa. Prawidłowo zrekonstruowane drzewo przedstawia w sposób kompletny filogenezę i jako takie jest jedyną podstawą tworzenia klasyfikacji dla **taksonomii filogenetycznej**, czyli **kladystyki**, a uważane jest również za pomocne dla taksonomii **ewolucyjnej**, zezwalającej na mniej rygorystyczną interpretację filogenezy w systematyce. Choć włączanie do drzewa znanych kopalnych przodków jest możliwe, to jednak podstawą konstrukcji drzew są **stany** poszczególnych **cech u kolejnych taksonów**, zdarzyć się więc może, że obliczone drzewo przedstawiać będzie nie pokrewieństwa, a jedynie podobieństwa pomiędzy taksonami – wówczas będzie to drzewo **fenetyczne**, grupujące nie na podstawie pokrewieństw, a jedynie **ogólnego podobieństwa** (*overall similarity*), będącego podstawą **taksonomii fenetycznej**. Drzewo składa się z **gałęzi** (*branches*, rzadziej *edges*) łączących **węzły** (*nodes*). Węzły **terminalne** (*terminal nodes*, rzadziej *leaves*), czyli obiekty, dla których mamy dane – stany kolejnych cech – to inaczej **operacyjne jednostki taksonomiczne** (*Operational Taxonomic Units* – OTU).

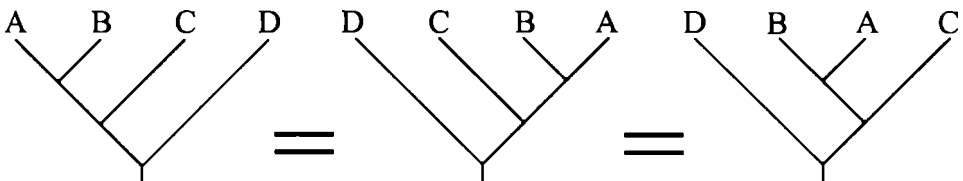


Ryc. 1.1. Drzewo przedstawiające filogenezę człowieka, z E. Haeckla (1874: *Anthropogenie: Keimes- und Stammes-Geschichte des Menschen*). Na „starym drzewie” o porozgałęzionych i powykręcanych konarach znajdujemy czym wyżej tym bardziej zaawansowane, „wyższe” organizmy, zwieńczone człowiekiem jako formą najwyższą, gdy w dolnych częściach samego pnia – stadia stworzone wyłącznie na podstawie danych embriologii porównawczej. Całość odzwierciedlała nie tyle ówczesny stan wiedzy, ile poglądy *Naturphilosophen*, których czołowym przedstawicielem był Haeckel



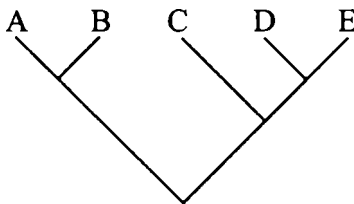
Ryc. 1.2. Podstawowe elementy dychotomicznego i politomicznego drzewa

OTU to określenie bardzo wygodne, bowiem nie muszą to być konieczne gatunki – choć genealogiczne następstwo gatunków w wyniku kolejnych specjacji najlepiej odpowiada takiemu przedstawieniu, a historia specjacji, czyli drzewo gatunków, to dla potrzeb taksonomii przedstawienie najpełniejsze. Możemy jednak za OTU uważać jednostki wyższe, jak rodzaje czy rodziny, bądź jednostki o niezdefiniowanej randze taksonomicznej. W praktyce jako OTU traktuje się też często populacje czy nawet osobniki: jest to wygodne, choć niezupełnie właściwe. Międzygatunkowa hybrydyzacja, choć u zwierząt nie tak rzadka jak niegdyś sądzono (Templeton 1989), a u roślin dość częsta, w sumie jest jednak na ogół zjawiskiem marginalnym i genealogiczne następstwo daje się przedstawić drzewem, niezawierającym **ok**, czyli **pętli (cycle)**, odzwierciedlających łączenie linii (hybrydyzację). Poniżej poziomu gatunku natomiast łączenie linii jest nie do pominięcia, więc drzewo nie będzie w stanie odzwierciedlić pokrewieństw w obrębie populacji. Pomiędzy populacjami na ogół też występuje przepływ genów, ponadto populacje wyróżniane są często arbitralnie, gdy znaczna odległość między badanymi stanowiskami uzasadnia założenie, że przepływ genów między nimi powinien być co najmniej skrajnie ograniczony, choć być tak nie musi. W sumie więc drzewa dla OTU wewnątrz gatunku zawsze będą odzwierciedlać ewolucję w sposób mocno przybliżony, chyba że będą to drzewa obliczone na podstawie mitochondrialnego DNA (mtDNA), dziedziczonego jedynie w linii żeńskiej (patrz Avise 2000).



Ryc. 1.3. Pomimo różnego wyglądu powyższe drzewa są identyczne, czyli mają tę samą topologię i odzwierciedlają tę samą historię procesów kladogenety

Węzły **wewnętrzne** (*internal*) to hipotetyczni przodkowie (*ancestors*) – których zestawy cech możemy odtworzyć dopiero na podstawie zrekonstruowanego drzewa – zwani też niekiedy **hipotetycznymi jednostkami taksonomicznymi** (*Hypothetical Taxonomic Units* – HTU). Przodka wszystkich OTU na drzewie określamy jako **korzeń** (*root*) drzewa. Węzły wewnętrzne symbolizują procesy **kladogenezy**, czyli powstawania nowych gałęzi (**kladów**), podczas gdy wzdłuż gałęzi zachodzi **anageneza**, czyli **ewolucja filetyczna**. Jeżeli w węźle z jednej gałęzi powstają dwie, to mówimy o rozgałęzieniu **dychotomicznym** (Ryc. 1.2), jeżeli więcej niż dwie, to mamy **politomię** (Ryc. 1.2). Za w **pełni zrekonstruowane drzewo** (*fully resolved tree*) uważa się drzewo, na którym brak politomii. Politomie występują powszechnie, tym częściej, im mniej mamy danych – lub im mniej „**sygnału filogenetycznego**” zawierają nasze dane. Wówczas uwzględnienie kolejnych, dodatkowych cech na ogół likwiduje politomie, a przynajmniej ich część. Politomie będące następstwem niepełnej rekonstrukcji drzewa, czyli takie, które teoretycznie można przekształcić w zestawy kolejnych dychotomicznych rozgałęzień, nazywamy **miękkimi** (*soft polytomy*). Zgodnie z klasyczną teorią taksonomii filogenetycznej, sformułowaną przez Henniga (1966), jedynie takie politomie są dopuszczalne. Współcześnie jednak zakłada się, że w jednej linii ewolucyjnej mogło jednocześnie zachodzić więcej niż jeden procesów specjacji, a przynajmniej, że kolejne specjacje zachodziły w mniejszych odstępach czasowych niż niezbędne dla modyfikacji stanów cech, na podstawie których prowadzi się analizę. Bywa tak choćby przy allopatrycznej specjacji grupy izolowanych populacji marginalnych. Wówczas politomia nie wynika z braku danych, a opisuje faktyczną historię ewolucji – jest to politomia **twarda** (*hard polytomy*).



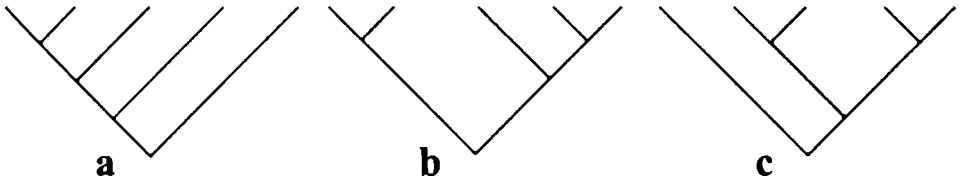
Ryc. 1.4. Topologię tego drzewa można zapisać: ((A,B),((D,E),C))

Topologia drzewa określa kolejność odgałęzień odzwierciedlającą genealogiczne związki, czyli filogenezę. Taką samą zawartość informacyjną – przedstawianie identycznej filogenezy – mogą mieć drzewa, które na pierwszy rzut oka wyglądają niepodobnie (Ryc. 1.3). Należy więc nauczyć się porównywania drzew. Niekiedy zachodzi potrzeba zapisania topologii drzewa, np. dla niektórych programów komputerowych lub w celu skrótowego przedstawienia większej liczby drzew, bez zajmujących wiele stron przedstawień graficznych. Istnieje prosty, powszechnie przyjęty standard, opisujący kolejne węzły za pomocą nawiasów. Dla drzewa przedstawionego na Ryc. 1.4 topologię można zapisać:

((A,B),((D,E),C)).

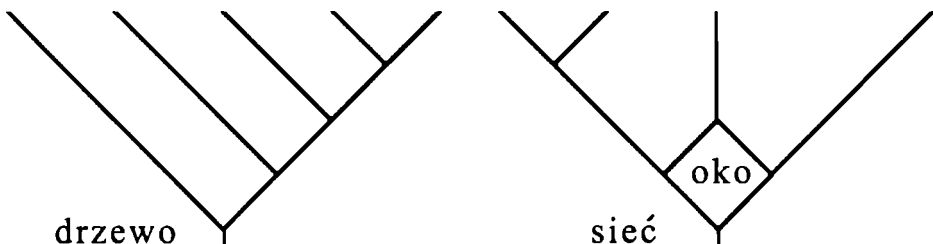
Niekiedy w tym zapisie pomija się przecinki. Pojęcia „topologia” używa się niekiedy w nieco innym, węższym sensie, ograniczając do kształtu drzewa, nie rozróżniając ta-

kiego lub innego układu taksonów terminalnych. Ryc. 1.5 przedstawia wszystkie możliwe topologie w tym węższym rozumieniu dla drzewa o pięciu taksonach terminalnych (spośród wszystkich 105 ukorzenionych możliwych dla tylu taksonów). Pamiętając o drzewach z Ryc. 1.3, pominięto trzy inne kształty, symetryczne do trzech przedstawionych, jako identyczne z nimi. Drzewa z Ryc. 1.5, choć mogą nie różnić się między sobą następstwem taksonów terminalnych, mają nieco inną zawartość informacyjną, przedstawiając inną historię kladogenez.



Ryc. 1.5. Wszystkie możliwe kształty, czyli topologie w wąskim rozumieniu, dla pięciotaksonowego drzewa (spośród łącznie 105 ukorzenionych drzew, różniących się topologią w szerszym rozumieniu, czyli także kolejnością taksonów terminalnych i miejscem ukorzenia, możliwych dla pięciu taksonów terminalnych)

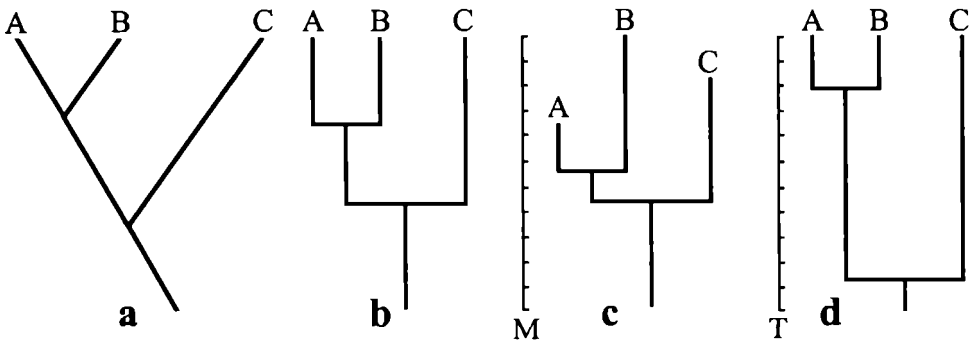
Rekonstruując bądź opisując genealogię na poziomie osobników, w obrębie populacji czy między populacjami, a także wówczas, gdy gatunki powstawały również w następstwie hybrydyzacji, zamiast drzewa wykorzystywać musimy sieć (*network*), w której obok rozwidleń występują także oka, czyli pętle (Ryc. 1.6). O ile wszystkie programy komputerowe operują na drzewach, to z sieciami większość sobie nie radzi, obecność ok bardzo więc utrudnia analizę. Warto też wspomnieć, że wielu taksonomów określa mianem sieci (*network*) drzewa nieukorzenione, co z punktu widzenia teorii grafów jest niewłaściwe, a w terminologii wprowadza niepotrzebne zamieszanie.



Ryc. 1.6. Drzewo odróżnia od sieci brak ok, czyli linii łamanych zamkniętych

1.2. Kladogramy, drzewa addytywne i ultrametryczne

Przedstawione dotąd drzewa (Ryc. 1.2–1.5) miały wszystkie formę **kladogramu** (*cladogram*), kladogramami są też drzewa a i b na Ryc. 1.7. Na takim drzewie taksony terminalne znajdują się u góry, korzeń u dołu, a więc chronologiczne następstwo kladogenez przedstawione jest od dołu do góry. Kolejność kladogenez wyczerpuje zawartość informacyjną drzewa, natomiast długości gałęzi, czyli odstęp między węzłami, niczego nie odzwierciedlają, rysowane są po prostu tak, aby terminalne OTU leżały na tym samym poziomie. Gdyby kladogram przedstawić w układzie współrzędnych, to zarówno oś odciętych, jak i oś rzędnych nie przedstawiałyby niczego, choć wzdłuż osi rzędnych upływał czas. W paleontologii (Stanley 1998) próbuje się niekiedy odzwierciedlać wielkość ewolucyjnych modyfikacji między taksonami długością poziomych odcinków drzewa (rysowanego w konwencji Ryc. 1.7b lub c), w rekonstrukcji filogenezy nie jest to jednak przyjęte i wzdłuż osi odciętych drzewa skaluje się dowolnie.



Ryc. 1.7. Różna forma graficzna i zawartość informacyjna drzew. Kladogramy (a–b) pomimo różnego wyglądu nie różnią się zawartością informacyjną; choć czas upływał z dołu do góry, ani oś odciętych, ani oś rzędnych nie niosą informacji, długość gałęzi nie odzwierciedla niczego, drzewo informuje jedynie o kolejności kladogenez. Także na drzewie addytywnym, czyli metrycznym (c), oś odciętych nie odzwierciedla niczego, natomiast oś rzędnych jest skalowana wielkością zmian anagenetycznych, czyli modyfikacji (M) na odpowiednich gałęziach. Na drzewie ultrametrycznym, czyli fenogramie (d), oś rzędnych jest skalowana czasem (T), który upłynął od powstania odpowiednich gałęzi

Drzewo **addytywne**, czyli **metryczne** (*additive* albo *metric tree* bądź *phylogram*), ma pionowe odcinki gałęzi skalowane proporcjonalnie do wielkości zmian ewolucyjnych, jakie zaszły wzdłuż danej gałęzi (Ryc. 1.7c). Oś rzędnych odzwierciedla tu więc wielkość tych zmian, do drzewa dołącza się niekiedy skalę (M), ale często tej skali brak. Jak widać, kolejne taksony terminalne znajdują się na różnej wysokości, bowiem kończą się nimi gałęzie tym dłuższe, im więcej zmian następowało w obrębie danej gałęzi, czyli im wyższe było tempo anagenety. Dopuszczalne na drzewie addytywnym zróżnicowanie tempa ewolucji w obrębie drzewa jest założeniem bezpiecznym i biologicznie realistycznym. Formalnie addytywność zdefiniujemy w Rozdziale 2.9.

Drzewo **ultrametryczne** (*ultrametric tree*), inaczej zwane **fenogramem** (*phenogram*) lub **dendrogramem** (*dendrogram*) zakłada ultrametryczność odległości odzwierciedlanych na nim, co również w sposób formalny zdefiniujemy w Rozdziale 2.9. W praktyce ultrametryczność oznacza, że każdy z taksonów terminalnych jest równoodległy od ko-

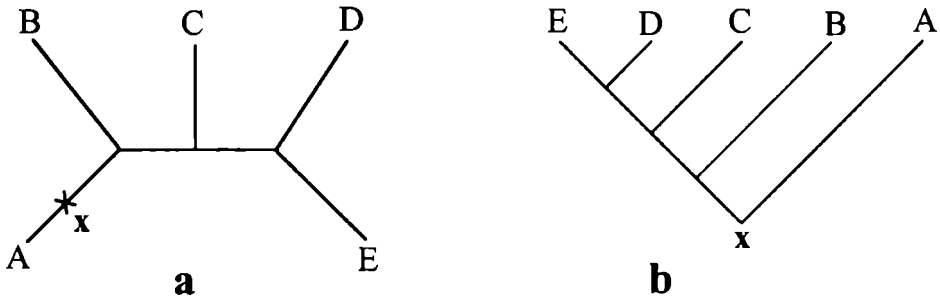
zenia drzewa (Ryc. 1.7d), a długość pionowych gałęzi odzwierciedla czas, jaki upłynął pomiędzy korzeniem a kolejnymi wewnętrznymi węzłami (HTU) i taksonami terminalnymi. Takiemu drzewu niemal zawsze towarzyszy skala (T), będąca skalą czasu. Pamiętajmy jednak, że skalowanie gałęzi czasem stosujemy do drzewa, które skonstruowano na podstawie zróżnicowania stanów cech, a więc drzewo będzie realistyczne jedynie wówczas, gdy tempo ewolucyjnych zmian było identyczne dla wszystkich gałęzi, czyli także niezmiennie w czasie. Dzieje się tak, gdy możemy założyć działanie **zegara molekularnego**. W rzeczywistości jednak zegar molekularny trudno zakładać dla cech morfologicznych, a i dla molekularnych tempo tego zegara bywa różne dla różnych molekuł i różnych organizmów (Page i Holmes 1998, Avise 2000). W sumie więc zastosowanie drzew ultrametrycznych w rekonstrukcji filogenezy jest ryzykowne, a dla danych ultrametrycznych drzewo addytywne będzie i tak identyczne z ultrametrycznym, coraz rzadziej więc używa się dendrogramów w rekonstrukcji filogenezy. Drzewa takie natomiast znakomicie odzwierciedlają ogólne podobieństwo, są więc powszechnie stosowane w taksonomii fenetycznej (patrz analiza skupisk: Rozdział 3.7), jednak wówczas skala nie ma odzwierciedlać czasu, a jedynie wielkość różnic.

1.3. Drzewa ukorzenione i nieukorzenione, liczba możliwych topologii

Jak pamiętamy, korzeniem drzewa jest hipotetyczny (a w wyjątkowych przypadkach znany) przodek wszystkich taksonów terminalnych, dla których obliczono to drzewo. Większość technik rekonstrukcji filogenezy umożliwia obliczenie drzewa jako określonego układu gałęzi i węzłów – niekiedy także długości gałęzi – ale bez określenia położenia korzenia. Drzewo takie określamy jako drzewo **nieukorzenione** (*unrooted tree*) i ma ono postać porozgałęzianego dendrytu (Ryc. 1.8a). Nie przedstawia ono chronologii rozgałęzień, bowiem nie potrafimy wskazać najstarszego fragmentu drzewa. Ponownie warto podkreślić, że drzewo nieukorzenione pozostaje drzewem, więc nie powinno być nazywane siecią (*network*), jak to często ma miejsce w literaturze taksonomicznej. Podobnie nieuzasadnione jest nieuznawanie takiego drzewa za filogenezę – obliczone technikami filogenetycznymi, pozbawione jest jedynie informacji o chronologii wydarzeń.

Jeżeli wiemy, że przodek wszystkich taksonów z analizowanego drzewa występował w punkcie (X), to możemy to drzewo ukorzenić w tym punkcie – zwykle nie mamy informacji o punkcie, a jedynie o gałęzi, więc ukorzeniamy w połowie długości tej gałęzi, przekształcając drzewo w **ukorzenione** (*rooted tree*) (Ryc. 1.8b). Jak widać, drzewo o pięciu taksonach terminalnych ukorzenić można na jednej z siedmiu gałęzi, dla pięciu taksonów drzew ukorzenionych będzie więc siedmiokrotnie więcej niż nieukorzenionych. Ukorzeniamy zarówno kladogramy, jak i drzewa addytywne, natomiast drzewa ultrametryczne są – z definicji – zawsze ukorzenione. Warto pamiętać, że typowa, dendrytowata forma drzewa nieukorzenionego bywa niechętnie rysowana przez licznych taksonomów, a większość programów rekonstruujących filogenezę takich drzew nie rysuje – operują drzewami w formie kladogramu bądź fylogramu (Ryc. 1.7a–c), także gdy są to drzewa nieukorzenione, wówczas drzewom towarzyszy uwaga o braku ukorzenienia. Zgodnie z klasyczną teorią kladystyki specjacja była równo-

znaczna z kladogenezą przynoszącą powstanie dwóch nowych gatunków i wymarcie gatunku macierzystego (Hennig 1966) – przodek nie mógł być więc współczesny gatunkom, którym dał początek. Obecnie jednak uważa się to założenie za nierealistyczne (choćby znów model specjacji izolowanych populacji peryferycznych), więc drzewo ukorzeń możemy na jednym z gatunków terminalnych. Może się też zdarzyć, że przodek badanej grupy jest znany z materiału kopalnego albo też bogaty materiał z morfologii porównawczej i zapisu paleontologicznego, z dobrze rozumianymi homologiami, umożliwi nam stworzenie precyzyjnego obrazu przodka, niejako archetypu analizowanej grupy. Wszystko to jednak zdarza się rzadko i zwykle drzewo ukorzeniać musimy na którejś z jego gałęzi.



Ryc. 1.8. Drzewo nieukorzone (a) po ukorzeniu w punkcie x (b)

Ukorzenie drzewa łatwo przeprowadzić, gdy znamy **polaryzację** (*polarity*) cech – oczywiście pierwotne stany cech występować będą w pobliżu korzenia (jeśli odwrócenie kierunku ewolucji brak bądź są nieczęste). Polaryzację cech możemy znać (bądź przynajmniej postulować) na podstawie danych porównawczych czy paleontologicznych, nieoceniona jest też znajomość ontogenezy – przy całej ostrożności z wykorzystywaniem ontogenezy jako klucza do filogenezy (Gould 1977, Wiley 1981, Falniowski 1993) stany cech pojawiające się w ontogenezie później zwykle – lecz nie w przypadku pedomorfozy – są stanami filogenetycznie zaawansowanymi w stosunku do pojawiających się wcześniej stanów pierwotnych. Polaryzacja cech na ogół jednak nie jest znana – ustala się ją dopiero na podstawie zrekonstruowanego drzewa – i ukorzenie drzewa przeprowadza się najczęściej metodą **grupy zewnętrznej** (*outgroup*). Po prostu do analizy włącza się takson (albo taksony, zależnie od techniki) zewnętrzny, czyli do grupy nie należący, lecz niezbyt od niej odległy. Teoretycznie najlepiej byłoby użyć **taksonu siostrzanego** (*sister taxon*), czyli grupy, której ostatni przodek był wspólny z przodkiem grupy badanej. Wtedy drzewo ukorzamy na gałęzi łączącej grupę zewnętrzną z grupą badaną (**wewnętrzną** – *ingroup*). W praktyce technika, choć najpowszechniej stosowana, kryje pułapki, bowiem grupę zewnętrzną łatwo błędnie wyznaczyć. Gdy należy ona do wewnętrznej, ukorzenie nie ma sensu (skądinąd położenie grupy zewnętrznej wśród taksonów badanej jest jednostronnym testem polifiletizmu grupy wewnętrznej); gdy jest odleglejsza, błędne ukorzenie jest tym prawdopodobiejsze, im odleglejsza grupa zewnętrzna, bowiem w miarę upływu czasu od oddzielenia linii rośnie udział nieskompensowanych wydarzeń, jak odwrócen, paralelizmów i konwergencji (więcej o tym w Rozdziale 4.5). Trzecią – stosowaną w ostatecz-

ności – techniką ukorzenia jest **ukorzenie na najdłuższej gałęzi** (*midpoint rooting*): choć można spekulować o odpowiadaniu największej liczby zmian najdłuższemu czasowi, a więc najstarszej części drzewa, to jest to uzasadnienie słabe, więc takie ukorzenie uznaje się za arbitralne i przeprowadza wtedy, gdy inaczej się nie da, a nie chcemy operować drzewem nieukorzenionym. Jak łatwo spostrzec, takie właśnie ukorzenie ma miejsce w fenogramie i jest odpowiednie wyłącznie wówczas, gdy dane są ultrametryczne w wyniku działania zegara molekularnego.

Liczba teoretycznie możliwych różniących się między sobą drzew zależy od liczby taksonów terminalnych n i od tego, czy drzewo jest ukorzone, czy nie. Jeszcze w XIX stuleciu znaleziono formułę określającą tę liczbę:

$$N_{\text{nieukorzenionych}} = \frac{(2n-5)!}{(n-3)! \times 2^{n-3}}, \quad N_{\text{ukorzenionych}} = \frac{(2n-3)!}{(n-2)! \times 2^{n-2}}.$$

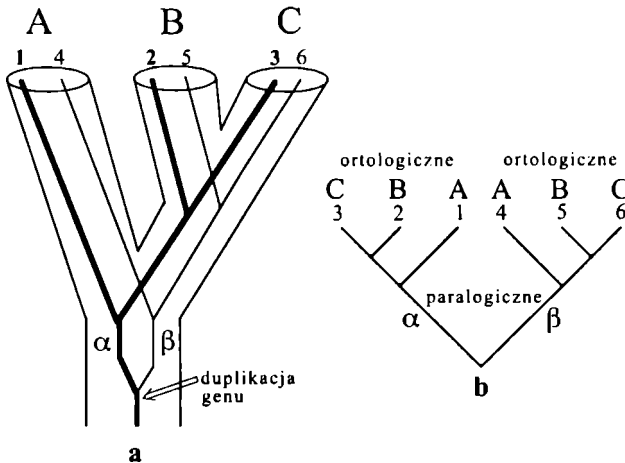
Jak widać, liczba drzew ukorzenionych dla n taksonów równa jest liczbie drzew nieukorzenionych dla $n+1$ taksonów. Poniżej zestawiono liczby możliwych drzew, obliczone dla 2–15 i 20 taksonów.

n taksonów terminalnych	N drzew nieukorzenionych	N drzew ukorzenionych
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425
11	34459425	654729075
12	654729075	1374931×10^4
13	1374931×10^6	3162341×10^5
14	3162341×10^5	7905853×10^6
15	7905853×10^6	2134580×10^8
20	2216431×10^{14}	8200795×10^{15}

Liczba możliwych drzew rośnie gwałtownie, wcześniej osiągając zupełnie niewyobrażalne wartości. Warto wskazać, że dla 135 OTU liczba możliwych drzew ukorzenionych równa jest $2,113 \times 10^{267}$, co przekracza liczbę cząsteczek w znanym Wszechświecie (Page i Holmes 1998). Można by to uznać za kuriozalną ciekawostkę, gdyby nie to, że większość metod rekonstrukcji filogenezy nie tworzy drzew, a jedynie umożliwia ocenę (przy lepiej lub gorzej dobranym modelu oraz lepszej czy gorszej technice jego weryfikacji), które z dwóch drzew jest lepsze, bardziej zgodne z danymi i modelem. Teoretycznie należałoby więc porównać wszystkie drzewa, a to już przy niewielkiej liczbie OTU przestaje być możliwe. Charakter formuły obliczającej liczbę drzew wskazuje też jednoznacznie, że stałe zwiększanie mocy obliczeniowej komputerów też wiele nie pomoże – już dla 16 taksonów drzew jest 27 razy więcej niż dla 15, a dla 21 – 37 razy więcej niż dla 20. Nieuniknione więc jest i będzie stosowanie technik analizy części drzew (Rozdział 4.3).

1.4. Filogeneza genów i organizmów, wewnątrz i powyżej gatunku

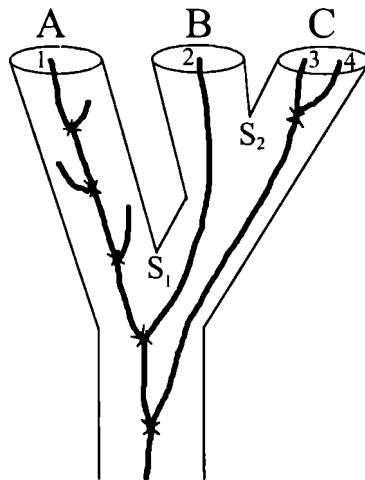
Ewolucja jakiejś struktury czy molekuly nie musi odzwierciedlać ewolucji całego organizmu, u którego ta struktura czy molekula występuje. Możemy zupełnie prawidłowo zrekonstruować filogenezę jakiegoś genu – sekwencji DNA – a nie będzie ona prawidłowo przedstawiała filogenezy organizmu mającego ten gen. Jedną z przyczyn jest – częsta – duplikacja genów. Każdy z trzech gatunków A, B, C (Ryc. 1.9a) ma dwie kopie tego samego genu, które od momentu duplikacji ewoluowały niezależnie. **Homologiczne** (Rozdział 1.5, 2.2 i 2.12) w węższym znaczeniu – leżące na tej samej kopii genu, czyli **ortologiczne** – będą więc wzajemnie sekwencje 1, 2 i 3, a także 4, 5 i 6, natomiast sekwencje 1, 2 i 3 będą w stosunku do 4, 5 i 6 **paralogiczne**. Znajomość wszystkich sześciu sekwencji umożliwi prawidłową rekonstrukcję filogenezy (Ryc. 1.9b), natomiast nie musi tak być wówczas, gdy nie wiemy o duplikacji i zsekwenjowaliśmy, powiedzmy, geny 2, 4 i 6: wynikiem będzie drzewo o topologii ((A,C),B), bowiem sekwencje 4 i 6 są bliższe sobie niż sekwencji 2, pomimo że gatunek B jest bliższy gatunkowi A niż gatunek C.



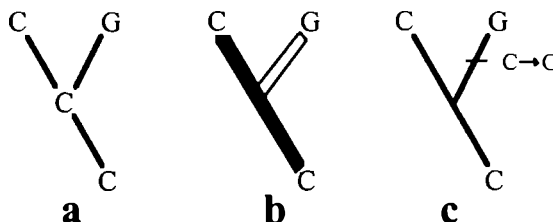
Ryc. 1.9. Duplikacja genu (a), w wyniku której powstały paralogiczne geny α i β , może być przyczyną błędnej rekonstrukcji. Wiedząc o duplikacji i znając wszystkie sześć sekwencji, prawidłowo zrekonstruujemy filogenezę (b), natomiast znajomość jedynie paralogicznych sekwencji pociągnie za sobą błędną rekonstrukcję (patrz tekst)

W obrębie gatunku następuje **sortowanie rodów** (*lineage sorting*), co w połączeniu z występującym u przodka **ancestralnym polimorfizmem** (*ancestral polymorphism*) może prowadzić – również dla homologicznych genów – do obrazów filogenezy genu niezgodnych z filogenezą gatunku. Filogenezę genu śledzić możemy na zrekonstruowanym drzewie, posuwając się w dół, czyli od czasów nowszych ku starszym, znajdując ich wspólnych przodków w miejscach **złań linii ewolucyjnych** (*coalescence*). Jak już mówiliśmy, filogeneza na poziomie osobników w obrębie rozmnażającego się płciowo gatunku nie daje się wierniej przedstawić drzewem, bowiem stałym elemen-

tem jest krzyżowanie. Można jednak takie drzewo przedstawiać dla genów, rozpatrywanych oczywiście jako **allele**. Praktycznie rozważania takie najlepiej prowadzić na danych z natury haploidalnych – jak mitochondrialne DNA czy DNA z chromosomu X (Avice 2000). Dysponując jedynie filogenezą pojedynczego genu, przedstawioną na Ryc. 1.10, nieuchronnie zrekonstruujemy topologię ((A,B),C) zamiast prawdziwej (A,(B,C)). Stanie się tak, bowiem allel 2 jest bliższy allelowi 1 niż allelom 3 i 4. Zauważmy, że allel ancestralny dla alleli 3 i 4 oraz allel 2 oddzieliły się kolejno od linii zakończonej allelem 1 jeszcze przed specjacją S_1 , choć występowały u jednego gatunku aż do znacznie późniejszej specjacji S_2 . Podobnych sytuacji będzie tym więcej, im mniejsza jest częstość zakończonych procesów specjacji, a im większa – mutacji zapewniających powstawanie nowych alleli. Ancestralny polimorfizm, połączony z nieuchronnym wykorzystywaniem w rozrodzie przez kolejne generacje jedynie części dostępnych alleli, a także doбором naturalnym dla cech selekcyjnie nieneutralnych, często prowadzi do sytuacji, gdy filogeneza genu będzie różna od filogenezy na poziomie gatunków. Dlatego też ryzykowne jest konstruowanie filogenez opartych na jednym czy nawet paru genach – im więcej genów uwzględniamy, tym większym zaufaniem obdarzyć możemy uzyskaną rekonstrukcję.



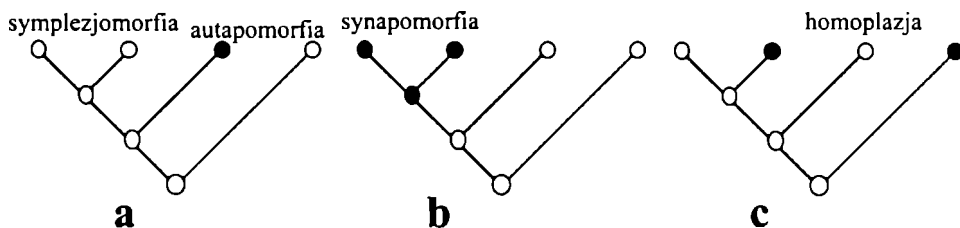
Ryc. 1.10. Filogeneza genu nie musi odpowiadać filogenezie organizmu. Dwukrotna duplikacja genu poprzedza specjacją S_1 , a trzecia duplikacja ma miejsce po zakończeniu specjacji S_2 , szereg linii genu wygasa. Rekonstrukcja filogenezy organizmów na podstawie genu w tym przypadku musi dać błędny wynik (patrz tekst)



Ryc. 1.11. Trzy różne konwencje (a, b, c) znaczenia na drzewie stanów cech i zmian tych stanów

1.5. Rekonstrukcja ewolucji cech, grupowanie kladystyczne i fenetyczne

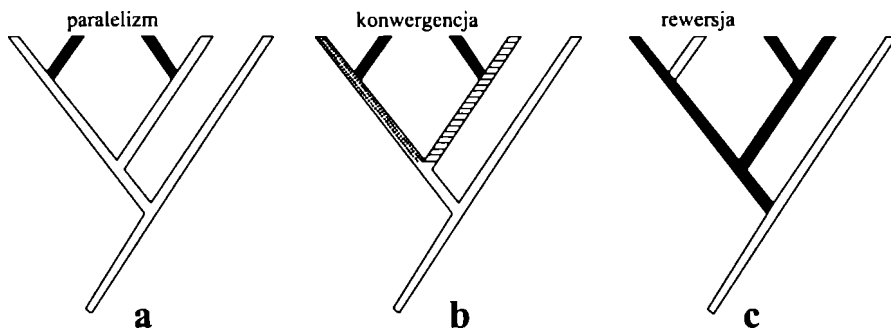
Jak już mówiliśmy, znajomość stanów cech taksonów terminalnych umożliwia rekonstrukcję filogenezy dla tych taksonów, a z kolei zrekonstruowana filogeneza – w postaci drzewa – umożliwia śledzenie ewolucji stanów tych cech – odtwarzanie historii ich zmian. Na zrekonstruowane drzewo można nanieść stany cech taksonów terminalnych, a następnie, posuwając się w kierunku przeciwnym do upływu czasu – czyli w dół drzewa – zrekonstruować kolejne stany cech i ich zmiany, zaznaczone graficznie na jeden z trzech przyjętych sposobów (Ryc. 1.11). Rekonstrukcja taka jest często niejednoznaczna, zajmiemy się tym w Rozdziale 4.5. Jeżeli np. na jednej z gałęzi drzewa cytozynę zastąpiła guanina (Ryc. 1.11), to w takim przypadku cytozyna jest stanem pierwotnym, czyli – w terminologii kladystycznej – **plezjomorfia** (inaczej **plezjomorfizm**), a guanina – stanem ewolucyjnie zaawansowanym, czyli **apomorfia** (inaczej **apomorfizm**). Plezjomorfia wspólna dla więcej niż jednego taksonu to **symplezjomorfia** (**symplezjomorfizm**). Apomorfia występująca u jednego taksonu – w jednym kladzie – to **autapomorfia** (**autapomorfizm**), zaś występująca u więcej niż jednego taksonu – kladu – nazywamy **synapomorfia** (**synapomorfizm**). Aby jednak definicja synapomorfii była pełna, musimy dodać, że jednocześnie występowanie zaawansowanego stanu cechy u więcej niż jednego taksonu – kladu – musi być następstwem **homologii**, co oznacza, że ostatni wspólny przodek tych wszystkich taksonów mających ten stan cechy już ten stan posiadał (Ryc. 1.12). Jeżeli go nie posiadał, to zamiast o homologii mówimy o **homoplazji** (Ryc. 1.12), wykluczając tym samym synapomorfie.



Ryc. 1.12. Polaryzacja stanów cech: zgodnie z konwencją stan plezjomorficzny, czyli prymitywny,znaczmy okręgiem lub prostokątem niewypełnionym (jasnym), a apomorficzny, czyli zaawansowany, okręgiem lub prostokątem wypełnionym (ciemnym). Gdy apomorfia pojawia się u pojedynczego taksonu terminalnego, mówimy o autapomorfii (a), gdy u więcej niż jednego oraz ich ostatniego wspólnego przodka – mamy synapomorfie (b), gdy zaawansowane stany cechy są rozsiiane na drzewie, to odzwierciedlają nie homologie, a homoplazje (c)

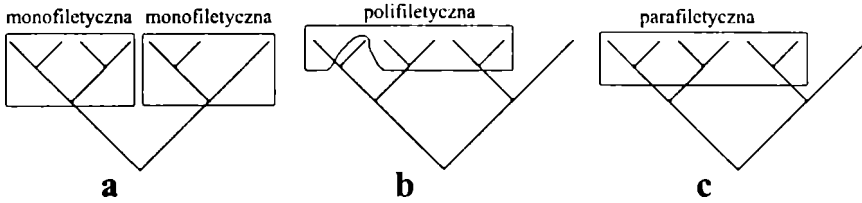
Wspólne posiadanie pierwotnego stanu cechy wskazywać może na powolną ewolucję, lecz niczego nie wnosi do obrazu pokrewieństw: przyjmuje się, że dla rekonstrukcji ewolucji użyteczne są jedynie synapomorfie, zaś taksonomia kladystyczna rygorystycznie wymaga synapomorfii jako warunku wyodrębnienia naturalnej grupy. Często jednak za synapomorfie błędnie uznaje się stany cech będące następstwem homoplazji. Homoplazje (Ryc. 1.13) dzielimy na **paralelizmy**, **konwergencje** i **odwrócenia** (czyli

rewersje, inaczej **wtórne utraty** albo **uwstecznienia**). Paralelizmem, czyli ewolucją równoległą, określamy niezależne uzyskanie takiego samego stanu zaawansowanego z tego samego stanu pierwotnego (Ryc. 1.13a). Dla bardziej złożonych cech morfologicznych zachodzi domniemanie, że skoro równolegle powstał taki sam stan, to ostatni wspólny przodek musiał być bliski jego uzyskania, a więc paralelizmy mają pewną wartość dla określania pokrewieństw. Jest to jednak właśnie domniemanie, a już zupełnie nie nadaje się dla cech molekularnych. Nawet takiej domniemanej wartości nie mają konwergencje, gdy taki sam stan cechy powstał z różnych stanów pierwotnych (Ryc. 1.13b). Odwrócenia to sytuacje, gdy stan zaawansowany zmienia się w stan pierwotny, taki sam jak przed uzyskaniem stanu zaawansowanego (Ryc. 1.13c). Pojęcia synapomorfii i homologii nie są bezwzględными atrybutami określonych stanów cech – zależą od **poziomu uniwersalności**: synapomorfia wyróżniająca np. rodzinę staje się symplezjomorfia dla rodzajów w obrębie tej rodziny, skrzydła ptaka i nietopecza są homologiczne jako przednie kończyny, lecz jako struktury służące do lotu – nie.



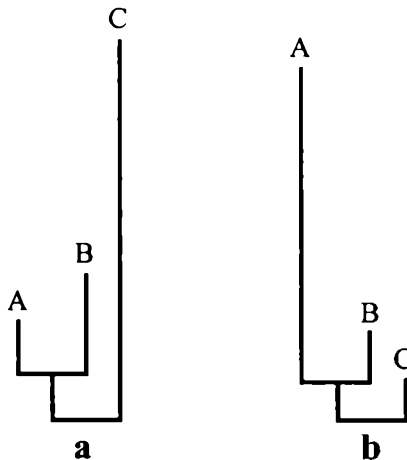
Ryc. 1.13. Różne rodzaje homoplazji: paralelizm (a), konwergencja (b) i rewersja, czyli odwrócenie (c)

Interpretacja zrekonstruowanych drzew filogenetycznych to obszerna dziedzina, którą w tej książce się nie zajmujemy. Konieczne jest jednak wprowadzenie paru pojęć. W klasyfikacji dąży się do uznawania jedynie **grup naturalnych**, a za takie powszechnie uznaje się wyłącznie grupy **monofiletyczne**. W tym miejscu jednak zgodność między taksonomami wyczerpuje się i monofiletyzm rozumiany bywa bardzo rozmaicie. Do skrajnych zaliczyć możemy podejście Simpsona (1961), dla którego takson monofiletyczny to taki, który można wyprowadzić od jednego przodka o randze taksonomicznej nie niższej niż taksonu, którego monofiletyczność jest rozpatrywana. Nietrudno wykazać, że stosując tę definicję, trudno byłoby znaleźć takson, którego monofiletyczności nie dałoby się dowieść. W dodatku przodkiem, powiedzmy, gromady nie może być gromada ani nawet rodzina, a jedynie gatunek, bowiem dzieje powstawania wszelkich jednostek systematycznych to ostatecznie nic więcej niż kompletny zestaw wszystkich zakończonych powodzeniem procesów specjacji – innych mechanizmów kladogenezy po prostu brak. Drugi biegun to koncepcja taksonomii filogenetycznej (Hennig 1966, Wiley 1981): monofiletyczny takson to takson zawierający wszystkich potomków przodka tego taksonu, wraz z tym przodkiem (Ryc. 1.14a).



Ryc. 1.14. Klasyfikacja w grupy: monofiletyczne (a), polifiletyczną (b) i parafiletyczną (c)

Wynikiem niedostatecznej znajomości homologii, przez co nie rozpoznano szeregu homoplazji, będzie wyróżnianie grup **polifiletycznych** (Ryc. 1.14b), czyli złożonych z fragmentów różnych kładów: ostatni wspólny przodek taksonów należących do grupy polifiletycznej był także przodkiem szeregu linii ewolucyjnych do tej grupy niewliczonych. Odrębny przypadek niemonofiletizmu to grupa **parafiletyczna** (Ryc. 1.14c), czyli obejmująca nie wszystkich potomków wspólnego przodka i/lub nieobejmująca tego przodka. Grupy parafiletyczne najczęściej łączą organizmy, które mają szereg symplezjomorfii, co odróżnia je od wyodrębnianych z nich linii odznaczających się autapomorfiami. Klasyczny przykład grupy parafiletycznej to gady, a także ślimaki przodoskrzelne zaliczane do Archaeogastropoda. Kładysty wykluczają uznawanie grup parafiletycznych, lecz nie systematycy ewolucyjni. Specjalny przypadek grupy parafiletycznej to grupa **ortofiletyczna**, czyli wspólny przodek wraz z potomkami z wyjątkiem jednej linii – taką grupą są choćby jednotarczowce Trybliidiida, a wyłączone z nich linie – ślimaki. Grupy ortofiletyczne są niekiedy wygodne i bywają uznawane przez mniej ortodoksyjnych kładystów.



Ryc. 1.15. Pomimo zdecydowanie wyższego tempa ewolucji w kładzie C (a) nie zachodzi tu niebezpieczeństwo klasyfikacji horyzontalnej, bowiem ogólne podobieństwo w tym przypadku daje grupowanie ((A,B),C), takie samo jak analiza filogenetyczna. Jeżeli jednak wysokim tempem ewolucji odznaczać się będzie kład A (b), to przy niedostatecznej wiedzy o homologiach i ewolucji cech nieuchronnie rozpozna się błędnie grupowanie (A,(B,C)): będzie to grupowanie na podstawie ogólnego podobieństwa, czyli fenetyczne, a klasyfikacja będzie horyzontalna

Zróznicowanie tempa ewolucji w obrębie rekonstruowanej filogenezy zdarza się często i może przysparzać poważnych problemów. Bywa, że wolniejsza ewolucja w obrębie kladu (Ryc. 1.15a) dodatkowo podkreśla jego odrębność i wówczas błąd w rekonstrukcji filogenezy popełnić trudno. Bywa jednak odwrotnie (Ryc. 1.15b): wolne tempo ewolucji wzdłuż gałęzi zakończonych taksonami B i C nieuchronnie prowadzi do większego podobieństwa – w wyniku zachowania wielu plezjomorfii – pomiędzy tymi taksonami niż pomiędzy B a kończącym szybko ewoluującą linię taksonem A. Pomimo to kladystyczne grupowanie, odzwierciedlające związki genealogiczne, przedstawia drzewo ((A,B),C), choć analiza różnic stanów cech, bez prawidłowo rozpoznanych homologii, doprowadzi najprawdopodobniej do drzewa (A,(B,C)). Będzie to grupowanie fenetyczne, oparte na ogólnym podobieństwie, a stworzona w ten sposób systematyka będzie **systematyką poziomą**, czyli **horyzontalną**, odzwierciedlającą nie **linie ewolucyjne** (*clades*), lecz stopnie **ewolucyjnego zaawansowania** (*grades*).

2. Dane

2.1. Cechy „kladystyczne” i „fenetyczne”

Podstawą jakiegokolwiek grupowania organizmów, również rekonstrukcji filogenetycznych związków pomiędzy nimi, jest znajomość **cech**, opisujących te organizmy. Często wyróżnia się dwa pojęcia: cechy (*character*) i **stanu cechy** (*character state*), co ułatwia precyzyjny opis. A więc, zamiast mówić, że zielona barwa oczu to cecha określonego taksonu, jako cechę definiujemy barwę oczu, a cecha ta przyjmować może stany: zielony, niebieski, różowy, itd. Oczywiście u podstaw wykorzystania stanów cech dla rekonstrukcji filogenezy leży założenie, że specjacji towarzyszyła modyfikacja stanów cech i/lub tempo specjacji nie przewyższało tempa modyfikacji stanów cech. Ponadto cechy użyteczne dla taksonomii powinny odznaczać się możliwie niewielką zmiennością wewnątrz taksonów, a jak największą pomiędzy taksonami. W praktyce bywa różnie i niejednokrotnie zmuszeni jesteśmy korzystać z cech niezupełnie spełniających powyższe warunki.

Choć systematyka organizmów opiera się głównie na morfologii i – w coraz większym stopniu – cechach molekularnych, to właściwie każda biologiczna informacja może być przydatna dla taksonomii, jeżeli tylko wykazuje różnice pomiędzy organizmami. Można więc mówić o cechach morfologicznych, kariologicznych, biochemicznych, fizjologicznych, behawioralnych, ekologicznych czy biogeograficznych. Linneuszowska idea opierania systematyki na cechach widocznych i łatwych do badania oraz jednoznacznego opisu, pomimo niewątpliwych walorów praktycznych, utrzymać się nie dała. Nie ma cech ogólnie „dobrych” czy „gorszych”, a systematykę często opierać musimy na cechach bardzo trudnych technicznie do badania i/lub wykazujących znaczną zmienność wewnątrz taksonów przy małych różnicach pomiędzy taksonami. Istotną częścią pracy taksonoma jest stałe testowanie przydatności cech stosowanych w systematyce danej grupy oraz poszukiwanie nowych, wcześniej niewykorzystywanych. Ogólne zasady wyboru cech omawiają choćby Simpson (1961), Mayr (1969) i Wiley (1981), zagadnienie to wykracza zresztą poza zakres tej książki.

Cechy stanowią **zmienne**, które z formalnego punktu widzenia podzielić można na **ilościowe** i **jakościowe**. Zmienna ilościowa to cecha, której stanami, czyli realizacjami zmiennej są wartości liczbowe. Zmienne ilościowe podzielić można na **ciągłe** i **dyskretne**. Ciągła zmienna przyjmować może nieprzeliczalną liczbę wartości, często pochodzących z określonego przedziału liczbowego; przykładem może być długość,

wysokość, asymilacja tlenu w jednostce czasu, temperatura. Zmienna dyskretna przyjmować może przeliczalną, w praktyce skończoną liczbę wartości, jak liczba palców kończyny, liczba kolców, liczba segmentów. W praktyce jednak pomiarów dokonujemy zawsze z określoną dokładnością, a więc w badanym zakresie mieści się skończona liczba jednostek pomiaru (np. 0,01 mm), czyli traktujemy zmienną ciągłą jak dyskretną.

Zmienna jakościowa to cecha, której stanami są pewne kategorie, a nie wartości liczbowe. Choć w praktyce często koduje się ich stany za pomocą liczb, to liczby te mają charakter umowny i nie wolno dokonywać na nich działań arytmetycznych. Zmienne jakościowe podzielić można na **nominalne** i **porządkowe**. Zmienne nominalne to takie zmienne jakościowe, dla których stanów powiedzieć możemy jedynie, że są takie same lub różne, a jakiegokolwiek inne porównanie stanów cechy nie jest dozwolone. Takimi zmiennymi są np. płeć, obecność lub brak płuca, barwa piór. Szczególnym przypadkiem zmiennej nominalnej jest zmienna **binarna**, czyli **zero-jedynkowa**. Przykładem obecność lub brak określonej struktury. Zmienna porządkowa to taka zmienna jakościowa, która określa porządek od „najniższej” do „najwyższej”, np. kolejność wykluwania się młodych z jaj. Warto podkreślić, że choć stany cechy tego rodzaju możemy zapisać w postaci: 1, 2, 3, 4, 5, 6, ..., n , to jednak nie możemy zakładać, że różnica pomiędzy 1 a 2 jest taka sama czy podobna jak pomiędzy 5 a 6 ani też dwukrotnie mniejsza niż między 4 a 6.

Aparat matematyczny jest najlepiej opracowany dla zmiennych ilościowych ciągłych, dla innych cech występują rozmaite ograniczenia. Oczywiście różne rodzaje cech wymagają użycia różnych technik i będziemy o tym w dalszej części tej książki szerzej mówili. W tym miejscu musimy jednak zwrócić uwagę na inny, biologiczny aspekt wykorzystywania różnego rodzaju cech. Otóż taksonomia filogenetyczna opiera się niemal wyłącznie na cechach jakościowych, ewentualnie wykorzystując też cechy ilościowe dyskretnie. Większość technik kladystycznych wymaga takich właśnie cech, ale nie to jest najważniejsze. W końcu zmienną ilościową ciągłą przekodować możemy na nieciągłą. Problem leży gdzie indziej. Otóż zmienne jakościowe opisują występowanie lub brak, lub znaczące modyfikacje określonych struktur, czemu przypisać zwykle możemy (lub moglibyśmy, mając bardziej pełną wiedzę) konkretne znaczenie biologiczne, określony kontekst ewolucyjny. Trudno tymczasem to samo powiedzieć o zmianie wartości jakiegoś wymiaru czy np. proporcji, dla której w dodatku klasy określono arbitralnie. Zestawianie obok siebie choćby obecności/braku kolca z wartością pomiaru w przedziale $\langle 0,50-0,59 \rangle / \langle 0,60-0,69 \rangle$ budzić musi zrozumiałą sprzeciw, nawet jeżeli tej drugiej cesze przypiszemy znacznie niższą wagę. Tak więc najprościej byłoby – choć często ubóstwo dostępnych cech uniemożliwia to – cechy ilościowe, zwłaszcza ciągłe, pozostawić analizie fenetycznej, a analizę filogenetyczną opierać na cechach jakościowych.

Wszystkie metody rekonstrukcji filogenezy, lecz nie fenetyczne, zakładają wzajemną niezależność cech. Upraszcza to aparat matematyczny, bowiem umożliwia nieuwzględnianie macierzy korelacji bądź kowariancji pomiędzy cechami, pozwala też na rozpatrywanie złożonych rekonstrukcji jako zestawu prostszych, które zwyczajnie się sumują. Odstępstwa od niezależności cech, choć niepożądane, nie muszą pociągać za sobą poważniejszych następstw, jeżeli wykorzystujemy metodę **kladystyczną**, czyli **redukcjonistyczną** (*parsimony*), natomiast zupełnie fałszują wynik w przypadku metod opartych na **maksymalizacji wiarygodności** (*maximum likelihood*: Felsenstein 1993). Oczywiście spełnienie warunku wzajemnej niezależności cech w praktyce nie

jest właściwie możliwe. Brak statystycznego testu na niezależność: możemy jedynie badać istotność zależności określonego typu, a więc wykluczyć niezależność, ale nie wykazać zależności np. liniowej nie oznacza braku zależności choćby logarytmicznej. Na koniec, nawet jeżeli przekształcimy nasze dane w zmienne kanoniczne, formalnie całkowicie niezależne, to i tak nie możemy wykluczyć jakiegś biologicznej zależności pomiędzy nimi, w dodatku sens takiej zależności pozostanie zupełnie niezrozumiały, bowiem rozpatrujemy wówczas całkowicie abstrakcyjne zmienne, których biologicznej interpretacji (jeżeli w ogóle taka istnieje) zupełnie nie znamy.

Tak więc, pomimo formalnego wymogu niezależności cech, musimy się godzić z niespełnianiem tego warunku, w mniejszym lub większym stopniu. Nie znaczy to jednak, że warunek ten zupełnie ignorujemy: po prostu kierujemy się zdrowym rozsądkiem i unikamy cech, które muszą być z sobą skorelowane. Oczywiście duże wymiary całego organizmu muszą być skorelowane z wymiarami jakiegś jego części, obecność czułka jest warunkiem obecności urzęsienia na tym czułku, zakopywanie się w ziemi i kończyzna przystosowana do kopania też będą z sobą związane. Niezależność cech nie jest natomiast warunkiem stosowania metod fenetycznych. Właściwie jest wręcz odwrotnie: większość metod **statystycznej analizy wielowymiarowej** (*multivariate statistical analysis*) opiera się na badaniu zależności pomiędzy cechami i dla cech całkowicie niezależnych analiza taka zwyczajnie nie byłaby możliwa. W praktyce jednak takiej niezależności nigdy nie spotkamy – do zagadnienia tego jeszcze wrócimy.

2.2. Homologie i ich testowanie

Jeżeli porównujemy przydatność cech dla taksonomii fenetycznej z jednej strony, a filogenetycznej czy ewolucyjnej z drugiej, to jeszcze ważniejsza jest inna różnica. Otóż dla taksonomii fenetycznej przydatna jest każda cecha i wszystkie cechy ważone są tak samo. Dla taksonomii filogenetycznej czy ewolucyjnej natomiast znaczenie mają jedynie cechy **homologiczne**, a ważone mogą być różnie, do czego wrócimy w dalszych rozdziałach. Rekonstrukcja filogenezy wymaga użycia wyłącznie cech homologicznych, a nieprawidłowe rozpoznanie lub nieznanostwo homologii pociąga za sobą klasyfikację opartą na ogólnym podobieństwie, czyli fenetyczną, bez względu na zastosowaną metodę rekonstrukcji (Falniowski i Szarowska 1995).

Homologią nazywamy zgodność organu lub jego części u dwóch lub więcej taksonów, jeżeli mamy podstawy do założenia, że zgodność ta jest następstwem odziedziczenia cechy, z ewentualną późniejszą modyfikacją, po wspólnym przodku tych taksonów. Homologia jest więc pojęciem **filogenetycznym**, choć ustala się ją na podstawie kryteriów **morfologicznych**. Ogromnie pomocne w ustalaniu homologii są informacje paleontologiczne i embriologiczne. Istnieje szereg **kryteriów**, pomocnych w rozpoznawaniu homologii: **budowy**, **położenia** w stosunku do innych struktur, **współwystępowania** z innymi strukturami oraz **szeregu przejściowego** od jednego stanu cechy do drugiego, występującego w materiale kopalnym lub wśród gatunków współczesnych (Remane 1952, Riedl 1978, Wiley 1981, Wagner 1989). Im bardziej złożona jest budowa rozpatrywanej struktury, tym większe prawdopodobieństwo, że tak samo zbudowana struktura jest homologiczna. Warto więc pamiętać, że uderzające podobieństwo budowy prostych układów nie musi dowodzić homologii. Z drugiej strony, im bardziej niezwykła, **unikatowa** jest dana struktura, tym większa szansa, że nie powstała w filo-

genezie więcej niż jeden raz, a więc jest homologiczna. Struktury, których homologie jest sens rozpatrywać, muszą być **wytwarzane indywidualnie** w rozwoju ontogenetycznym: jeżeli np. u jakichś organizmów występują pola ząbków czy kolców, to nonsensowne byłoby próbowanie homologizacji poszczególnych ząbków (kolców). Ostatecznie homologie, proponowane na podstawie powyższych kryteriów, testowane są zrekonstruowaną filogenezą.

To, czy dane dwie struktury są homologiczne, zależy też od **poziomu uniwersalności**. Muszla ślimaka i małża to struktury homologiczne, wytwarzane w podobny sposób przez płaszcz, ale dwuskorupkowa, parzysta muszla ślimaka *Berthelinia* nie jest bynajmniej homologiczna z parzystymi muszlami małży, w większym stopniu niż „normalne”, pojedyncze muszle ślimaków. Tylnie kończyny wszystkich kręgowców są oczywiście homologiczne, ale przystosowana do przemieszczania się skokami tylna kończyna kangura nie jest homologiczna z podobnie zbudowaną kończyną zająca w większym stopniu niż np. z tylną kończyną słonia. Przykłady można by mnożyć. Dla potrzeb taksonomii nie wystarcza informacja, że dane struktury na jakimś, wysokim poziomie uniwersalności są homologiczne – zwykle interesuje nas, czy dany stan cechy, identyczny lub bardzo podobny do stanu tejże cechy u innego organizmu, jest z nim homologiczny. Jest homologiczny wówczas, gdy ostatni wspólny przodek obu organizmów już ten stan cechy posiadał – a więc tak rozumiana homologia jest równoznaczna z **synapomorfia**, a synapomorfie możemy testować, rekonstruując filogenezę.

Nie zawsze struktury podobne, nawet bardzo, są homologiczne. Wówczas mówimy o **homoplazjach**. Może być tak, że ostatni wspólny przodek jeszcze tego stanu cechy nie miał, choć był bliski jego uzyskania, a stan cechy pojawił się wkrótce po rozdzieleniu kładów – mówimy wówczas o **paralelizmie**. Jeżeli natomiast obserwowane podobieństwo struktur jest wynikiem upodobnienia struktur wyjściowo różnych, w następstwie pełnienia podobnych funkcji, mówimy o **konwergencji**. Znowu: choć skrzydło ptaka i nietoperza, rozpatrywane na poziomie kręgowców, są strukturami homologicznymi, to jednak na niższym poziomie to typowy przykład konwergencji. Wreszcie, ewolucja może zmieniać kierunek, tak że ze stanu zaawansowanego powstaje wtórnie stan pierwotny – mówimy wówczas o **odwróceniach**, czyli **rewersjach**, i znów podobieństwo struktur nie dowodzi pokrewieństw.

Terminu „homologia” w biologii molekularnej używa się niekiedy w znaczeniu podobieństwa czy identyczności (Moritz i Hillis 1996); mówi się wówczas o, powiedzmy, „90% homologiczności”. Dla rekonstrukcji filogenezy jednak również sekwencje białek czy aminokwasów muszą być podobne lub identyczne w następstwie pochodzenia od wspólnego przodka, a nie tylko podobne. Podobieństwo może być bowiem następstwem konwergencji lub konwersji genu. Na poziomie molekularnym całkowita identyczność jedynie sugeruje homologię – nukleotydów jest zaledwie cztery, aminokwasów 20, więc obecność tego samego nukleotydu lub aminokwasu na danej pozycji w porównywanych sekwencjach jest bardzo prawdopodobna, choć historia porównywanych pozycji może być zupełnie różna dla obu sekwencji. Homologia, czyli podobieństwo sekwencji w wyniku wspólnej historii, to warunek konieczny, lecz bynajmniej nie dostateczny dla prawidłowej rekonstrukcji filogenezy na podstawie różnic pomiędzy tymi sekwencjami. W taksonomii bowiem dążymy do odtworzenia pokrewieństw pomiędzy organizmami, a nie pomiędzy samymi genami. Interesują nas więc jedynie sekwencje **ortologiczne** (Fitch 1970), czyli takie, których historię możemy zrekonstruować wstecz aż do specjacji, w wyniku której linie ewolucyjne rozdzieliły

się i w każdej z nich niezależnie zachodziły zmiany w sekwencji. Nie możemy natomiast wykorzystywać sekwencji **paralogicznych**, czyli powstałych w następstwie duplikacji genu (Fitch 1970). Podobnie bezużyteczne dla rekonstrukcji filogenezy są sekwencje **kzenologiczne** (Page i Holmes 1998), czyli obecne w następstwie horyzontalnego transferu (np. przez retrowirusy). Szerzej homologię sekwencji omawiają Moritz i Hillis (1996).

Prawidłowo rozpoznane dwie homologiczne, ortologiczne sekwencje przy porównaniu następczą kolejnych problemów z homologią. Teoretycznie każda zasada w kwasie nukleinowym czy każdy aminokwas w białku może być zastąpiony innym (choć biologicznie jest tu szereg ograniczeń), toteż uznaje się każdą zasadę/amino kwas za odrębną cechę, niezależną od innych. A więc musimy porównywać homologiczne pary zasad/amino kwasów: jest to tzw. **homologia pozycyjna**. Gdyby kwasy nukleinowe podlegały w trakcie ewolucji jedynie punktowym mutacjom, wystarczyłoby zwykłe porównanie pozycji za pozycją. Tak też na ogół jest, jeżeli porównujemy sekwencje bliskich sobie taksonów, których ostatni wspólny przodek występował stosunkowo niedawno. Przy wszelkich porównaniach taksonów mniej sobie bliskich pojawia się natomiast problem deficytów, duplikacji, translokacji i inwersji. Wówczas ustalenie homologii pozycyjnej przestaje być łatwe i często nie może być jednoznaczne. Konieczne staje się **współosiowanie**, czyli zestawianie pozycji homologicznych (*alignment*) dla porównywanych sekwencji. Jest to najbardziej niepewny, potencjalnie generujący największe błędy etap rekonstrukcji filogenezy, jego metody przedstawiają Hillis i inni (1996); wrócimy do tego jeszcze omawiając transformację sekwencji w odległości (Rozdział 2.12).

Homologię dla białek allozymatycznych, będących ekspresją określonych *loci*, ustala się na podstawie szeregu kryteriów, dotyczących ich budowy, funkcji i ekspresji. Większość paralogicznych protein łatwo rozpoznać jako produkty odrębnych *loci*, na podstawie obecności w innych tkankach czy innej migracji elektroforetycznej. Pamiętajmy zresztą, że elektroforeza allozymatyczna jest techniką porównawczą, a porównanie dwóch taksonów może być wykorzystane w rekonstrukcji filogenezy jedynie wówczas, gdy w którychś z *loci* występują te same allozomy – technika ta więc z konieczności użyteczna jest w dość wąskim zakresie, od wewnątrzgatunkowego po bliskie sobie gatunki, choć „bliskość” ma krańcowo różne znaczenie, zależnie od grupy organizmów.

2.3. Cechy jakościowe: polaryzacja, serie transformacyjne, wagi

Jak już wspominaliśmy, dla analizy kładystycznej najdogodniejsze są dane jakościowe: binarne lub wielostanowe, a formalnie podobnie przydatne są cechy ilościowe dyskretne, choć biologiczne znaczenie różnic pomiędzy kolejnymi stanami cechy dyskretnej zazwyczaj jest mniejsze niż pomiędzy różnymi stanami cechy jakościowej. Z drugiej strony warto zauważyć, że rozróżnienie pomiędzy cechami jakościowymi a ilościowymi, tak jednoznaczne w matematyce, w biologii ma często charakter formalny: jeżeli za stany cechy uznamy przykładowo jajowaty albo wstęgowaty kształt liścia, to w gruncie rzeczy opisujemy obrazowo cechę, dającą się wyrazić pomiarami. To jeszcze jeden przykład wskazujący, że cech nie możemy wybierać ani analizować

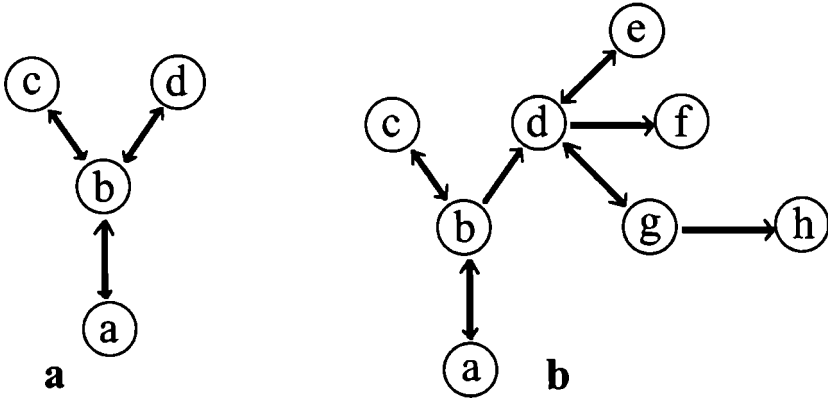
mechanicznie i każdorazowo znaczenie biologiczne obserwowanego zróżnicowania powinno być rozważone.

Początkowo programy komputerowe, używane do rekonstrukcji filogenezy, pozwalały na wykorzystywanie jedynie zmiennych binarnych (zero-jedynkowych), współczesne nie mają tego ograniczenia i z powodzeniem możemy stosować zmienne wielostanowe. Jeżeli zmienna ma dwa stany: a i b , to jeden z nich jest (w obrębie badanej grupy lub jej fragmentu – tu znów istotny jest poziom uniwersalności) **pierwotny**, a drugi **zaawansowany** ewolucyjnie i pochodzi od tego pierwszego. Jeżeli więc wskazujemy, że stan a jest pierwotny, gdy b zaawansowany, to określamy w ten sposób **polaryzację** cechy. Znajomość polaryzacji cechy jest pożądana i często możemy ją określić na podstawie danych paleontologicznych, embriologicznych lub poprzez porównanie z pokrewnymi grupami. Zwłaszcza kryterium embriologiczne jest często użyteczne: stan cechy pojawiający się w ontogenezie później jest stanem zaawansowanym filogenetycznie w stosunku do stanu pojawiającego się w ontogenezie wcześniej, o ile nie mamy do czynienia z pedomorfozą (Gould 1977). Podobnie znajomość polaryzacji cechy w blisko spokrewnionych grupach może – choć nie musi – wskazywać na polaryzację w grupie badanej. Nie jest natomiast prawdą, że znajomość polaryzacji cech jest warunkiem podjęcia rekonstrukcji filogenezy: rekonstrukcję możemy przeprowadzać, nie znając polaryzacji cech, a ustalić ją właśnie na podstawie zrekonstruowanej filogenezy, a co za tym idzie – zrekonstruowanej ewolucji cech. Wrócimy do tego jeszcze.

Oprócz polaryzacji znać też musimy możliwe **transformacje**, czyli przejścia jednego stanu w drugi. Wyróżniamy więc cechy **nieodwracalne** (*irreversible*): możliwe jest przejście ze stanu a w stan b , a następnie w c ($a \rightarrow b \rightarrow c$), natomiast przejście odwrotne ($c \rightarrow b \rightarrow a$) nie jest możliwe. Nie należy tego rodzaju cech mylić z **cechami typu Dollo**, choć mówi się o „prawie Dollo o nieodwracalności ewolucji”. Dla cech typu Dollo ewolucja jest też nieodwracalna, ale w inny sposób: niemożliwy jest powrót do stanu zaawansowanego po jego utraceniu. Czyli możliwe jest przejście: a (stan pierwotny) $\rightarrow b$ (stan zaawansowany), jak też $b \rightarrow a$, ale następnie ponowne przejście $a \rightarrow b$ możliwe nie jest, a jedynie $a \rightarrow c$, przy czym $c \neq b$, czyli ponownie uzyskany stan zaawansowany – choćby był nieodróżnialny od wcześniej utraconego – nie jest z nim identyczny, czyli nie jest homologiczny. Taki rodzaj dopuszczalnej transformacji cechy wygląda może dziwacznie i nierzeczywiście, jednak okazuje się użyteczny, zwłaszcza dla niektórych cech molekularnych. Kolejny rodzaj cech to cechy **odwracalne, uporządkowane** (*reversible, ordered*). Oznacza to, że istnieje określony porządek przejść pomiędzy stanami, możliwych wprawdzie w obu kierunkach, lecz z zachowaniem określonej kolejności stanów cechy. Czyli możliwe są przejścia: $a \leftrightarrow b \leftrightarrow c \leftrightarrow d$, natomiast niemożliwe są bezpośrednie przejścia: $a \leftrightarrow c$, $a \leftrightarrow d$ czy $b \leftrightarrow d$. Ostatnim rodzajem cech są cechy **odwracalne, nieuporządkowane** (*reversible, unordered*). W tym przypadku możliwe jest bezpośrednie przejście, w obu kierunkach, pomiędzy jakikolwiek dwoma stanami, czyli $a \leftrightarrow b \leftrightarrow c \leftrightarrow d$, ale tak samo i $a \leftrightarrow c$, $a \leftrightarrow d$ czy $b \leftrightarrow d$.

Wskazanie możliwych zmian stanów cechy to inaczej określenie **serii transformacyjnej** stanów tej cechy. Termin ten jest o tyle istotny, że w praktyce niejednokrotnie samo określenie cechy jako nieodwracalna, typu Dollo, odwracalna uporządkowana czy odwracalna nieuporządkowana nie będzie odpowiadało rzeczywistej ewolucji cechy. Cecha może być np. nieodwracalna przy przejściu ze stanu a w stan b , a odwracalna przy przejściu ze stanu b do stanu c : $a \rightarrow b \leftrightarrow c$. Albo też przejście ze stanu a do d

musi się odbyć za pośrednictwem stanu *b*, podczas gdy stan *c* może być pominięty lub nie. Nieraz dla opisu serii transformacyjnej musimy rysować **drzewo stanu cech**, gdy np. $a \leftrightarrow b$, ale następnie $b \leftrightarrow c$ lub $b \leftrightarrow d$ (Ryc. 2.1a), albo seria jest jeszcze bardziej złożona (Ryc. 2.1b).



Ryc. 2.1. Przykłady dwóch hipotetycznych drzew stanów cech. Ze stanu *a* możliwe jest przejście w *c* lub *d*, wszystkie przejścia są odwracalne (a). Znacznie bardziej skomplikowany schemat przejść, przy czym nie wszystkie z nich są odwracalne (b)

To oczywiście, że znajomość serii transformacyjnej jest niezbędna dla prawidłowej rekonstrukcji ewolucji. Zazwyczaj jednak niewiele wiemy o ewolucji cech, a dokładniej opisać tę ewolucję możemy dopiero po zrekonstruowaniu filogenezy, do czego z kolei niezbędna jest znajomość ewolucji cech, a więc błędne koło się zamyka. W dalszej części tej książki (Rozdział 4.5) powrócimy do rekonstrukcji ewolucji cech, wskazując, jak zminimalizować niebezpieczeństwo zastosowania logiki błędnego koła. Tutaj poprzestańmy na podkreśleniu, że dla przyjęcia określonej serii transformacyjnej staramy się wykorzystać wszystko, co wiemy o ewolucji, pokrewieństwach, biologii itd. badanej grupy, starając się przyjmować możliwie najprostsze serie transformacyjne, o ile nie mamy konkretnych podstaw do założenia bardziej skomplikowanej ewolucji. Kolejnym problemem są **wagi cech**, czyli ich znaczenie w rekonstruowanej ewolucji. Trudno przecież pojawienie się płuca uznać za tak samo ważne jak np. powstanie barwnej pigmentacji ciała. Najogólniej mówiąc, im trudniejsza do nastąpienia, a więc im rzadsza była zmiana, tym wyższą wagę powinna mieć dla rekonstrukcji, i na odwrót. I tutaj istnieją techniki **ważenia**, którymi zajmiemy się dalej (Rozdział 4.5). I one jednak nie są nigdy zupełnie wolne od logiki błędnego koła. Wydawać by się mogło, że najobiektywniej byłoby wszystkim cechom dać tę samą wagę, tak też się postępuje przy analizach fenetycznych. Dla taksonomii filogenetycznej czy ewolucyjnej takie podejście jest jednak nie do przyjęcia. Już zwykły zdrowy rozsądek każe nam różnicować wagi, choć oczywiście mając do tego biologiczne przesłanki.

2.4. Kodowanie cech ilościowych ciągłych jako dyskretnych

Jak już wspominaliśmy, zastosowanie cech ilościowych ciągłych w analizie kladystycznej budzi szereg zastrzeżeń i wątpliwości, skłaniając wielu kladystów do zupełnego ich odrzucania, jako bezwartościowych (Kitching i inni 1998). To jednak stanowisko skrajne: niepotrzebne i niezbyt sensowne byłoby korzystanie z cech ilościowych, czyli w praktyce danych morfometrycznych do badania pokrewieństw na poziomie typu, gromady, rzędu czy nawet rodziny. Tam różnic jest dosyć. Dla odmiany jednak rekonstrukcja filogenezy pomiędzy bliskimi sobie gatunkami zwykle nie może się obejść bez cech morfometrycznych, bowiem często są to niemal jedyne różnice, jakie dają się stwierdzić. Tymczasem stosowane w kladystyce techniki i programy komputerowe niemal zawsze wymagają cech jakościowych lub ilościowych dyskretnych, kodowanych jako zmienne wielostanowe dyskretnie.

Istnieją metody filogenetyczne, analizujące cechy ciągłe (np. Farris 1970, Rogers 1984, Huey i Bennett 1987, Swofford i Berlocher 1987, Swofford i Maddison 1987), przydatne zwłaszcza dla częstości alleli, ale także dla innego rodzaju cech ciągłych, jak w programie MacClade (Maddison i Maddison 1992). Opierają się na minimalizacji sumy wartości absolutnych różnic (metoda liniowa: *linear*, *Manhattan* lub *Wagner parsimony*) albo też minimalizacji sumy kwadratów różnic (*squared-change parsimony*) pomiędzy wartościami cech, danymi lub obliczonymi na końcach gałęzi drzewa. Choć zasada jest prosta, to obliczenia są czasochłonne. Metod tych zresztą nie stosuje się zbyt często, tym bardziej że wiele cech wykorzystywanych w kladystyce to zmienne dyskretnie i pożądanym jest łączne użycie wszystkich cech jednocześnie.

W tej sytuacji pozostaje więc **przekodowanie** (Archie 1985, Baum 1988, Felsenstein 1988, Goldman 1988, Chappill 1989, Falniowski i inni 1997). Dla cech ilościowych ciągłych bywa regułą zachodzenie na siebie zakresów zmienności dla poszczególnych taksonów, a zdarza się to i dla cech jakościowych oraz ilościowych dyskretnych. Oczywiście najlepsze cechy to takie, dla których zakresy zmienności są rozłączne, a najgorsze – o niemal identycznych zakresach zmienności. Nie oznacza to jednak, że cechy możemy z tego punktu widzenia podzielić na „złe” i „dobre” – mamy raczej ciągłą skalę dobroci cech, odpowiadającą stopniowi rozłączności zakresów zmienności. Skala taka jednakże klasyfikuje „jakość” cech jedynie formalnie – można przecież bez trudu wskazać cechy o całkowicie rozłącznych przedziałach zmienności dla poszczególnych taksonów, a pomimo to niezbyt przydatne, jako pozbawione głębszego filogenetycznego znaczenia, a z drugiej strony cechy o zakresach zmienności silnie zachodzących, lecz ważne ewolucyjnie (np. odzwierciedlające ewolucyjny trend redukcji albo zwiększania proporcjonalnych wymiarów jakiejś struktury). Warto też nie zapominać o różnicy między cechami **diagnostycznymi** – charakteryzującymi się stanami łatwo wyróżnialnymi i niezachodzeniem zakresów zmienności pomiędzy taksonami, lecz niekoniecznie odzwierciedlającymi pokrewieństwa – a synapomorfiami, nawet jeżeli ich stany wyróżnia się z trudnością.

Cechę ilościową ciągłą zazwyczaj przekodowuje się w dyskretną wielostanową, uporządkowaną i odwracalną (*ordered*, *reversible*), tym samym zakładając określoną serię transformacyjną, a więc i określony model zmiany stanu cechy. Logicznie wydaje się to uzasadnione, choć odwracalność zmiany budzi wątpliwości, bowiem np. w następstwie stopniowej redukcji struktura może być utracona nieodwracalnie. Ponadto średnie rozmiary niekoniecznie muszą w trakcie ewolucji być przejściem pomię-

dzy dużymi i małymi, choć najczęściej są. W każdym razie warto każdorazowo rozpa-
 trzyć, co wiemy o zmienności, ontogenezie, dziedziczeniu i ewolucji danej cechy. Ko-
 lejną sprawą to liczba stanów, jakie wyróżnimy. Powinno ich być mniej niż taksonów,
 których filogenezę rekonstruujemy z wykorzystaniem tej cechy. Ponadto programy
 komputerowe mają ograniczenia liczby stanów cechy – od 10 po ponad 30, zależnie od
 programu: liczb tych nie możemy przekroczyć. Mówiliśmy już, że kontrowersyjne jest
 przypisywanie zmianie jakiejś proporcji takiej samej rangi jak utracie lub uzyskaniu
 np. torebki kopolacyjnej w obrębie narządów rozrodczych – musimy stosownie ważyć
 cechy. I to jednak nie wystarczy, gdy występowanie/utratę tejże torebki kodujemy bi-
 namiem: 0/1, a pełen zakres zmienności tejże proporcji pomiędzy badanymi taksonami
 zawiera się między stanami 0 a 9. Wówczas pełna różnica będzie wydłużała drzewo
 (przy metodzie kladystycznej – *parsimony*) dziesięciokrotnie bardziej niż utrata albo
 uzyskanie torebki. Aby cechy ważyły podobnie, musimy każdorazowo liczbę kroków
 dla przekodowanej cechy ciągłej „normalizować”, dzieląc ją przez pełną liczbę kro-
 ków, czyli w tym wypadku przez 10.

Bywa, że zmienne ciągle wykazują nieciągłości pomiędzy przynajmniej niektórymi
 taksonami i wówczas przypisanie określonym wartościom stanu cechy odpowiadają-
 cych im kodów zmiennej dyskretnej wielostanowej jest oczywiste. Najczęściej jednak
 takich nieciągłości brak i wówczas musimy je wyznaczyć. Nie powinno się to odbywać
 mechanicznie, po prostu dzieląc znany, pełny zakres zmienności cechy na odpowiednią
 liczbę przedziałów. W jednym z dalszych rozdziałów zajmiemy się wstępną analizą
 cech ilościowych ciągłych, dla których nie wystarcza sama znajomość zakresu – należy
 też obliczyć średnią, medianę i wariancję. Powinno się też, wykorzystując analizę wa-
 riancji (ANOVA), sprawdzić, czy różnice wartości cech pomiędzy taksonami są staty-
 stycznie istotne. Następnie dopiero przystępujemy do znalezienia nieciągłości. Metod
 jest wiele (Mickey i Johnson 1976: *simple gap-coding*; Colless 1980: *segment cod-
 ing*; Thorpe 1984, Almeida i Bisby 1984: *divergence coding*; Archie 1985, Goldman
 1988: *generalized gap-coding*; Baum 1988: *range coding*; Thiele 1993: *gap-
 weighting*). Wszystkie polegają na zastosowaniu prostego algorytmu dla uzyskania
 nieciągłości, by móc kolejnym odcinkom wyjściowo ciągłej lub zachodzącej zmienno-
 ści, oddzielonym od siebie tymi nieciągłościami, przypisać kolejne kody, jak dla kolej-
 nych stanów cechy dyskretnej. *Simple gap-coding* (Mickey i Johnson 1976) dzieli
 zakres zmienności w miejscach, gdzie nie ma wartości, a jeżeli takich miejsc brak, to
 w arbitralnie przyjętej odległości (zwykle równej odchyleniu standardowemu albo
 dwóm odchyleniom standardowym) od wartości średniej. Metodę tę łatwo zastosować,
 dysponując jedynie kalkulatorem; często metoda ta wystarcza.

W praktyce niejednokrotnie okaże się, że nieciągłości – czy to rzeczywiste, czy
 mierzone odległością między średnimi – będą różnej wielkości pomiędzy różnymi sta-
 nami przekodowanej cechy, a już intuicyjnie oczywiste wydaje się przypisanie więk-
 szego kosztu zmianie większej. Takie kodowanie – uwzględniające nie tylko kolejność,
 ale i wielkość **nieciągłości** (*gaps*) – zakłada **ważenie nieciągłości** *gap-weighting*
 (Thiele 1993). Dane wyjściowe najpierw szeregujemy jako uporządkowany zestaw
 stanów, którymi są zwykle **średnie** lub **mediany** dla kolejnych taksonów. Następnie je
transformujemy i **normalizujemy** (patrz Rozdział 2.7). Tak przekształcone wartości
 zaokrąglamy do liczb całkowitych i wykorzystujemy jako stany cechy dyskretnej.
 Wówczas może się zdarzyć, że w uzyskanej macierzy będą stany: 0, 1, 2, 5, 7 i 9, a nie
 będzie 3, 4, 6 i 8, bowiem tam właśnie nieciągłości były większe.

2.5. Kodowanie cech jakościowych i ilościowych dyskretnych, brakujące dane, polimorfizm

Dla cech jakościowych i ilościowych dyskretnych opisane wyżej procedury nie są potrzebne, co jednak nie oznacza, że kodowanie takich cech jest rzeczą prostą, tym bardziej że sposób kodowania wpływa znacząco na wyniki rekonstrukcji filogenezy. Załóżmy, że u pięciu gatunków: A, B, C, D i E, mamy następujące stany cech:

	A	B	C	D	E
występowanie łusek	brak	obecne	obecne	obecne	obecne
kształt łusek		kolisty	kolisty	trójkątny	trójkątny
pigmentacja łusek		czarna	brak	czarna	brak

Możemy uznać, że najlepiej powyższe różnice opisać za pomocą jednej, wielostanowej cechy: brak (0), koliste czarno pigmentowane (1), koliste pozbawione pigmentu (2), trójkątne czarno pigmentowane (3) i trójkątne pozbawione pigmentu (4) [metoda I]. Można też przyjąć, że kształt i obecność pigmentu to dwie niezależne cechy: 1. kształt łusek: brak (0), kolisty (1), trójkątny (2); 2. pigmentacja łusek: brak łusek (0), czarna (1), brak pigmentu na łuskach (2) [metoda II]. Inna możliwość to potraktowanie kształtu i pigmentacji jako dwóch niezależnych cech, a obecności łusek jako trzeciej niezależnej: 1. łuski: brak (0), obecne (1); 2. kształt łusek: kolisty (0), trójkątny (1); 3. pigment na łuskach: brak (0), obecny (1) [metoda III]. Kolejna metoda to kodowanie niezależne zmiennych binarnych, zakładające brak transformacji: 1. łusek brak (0), łuski obecne (1); 2. kształtu kolistego łusek brak (0), kształt kolisty łusek obecny (1); 3. kształtu trójkątnego łusek brak (0), kształt trójkątny łusek obecny; 4. czarny pigment na łuskach: nieobecny (0), obecny (1); 5. pozbawione pigmentu łuski: nieobecne (0), obecne (1). Tak więc kodowanie dla poszczególnych gatunków byłoby następujące:

Gatunek	metoda kodowania			
	I	II	III	IV
A	0	00	0--	00000
B	1	11	101	11010
C	2	12	100	11001
D	3	21	111	10110
E	4	22	110	10101

Przedstawione powyżej metody nie wyczerpują bynajmniej wszystkich możliwości, jednak wystarczająco dobrze ilustrują, jak różnie można przeprowadzić kodowanie. Już sama liczba cech – od jednej do pięciu – nie mówiąc o związanych z tym różnicach określenia możliwych serii transformacyjnych, nie pozwala nam wątpić, że obraz zrekonstruowanej ewolucji będzie niejednokrotnie zupełnie inny, zależnie od przyjętego sposobu kodowania. Metoda I zakłada zależność wszystkich stanów wszystkich cech, traktując je jako stany jednej cechy, metoda II uznaje kształt i pigmentację za cechy niezależne, w III odrębną cechą jest obecność lub brak łusek, wreszcie metoda IV przyjmuje niezależność występowania każdego z pięciu stanów, traktowanych jako odrębne cechy binarne. Zachodzi pytanie: którą metodę wybrać w danym przypadku? Jak w całej rekonstrukcji filogenezy – a kodowanie to właśnie jeden z bardziej czułych, błędotrwałych etapów – metod *foolproof*, czyli odpornych na inteligentnych inaczej,

brak. Nie istnieje jakaś „czarna skrzynka”, do której wrzucamy dane, by z drugiej strony wypadło gotowe drzewo, bezbłędnie przedstawiające ewolucję w obrębie badanej grupy. Często więc będziemy musieli próbować różnego kodowania i porównywać uzyskane rekonstrukcje filogenezy oraz ewolucji cech, zanim wybierzemy ten najwłaściwszy, naszym zdaniem i w danym przypadku. Można natomiast sformułować szereg wskazówek, którymi kierujemy się podczas kodowania.

W analizie kladystycznej powszechnie przyjmuje się, że każdą z cech traktować należy jako niezależną hipotezę o pokrewieństwach, testowaną rozkładem stanów cechy na zrekonstruowanym kladogramie. Gdy brak niezależności, metoda staje się mniej czuła, a rosnąca liczba cech sprzężonych (*linked*) zwiększa ryzyko, że jedna błędnie rozpoznana homologia „zamaskuje” topologie właściwe, wynikające z prawidłowo rozpoznanych homologii. Synapomorfie, czyli przypadki, gdzie podobne stany cech uznano słusznie za wynik homologii, powinny wystąpić jedynie w wycinkach kladogramu, spełniających warunki monofiletizmu; różne synapomorfie powinny zgodnie potwierdzać jeden (lub niewiele) kladogram najlepszy pod względem przyjętego kryterium optymalizacji: „najlepsze” drzewo dla cechy, której homologię chcemy sprawdzić, powinno mieć topologię identyczną jak „najlepsze” dla pozostałych cech o sprawdzonych homologiach, czyli dołożenie do analizy rozpatrywanej cechy nie może pogarszać wartości stosowanego kryterium optymalizacji. Homoplazje natomiast powinny być rozmieszczone nieregularnie na zrekonstruowanym drzewie. Tak więc, jak już pisaliśmy, homologie testuje się ostatecznie zrekonstruowaną filogenezą. Takie testowanie może być jednak niemożliwe przy niektórych sposobach kodowania. Metoda I wyklucza niezależne testowanie homologii kształtu łusek i ich pigmentacji, w metodzie II i III z konieczności zawarte są niesprawdzalne testem zgodności założenia serii transformacyjnych i dane *a priori* homologie, metoda IV natomiast pozwala na badanie każdego stanu cech z osobna – pomija jednakże logiczne związki pomiędzy stanami cech i ignoruje wstępne rozpoznanie homologii na podstawie podobieństwa, może więc prowadzić do dziwacznych rekonstrukcji, tylko formalnie najlepszych pod względem wartości kryterium optymalizacji (Meier 1994, Kitching i inni 1998).

Część teoretyków kladyzmu uznaje metodę IV, czyli kodowanie brak/obecny, za metodę najlepszą, czy wręcz jedynie słuszną, pomimo powyższych zastrzeżeń, jak też **redundancji** (rozwlekłości) kodowania, obciążonego dodatkowymi symbolami nieobecności. Uważają tak, gdyż metoda ta nie czyni nieuprawnionych założeń *a priori* o ewolucji cech, choć skądinąd pomija też i uprawnione założenia, odzwierciedlające naszą wiedzę na temat ewolucji, ontogenezy, itp. Metodę tę rzadko spotkamy w literaturze, zapewne głównie w wyniku trudności z rozwiązaniem problemów wynikających ze sprzężeń pomiędzy cechami. Ryzyko wynikające ze sprzężeń z cechą niehomologiczną minimalizuje metoda I, za to wymaga przyjęcia założeń o seriach transformacyjnych, czyli ewolucji cech. Jedną z prób rozwiązania problemu jest **analiza serii transformacyjnych** (*transformation series analysis* – TSA: Micklewich 1982), gdzie ewolucja cech badana jest niezależnie od rekonstrukcji filogenezy. W metodzie II i III dodatkowym problemem jest podział: czy kształt łusek odzwierciedla podobieństwa niezależnie od obecności pigmentu na łuskach?

Prawidłowe kodowanie powinno zawierać całą dostępną informację, a zarazem minimalizować nieuprawnione hipotezy *a priori* o ewolucji cech, homologiach, niezależności cech. Pamiętajmy też, że homologia określona dla jednego poziomu uniwersalności nie musi być homologią dla innego poziomu – jest więc oczywiste, że kodowanie

odpowiednie dla rekonstrukcji filogenezy w obrębie np. rodziny może być nieodpowiednie dla analizy pokrewieństw pomiędzy rodzinami. Niewątpliwie powinniśmy dobrać kodowanie tak, aby precyzyjnie odzwierciedlało ewolucję cech tam, gdzie nasza wiedza na to pozwala. Dla pozostałych cech należy w miarę możliwości unikać kodowania, które utrudniałoby późniejsze testowanie homologii. Pimentel i Riggins (1987) uważają, że w analizie kladystycznej nie można traktować stanów cech jako prostych zmiennych nominalnych, bowiem traci się wówczas część informacji. Zamiast tego proponują traktowanie cech jako stanów wielostanowej cechy, kodowanych jako **addytywne binarne**, aby możliwe było odróżnienie liniowych i rozgałęzionych serii transformacyjnych stanów cech. **Kodowanie addytywne** polega na zapisie n -stanowej cechy wielostanowej jako $n - 1$ cech binarnych. Przykładowo: brak kolca (000), kolec mały (100), kolec średni (110), kolec duży (111). Jak widać, tu też wadą jest redundancja: gdy pierwszą liczbą jest 0, to wiemy, że i następne dwie też będą 0; gdy ostatnią jest 1, to dwie pierwsze też muszą być 1. Kodowanie takie warto jednak zrobić dla wstępnej analizy, aby rozpoznać homologie i serie transformacyjne; później możemy zastąpić je innym, odzwierciedlającym wiedzę nabytą w trakcie wstępnej analizy. Kodowanie addytywne może być też użyteczne, gdy dysponujemy starszymi programami, wymagającymi wyłącznie cech binarnych.

Odrębnego omówienia wymagają **brakujące dane** i **polimorfizm** (Nixon i Davis 1991, Platnick i inni 1991, Maddison i Maddison 1992). Choć większość programów komputerowych umożliwia wówczas kodowanie „?” , a niektóre mają specjalne kody (np. „0&1”) dla polimorfizmu, bynajmniej nie rozwiązuje to problemu. Jeżeli nie wiemy, czy u danego taksonu i dla danej cechy występuje stan 0, czy stan 1, to jest to wynikiem: (1) niekompletnych informacji, (2) niedotyczenia tej cechy tego taksonu, (3) polimorfizmu, czyli występowania obu stanów cechy. Którykolwiek z tych przypadków ma miejsce, wynikiem jest wzrost liczby tak samo uprawnionych, „najlepszych” rekonstrukcji filogenezy, jak też zmniejszenie się rozdzielczości rekonstrukcji, wyrażające się wzrostem liczby nierozwiązanych politomii na kladogramie. To jednak nie wszystko. Jeżeli takson jest polimorficzny pod względem więcej niż jednej cechy, to może, choć nie musi, dojść do wzajemnych oddziaływań pomiędzy cechami, których rezultatem jest znajdowanie „najlepszych” rekonstrukcji filogenezy, dla których niemożliwe jest zrekonstruowanie ewolucji cech w sposób sensowny. Bywa też, że wynikiem przyjęcia polimorfizmu taksonów terminalnych jest wyższa liczba jednakowo uprawnionych „najlepszych” rekonstrukcji niż dla wszystkich możliwych kombinacji stanów tych cech łącznie.

Odsyłając do cytowanych wyżej publikacji, poprzestańmy na konkluzji, że polimorfizmu i brakujących danych powinniśmy unikać – inaczej zawsze liczyć się trzeba ze zmniejszeniem wiarygodności rekonstrukcji. Oczywiście zdarzyć się może, że jakichś danych rzeczywiście brak, a jest to zwłaszcza częste w paleontologii. Możemy wówczas wyłączyć dany takson z analizy, lecz nie jest to właściwe. Zazwyczaj takie pominięcie nie oznacza jedynie rezygnacji ze znajomości położenia tego właśnie taksonu w obrębie filogenezy, lecz wpływa też na pozostałą część rekonstrukcji, bowiem zestaw cech tego taksonu również wnosi informacje, dotyczące ewolucji cech w badanej grupie. Stąd nie warto rezygnować z taksonu jedynie dlatego, że stanu jednej cechy nie znamy. Warto jednakże dołożyć starań, by takich przypadków było jak najmniej. Unikać natomiast należy brakujących stanów cech, będących następstwem niedotyczenia danej cechy danego organizmu (choćby cechy kończyn u węży, piór u krokodyli, łusek

u płazów). Po prostu tak dobieramy cechy i sposób ich kodowania, by tego uniknąć – z tego punktu widzenia nie powinniśmy korzystać z metody III. Konieczność unikania brakujących danych ma jednak poważniejsze następstwa: właściwie nie można rekonstruować filogenezy taksonów, dla których dysponujemy różnymi zestawami danych. Symulacje wykazały na szczęście, że brakujące dane słabiej wpływają na rekonstrukcję filogenezy, gdy skoncentrowane są w pewnych częściach macierzy (Rozdział 4.10). Jeżeli przykładowo mamy dane morfologiczne dla wszystkich gatunków, a dodatkowo molekularne dla części, to wówczas wszystkie dane molekularne musiałyby być kodowane jako niedotyczące dla tych taksonów, dla których ich nie badano. Podobnie będzie np. przy jednoczesnym uwzględnianiu taksonów współczesnych i kopalnych – dla tych ostatnich zestaw danych jest daleko mniejszy. Pozostaje polimorfizm. Najlepiej go obejść poprzez kodowanie polimorfizmu jako dodatkowego stanu cechy albo uznanie jego stanów za cechy odrębne. Jeżeli więc oczy mogą być zielone lub niebieskie, to możliwe jest kodowanie cechy barwa oczu: zielona (0), zielona lub niebieska (1), niebieska (2) albo kodowanie dwóch cech: oczy zielone: brak (0), obecne (1); oczy niebieskie: brak (0), obecne (1).

2.6. Wstępna analiza cech ilościowych; ujęcie opisowe i stochastyczne

Pomimo wszystkich zastrzeżeń wyrażonych wcześniej, cechy ilościowe ciągle zmuszeni jesteśmy wykorzystywać w analizie filogenetycznej, a analiza fenetyczna opiera się głównie na nich. O ile cechy jakościowe, choć bywają polimorficzne, dają się łatwo określić – albo są łuski, albo ich brak, oczy albo są zielone, albo niebieskie – to cechy ilościowe wykazują zawsze zmienność: zarówno wewnątrz taksonów, jak i pomiędzy nimi. Najogólniej rzecz biorąc, zmienność między taksonami jest podstawą taksonomii, podczas gdy zmienność wewnątrz taksonów to zhora nekająca taksonomów. Nawet powtarzane pomiary tej samej wielkości dadzą różne wyniki, w następstwie błędów pomiaru. Właściwie każde naukowe badanie danych, dotyczących najszerzej rozumianej zmienności naturalnej, nazwać można statystyką, i w tym sensie wszystkie metody opisane w tej książce uznać by można za statystyczne. Zwykle jednak pojęcie statystyki rozumiemy wężiej, jako nauki powstałej z połączenia coraz to bardziej skomplikowanych metod opisu danych (wyjściowo dotyczących państwa – stąd nazwa) z teorią rachunku prawdopodobieństwa, na początku zajmującej się gramami. Stopniowo mechanistyczny, przyczynowo-skutkowy model biologii ustąpił w wielu jej dziedzinach modelowi statystyczno-probabilistycznemu. Statystykę określić możemy jako teorię i praktykę podejmowania rozsądnych decyzji w warunkach niepewności (Rao 1994). Używamy jej, gdy badana rzeczywistość jest wynikiem oddziaływania czynników zbyt wielu bądź za mało znanych, aby opis przyczynowo-skutkowy był możliwy. Pozostawiamy otwarte pytanie, czy otaczający nas świat jako taki ma charakter przyczynowo-skutkowy czy probabilistyczny, choć warto podkreślić, że dotychczasowego przebiegu ewolucji na Ziemi nie da się wytłumaczyć bez oddziaływania w wielu momentach czynników losowych, a gdyby doszło do jej powtórzenia, musiałaby ona mieć nie taki sam przebieg. Opisane w dalszej części metody rekonstrukcji filogenezy najczęściej do tak rozumianej statystyki nie należą – np. metoda kladystyczna dąży do odtworzenia jednego, tego właściwego przebiegu filogenezy, choć rekonstrukcje do pewnego stop-

nia wartościuje się metodami statystycznymi. Z drugiej strony takie metody jak maksymalizacji wiarygodności (*maximum likelihood*) należą do statystyki.

Istnieje wiele dobrych podręczników statystyki jednowymiarowej, także adresowanych do biologów (Sokal i Rohlf 1987, 1995, Kachigan 1991, Łomnicki 1995), tu więc całkowicie te zagadnienia pominiemy, poprzestając na niewielu uwagach praktycznych. Zacząć trzeba jednakże od ważnego rozgraniczenia, zaznaczającego się już na etapie zbierania danych. Zależnie od charakteru badanego zagadnienia i sposobu uzyskiwania danych odpowiednie będzie podejście **stochastyczne** lub **opisowe**. W podejściu stochastycznym zakłada się, że badany zbiór obiektów jest **losowo** pobraną **próbą** ze znacznie większego (najlepiej nieskończonego, a przynajmniej bardzo licznego) zbioru obiektów, zwanego populacją (w rozumieniu statystyki, choć niekiedy będzie to tożsame z populacją biologiczną). Zmienne są tu **zmiennymi losowymi**. Wówczas możemy założyć, że badany przez nas zbiór obserwacji odzwierciedla właściwości całej populacji. W podejściu opisowym o losowości próby nie mówi się nic, a rozpatrywane zmienne nie są zmiennymi losowymi, nie interesują nas właściwości stochastyczne zbioru obserwacji. Analizujemy zbiór posiadanych danych, nie wchodząc w problemy ich pozyskiwania. Interesują nas podstawowe charakterystyki opisujące badany zbiór.

Najbardziej klasycznym przykładem danych nadających się do zastosowania ujęcia stochastycznego są wyniki doświadczeń, które można (teoretycznie) powtarzać dowolną liczbę razy. Ujęcie stochastyczne jest też przydatne, gdy chcemy scharakteryzować np. cały gatunek albo jakąś jego populację. Dla odmiany, charakteryzując drzewa rosnące w danym parku, stosować należy ujęcie opisowe. Ujęcia opisowego – chcąc nie chcąc – używać musimy też zawsze, gdy zbioru obiektów badanych nie pobrano losowo (patrz Rozdział 2.8). Niestety więc podejście stochastyczne z reguły nie będzie możliwe, gdy pracować będziemy na zbiorach muzealnych, materiałach złowionych przez kolekcjonerów, itp. Niekiedy zresztą podejście stochastyczne nie ma sensu: jeżeli np. badamy cały garnitur ($n = 24$) zębów w czaszce jakiegoś ssaka, które wszystkie pomierzyliśmy, to np. średnia obliczona dla obserwacji charakteryzuje też cały zbiór. Pytanie o ujęcie warto postawić na samym początku. Choć dla taksonoma najbardziej pożądane byłoby stosowanie podejścia stochastycznego w każdym przypadku, to jednak bardzo często już charakter posiadanych (bądź nawet osiągalnych) danych takie podejście wykluczy. W czasach komputerów niezwykle kuszące jest wybranie w pakiecie statystycznym jeszcze kilku opcji, aby wyliczyć szereg współczynników, tyle że najczęściej uczynimy to w sposób nieuprawniony, a uzyskane parametry nie będą odzwierciedlały (niemal) niczego. Metody stochastyczne są też niezmiernie wrażliwe na **obserwacje nietypowe**, a szerzej – na zachowanie **jednorodności i normalności (eliptyczności)** rozkładu. Powrócimy do tego jeszcze, tutaj poprzestając na podkreśleniu, że podejście opisowe stosować możemy zawsze i jest ono znacznie bardziej **odporne (robust)** na obserwacje nietypowe i niespełnianie szeregu warunków danej techniki.

Najczęstszym błędem przy stosowaniu rozmaitych metod statystycznych, czy ogólnej numerycznych, jest niebaczenie na **warunki**, jakie muszą być spełnione, aby dana technika działała prawidłowo. W dalszej części przy każdej metodzie będziemy się starali określić każdorazowo warunki jej użycia. Tu warto wstępnie stwierdzić, że różne metody mają różne **założenia wstępne** i są bardziej lub mniej wrażliwe na niespełnianie tych założeń. Oczywiście nie ma w praktyce ani danych doskonale niespełniających założeń, ani też doskonale je spełniających. Spełnianie lub nie założeń jest więc względne: teoretycznie istnieje cały zakres możliwości, od doskonałego spełniania po

doskonale niespełnianie. W miarę jak założenia techniki są coraz bardziej naruszane, maleje czułość metody, wzrasta możliwość błędu. Oczywiście może się zdarzyć, że nawet przy bardzo zdecydowanych odchyleniach od spełniania warunków techniki wyniki będą prawidłowe, gdy po prostu kolejne odchylenia wzajemnie się zniosą, jednak trudno na to liczyć i bezpieczniej założyć, że odchylenia się zsumują.

Dla zmiennych ilościowych ciągłych istotna jest **dokładność** wykonywania pomiarów. Oczywiście nie może być za mała, ale nie ma też potrzeby dokonywania pomiarów z dokładnością najwyższą, jaką zapewnia posiadany sprzęt. I same pomiary będą trudniejsze, i obliczenia bardziej uciążliwe, pomimo łatwości operowania przez współczesne komputery zmiennymi o dużej liczbie miejsc dziesiętnych. Zwykle już kilka początkowych pomiarów pozwala ustalić zakres zmienności danej zmiennej. Dobrą praktyką jest, aby dokładność pomiarów zawierała się w przedziale $<1/300, 1/30>$ tego zakresu. Jeżeli więc wysokość muszli ślimaka mieści się w przedziale (10, 30) mm, to dokładność pomiarów powinna wynosić 0,1 mm. Dla odmiany, jeżeli pęd rośliny mierzy 0,7–1,2 m wysokości, to wystarczy dokładność 1 cm, najwyżej 1 mm. Kolejna sprawa to **wielkość próby**, czyli liczba zmierzonych osobników. W ujęciu opisowym mierzymy albo wszystkie osobniki, albo pewną ich liczbę. W ujęciu stochastycznym mierzymy fragment populacji, który musi być **reprezentatywny** dla całej populacji. Z pozoru im próba większa, tym bardziej reprezentatywna, jednak niezupełnie tak jest. Aby próba była reprezentatywna, musi być **pobrana losowo**, w taki sposób, że dla każdego z osobników całej populacji (statystycznej) zachodzi takie samo prawdopodobieństwo znalezienia się w próbie mierzonej. Jeżeli warunek losowości jest spełniony, to symulacje komputerowe wykazały, że zwykle wystarcza liczebność $n = 30$ osobników, aby próba była reprezentatywna. Nie zwiększy natomiast reprezentatywności zwiększanie próby, pobranej choćby tylko niezupełnie losowo, choć niewątpliwie zwiększy nakład zbędnej pracy.

Zmierzone wartości poddajemy wstępnej obróbce przy zastosowaniu **statystyki opisowej**. Należy więc, obok **liczebności próby**, podać **granice** i **zakres** mierzonej wielkości (np. 1–5 mm; 4 mm), **wartość średnią**, **wariancję** lub **odchylenie standardowe** (w obu przypadkach pamiętając o różnicy pomiędzy ujęciem stochastycznym i opisowym, czyli mianowniku równym odpowiednio $n - 1$ bądź n), może też **współczynnik zmienności**. Warto sprawdzić istotność różnic wartości danej cechy pomiędzy badanymi taksonami, wykorzystując **analizę wariancji** (ANOVA), pamiętając jednakże o zbadaniu wcześniej, czy spełnione są warunki tego testu, jak **normalność rozkładów**, **jednorodność** grup, itd. Analiza wariancji jest szeroko omawiana w wielu podręcznikach (Sokal i Rohlf 1987, 1995, Kachigan 1991, Łomnicki 1995). Niestety w literaturze znaleźć można liczne przykłady zastosowania tej techniki dla danych zupełnie niespełniających warunków jej użycia. ANOVA niemal z reguły powinna poprzedzać wielowymiarowe analizy fenetyczne. Jeżeli natomiast zmienne ciągle wykorzystywać zamierzamy do analiz filogenetycznych, analizy wariancji na ogół nie przeprowadzamy. Najczęściej korzystamy po prostu z wartości średnich. W praktyce jednak bardzo często wyniki pomiarów nie będą tworzyć dla poszczególnych taksonów grup jednorodnych, wśród obserwacji znajdują się **obserwacje nietypowe** (*outliers*). W takich sytuacjach korzystne jest wykorzystywanie technik **odpornych** (*robust*) na obserwacje nietypowe. Średnia wartością odporną nie jest: dla próby o liczebności n wystarczy $1/n$, czyli jedna wartość nietypowa, aby średnia „załamała się”, czyli przestała być reprezentatywna dla opisywanego zbioru. W przeciwieństwie do średniej, odporna jest me-

diana: jej współczynnik załamania wynosi 0,5, czyli potrzeba aż 50% obserwacji nietypowych dla załamania tego estymatora. Często więc mediana będzie odpowiedniejsza, gdy na tym kończymy analizę lub gdy następnym krokiem jest przekodowanie. W innych przypadkach mediany wykorzystać się nie da – właśnie na średniej, pomimo jej braku odporności, opiera się cała statystyka.

2.7. Transformacja i standaryzacja danych ilościowych

Analiza wariancji, a także większość technik statystycznej analizy wielowymiarowej (SAW – *statistical multivariate analysis*) omawianych dalej, wymaga jednorodności grup i normalności rozkładu. Dane będące wynikiem pomiarów żywych organizmów, właściwie z reguły warunków tych nie spełniają. Należy więc albo zrezygnować ze stosowania tych metod, albo spróbować zmodyfikować dane tak, aby warunki te spełniały. Możemy pominąć obserwacje nietypowe i/lub transformować dane. Oczywiście w tym miejscu Czytelnik zareagować może niechęcią i zwątpieniem w „naukowość” takich zabiegów. Kojarzyć się to może bowiem z celowymi manipulacjami danymi. Zastrzeżenia takie nie są jednak zwykle uzasadnione. Oczywiście jest technicznie możliwe manipulowanie danymi poprzez selekcje i transformacje, tak długo, aż wybrany przez nas test przyniesie oczekiwane rezultaty – co byłoby jednak po prostu przemyślanym fałszerstwem – lecz nie wynika stąd, że jakiegokolwiek przekształcenia danych nie są dopuszczalne. Z każdą z obserwacji nietypowych należy się bliżej zapoznać. Gdy dotyczy ona np. jedyne samca, jedyne osobnika z jakiejś populacji, znanej skądinąd z większych/mniejszych rozmiarów, czy okazji ważnego z jakiegokolwiek innego powodu, to oczywiście obserwacji takiej pominąć nie możemy. Z drugiej strony, obserwacje nietypowe często są wynikiem błędnych pomiarów, dotyczą okazji teratycznych, uszkodzonych czy np. częstych u ślimaków gigantów, powstających w następstwie kastracji przez pasożytnicze przywry. W każdym z tych, jak też wielu innych przypadkach, pominięcie takich obserwacji, jako zaciemniających jedynie obraz taksonu, jest oczywiste.

Rzeczywista „wyższość” danych oryginalnych, nietransformowanych, wyrażać się może w ich zgodności ze skalą liniową. Skala taka zapewne najbardziej odpowiada ludzkiemu postrzeganiu rzeczywistości i stąd wydaje się oczywista, choć już nikt nie protestuje przeciwko choćby podawaniu pH w skali logarytmicznej. W rzeczywistości skala liniowa jest zaledwie jedną z bardzo wielu możliwych, a szereg zjawisk biologicznych najlepiej opisują inne skale. Na przykład wzrost masy ciała rosnącego organizmu najlepiej oddaje skala logarytmiczna. W praktyce najprościej jest oglądać wykresy częstości danych w poszczególnych klasach wielkości, zmieniając skalę dla wielkości mierzonej; skala, dla której rozkład będzie najbardziej zbliżony do normalnego i najbardziej jednorodny określi nam transformację, której powinniśmy dokonać. Pamiętajmy jednak, że choć pewnego rodzaju transformacji dokonuje się rutynowo, to jednak bynajmniej nie zapewniają one automatycznie uzyskiwania danych o rozkładzie normalnym, jak zdaje się wierzyć szereg autorów publikacji, zawierających analizy morfometryczne. Pomimo zastosowanej transformacji, średnie podajemy dla danych nietransformowanych, bowiem ich interpretacja jest oczywista – czego nie dałoby się powiedzieć np. o średniej logarytmów. Analiz dokonujemy dla wartości transformowanych, dla nich też obliczamy przedziały ufności, lecz następnie przedziały te musi-

my przeliczyć na oryginalne wartości, pamiętając, że dla transformacji nieliniowych przedział taki będzie asymetryczny.

Najczęściej, wręcz rutynowo, stosuje się **transformację logarytmiczną**. Właściwie zawsze można jej użyć dla danych, które ze swej natury powinny mieć rozkład normalny, czyli np. do wyników pomiarów wielkości różnych struktur – będzie to więc najwłaściwsza transformacja dla większości danych wykorzystywanych w taksonomii numerycznej. Transformacja logarytmiczna minimalizuje lub nawet likwiduje krzywoliniowe zależności pomiędzy danymi, prawoskośność rozkładu (dla rozkładów lewoskośnych wskazana jest transformacja przeciwllogarytmiczna: $y' = e^y$) i korelację wartości średniej z wariancją (szczególnie ważne dla ANOVA i szeregu technik wielowymiarowych). Transformacja logarytmiczna zwykle też zbliża obserwacje nietypowe ku pozostałym. Jeżeli wśród transformowanych wartości występuje zero, to musimy zastosować transformację: $y' = \log(y+1)$, bowiem dla zera logarytm dziesiętny równy jest minus nieskończoności. Gdy wśród danych są wartości z przedziału $(0, 1>$, dla których logarytmy dziesiętne są ujemne, dobrze jest zastosować transformację: $y' = \log(1000y)$. Użycie rutynowo transformacji logarytmicznej wszędzie tam, gdzie dane powinny mieć rozkład normalny, nie jest błędem; błędne jest natomiast zakładanie, że taka transformacja musi rozwiązać wszystkie problemy związane z normalnością i jednorodnością rozkładu – po transformacji należy normalność i jednorodność zbadać.

Jeżeli dane są wynikiem przeliczeń, jak liczba guzków, włosków, kolców, zębów, wyrostków itp., ich rozkład nie będzie normalny, a zbliżać się będzie do rozkładu Poissona. Dla takich danych odpowiednia jest **transformacja pierwiastkowa**, a gdy w macierzy występują zera, należy dodać do wszystkich danych wartość 0,5 przed wyłączeniem pierwiastka kwadratowego:

$$y' = \sqrt{y} \text{ lub } y' = \sqrt{y + \frac{1}{2}}.$$

Dane będące proporcjami lub częstościami wyrażonymi w procentach stosują się do rozkładu dwumianowego i powinny być **transformowane kątowno**, czyli z wykorzystaniem **transformacji arcussinus** pierwiastka kwadratowego wartości transformowanej (*nota bene* wyrażonej w stopniach kątowych, skąd nazwa transformacji): $y' = \arcsin \sqrt{y}$. Inna użyteczna transformacja dla proporcji to:

$$\text{logit}(p) = \frac{1}{2} \log\left(\frac{p}{1-p}\right), \text{ a dla korelacji: } z(r) = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right).$$

Wymienione transformacje powinny na ogół wystarczyć, choć istnieje i bywa używany szereg innych (patrz Sokal i Rohlf 1995).

Wyniki pomiarów są wielkościami mianowanymi, a ich średnie, zakresy zmienności i wariancje mają różne wartości. Obok siebie możemy więc mieć średnią 30 cm i zakres 10–60 cm, średnią 9 mm i zakres 5–11 mm, a także średnią 25 kolców i zakres 23–27 kolców. O ile milimetry możemy przeliczyć na centymetry, to zestawianie wraz z nimi liczby kolców trudno uznać za sensowne. W dodatku w wielu technikach analize wielowymiarowej, jak np. w analizie głównych składowych (PCA), zmienne ważą

proporcjonalnie do wielkości ich wariancji, a we wszystkich właściwie, poczynając już od odległości, zmienne ważą proporcjonalnie do wartości średniej. Jeżeli więc nie chcemy, aby większość zmierzonych danych w rzeczywistości mogła być równie dobrze pominięta w analizie wielowymiarowej, należy wszystkie dane standaryzować. **Standaryzacja** polega na takiej **liniowej transformacji** oryginalnych (bądź już wcześniej transformowanych) danych, aby każda wartość odnosiła się do średniej i rozproszenia dla danej zmiennej, nie zależąc natomiast ani od użytych przy pomiarze jednostek, ani od wielkości średniej czy wariancji. Istnieje szereg sposobów standaryzacji (Sokal i Rohlf 1987, 1995, Rohlf 1994). Oczywiście sensowna będzie standaryzacja w obrębie danej cechy dla wszystkich obiektów, a nie w obrębie danego obiektu dla wszystkich cech. Standaryzacji dokonuje się najczęściej zgodnie ze wzorem:

$$y' = \frac{y - \bar{y}}{s}$$

gdzie y' to wartość standaryzowana, y – wartość przed standaryzacją (transformowana lub nie), \bar{y} – wartość średnia, s – odchylenie standardowe. Standaryzacja ta, najczęściej stosowana i zwykle wystarczająca, należy do całej rodziny standaryzacji, gdzie zamiast średniej i/lub odchylenia standardowego stosować możemy inne wielkości. Od oryginalnej wartości odejmowana może być wartość minimalna dla danej cechy – wówczas ominiemy wartości ujemne, niekiedy kłopotliwe. W mianowniku natomiast, zamiast odchylenia standardowego, podstawia się wartość maksymalną dla danej cechy, zakres (maksymalna – minimalna wartość) albo wariancję.

Standaryzację wspomnianą w Rozdziale 2.4, konieczną w metodzie *gap weighting* (Thiele 1993), przeprowadza się zgodnie z formułą:

$$y' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} n$$

Jest to więc odejmowanie wartości minimalnej i dzielenie przez zakres, lecz uzyskaną proporcję mnożymy przez n , czyli liczbę stanów cechy wielostanowej, dopuszczalną przez stosowany program rekonstruujący filogenezę.

2.8. Rozkłady nieznane, techniki Monte Carlo, liczby losowe, *jackknife*, *bootstrap*; test Mantela

Zdarza się często, że pomimo zastosowania transformacji, dane nadal nie są jednorodne, a ich rozkład wyraźnie odbiega od normalnego, nie przypomina też jakiegokolwiek innego, dla którego dysponujemy opracowanym aparatem matematycznym. Bywa tak, gdy badana cecha powinna mieć określony rozkład, natomiast nie ma go sama próba (np. pobrana niezupełnie losowo, część wartości zmierzonych błędnie, itp.). W innych przypadkach nie wiemy nic o rozkładzie lub wiemy, że musi być inny od zakładanego dla określonych testów statystycznych. We wszystkich takich sytuacjach możemy skorzystać z **randomizacji**, a szerzej, z całej rodziny **technik Monte Carlo**. Ogólną zasadą technik Monte Carlo jest porównywanie danych pochodzących z ob-

serwacji z próbami losowymi, generowanymi zgodnie z założonym modelem (Manly 1998). Dla danych rzeczywistych i każdej z wytworzonych **pseudoprób** (**nibyprób**) obliczamy wartość określonego parametru, który nas interesuje. Gdy wygenerujemy, powiedzmy, 999 prób, z których w czterech wartość tego parametru będzie wyższa niż w próbie pochodzącej z obserwacji, to wiemy, że prawdopodobieństwo uzyskania tak wysokiej wartości parametru w następstwie przypadku jest mniejsze lub równe 0,005. Taka jest zasada technik Monte Carlo w największym możliwym skrócie.

W Rozdziale 2.6 wspominaliśmy już, że reprezentatywna próba pobrana musi być losowo. Skądinąd jednak wielokrotnie wykazano doświadczalnie, że człowiek jest niezdolny do losowego wyboru, a ogólniej – losowych decyzji (Kachigan 1991). Podkreślmy – wszelkie „losowe” wybieranie obiektów (okazów), oparte na świadomej eliminacji wszelkiego świadomego wyboru, nigdy losowym nie będzie. W tym i w szeregu innych przypadków niezbędne jest korzystanie z **liczb losowych**. Formalnie liczbami losowymi nazywamy ciąg, którego nie da się zapisać za pomocą algorytmu krótszego niż sam ciąg. Są to więc liczby, pomiędzy którymi nie da się wskazać jakichkolwiek zależności. Początkowo liczby losowe uzyskano jako środkowe cztery wartości powierzchni parafii w Wielkiej Brytanii lub z tablic logarytmicznych dwunastocyfrowych. Obecnie korzysta się zwykle z liczb tworzonych przez generatory liczb **pseudolosowych**. Określenie „pseudolosowe” oznacza, że liczby te tworzy generator, korzystając z algorytmu eliminującego autokorelacje w generowanym ciągu (coraz nowsze wersje generatorów coraz dokładniej to wykonują), nie zaś, że są one gorsze od „losowych”. Statystycy uważają je raczej za lepsze od losowych we wcześniejszym rozumieniu, bowiem wśród tamtych więcej było zależności.

Zastosowania liczb losowych są liczne i o wielu z nich będziemy mówić w tej książce. Wartości kolejnych liczb losowych można wykorzystać do losowego pobierania okazów do mierzenia. Wystarczy wszystkie ponumerować, a następnie brać te, które odpowiadają kolejnym wartościom liczb losowych. Jeżeli wybieramy obiekty drobne, jak np. chrząszcze czy ślimaki, to zamiast numerowania lepiej i prościej jest wysypać okazy na duże naczynie szklane, pod nim umieścić choćby papier milimetrowy, po czym wybierać zgodnie ze współrzędnymi (x, y) , odczytywanymi jako kolejne wartości liczb losowych. Liczby losowe służą też do wyboru wartości w metodach Monte Carlo. Metody te są ogólnie proste, jeżeli idzie o aparat matematyczny, są natomiast „obliczeniowo-intensywne”, czyli wymagają długich, powtarzanych wielokrotnie obliczeń. Choć więc znane są od przeszło 60 lat, to ich zastosowanie zaczęło się dopiero około 20 lat temu, wraz z rozpowszechnieniem komputerów.

W sytuacji, gdy mamy podstawy do zakładania, że mierzona cecha ma w populacji rozkład normalny lub choć mniej więcej normalny, natomiast nasza próba takiego rozkładu nie ma, posłużyć się możemy techniką **jackknife**, wprowadzoną przez Tukeya (1958). **Jackknife** to po polsku scyzoryk – podręczne narzędzie, którym wykonać można wiele czynności, choć każdą z nich nienajlepiej. Tak też działa **jackknife** – obliczeniowo prosty, nie intensywny, dla dużych prób raczej nie ustępuje innym technikom, dla małych bywa bardzo zawodny (Manly 1998). Założeniem **jackknife** jest, że wariancja czy średnia dla próby o liczebności n będą niemal identyczne jak dla próby o liczebności $n - 1$ (oczywiście będzie tak tym bardziej, im wyższa będzie wartość n). Z próby o liczebności n wybieramy więc n prób o liczebności $n - 1$, każdorazowo pomijając inną wartość. Dla każdej z tak utworzonych nibyprób obliczamy np. średnią, dzięki czemu uzyskujemy estymat błędu standardowego średniej. Jak wspomnieliśmy,

działa to lepiej lub gorzej, zwykle zresztą lepiej dla wartości wcześniej transformowanych logarytmicznie. Warto pamiętać, że *jackknife* nie jest odporne na obserwacje nietypowe, i to w takim samym stopniu, jak dane oryginalne.

Najprostsza techniką Monte Carlo jest **randomizacja**, czyli nieparametryczna technika porównania rozkładu, o którym wiemy mało lub nic, z odpowiednio dużą liczbą prób losowych, czyli powstałych w wyniku przypadku. „Modelem” jest tu więc losowe przestawianie wartości. Możliwe są dwa typy randomizacji: **dokładny** (*exact*), gdy na podstawie teorii rachunku prawdopodobieństwa lub wyliczenia (wyszczególnienia) można wskazać wszystkie możliwe przypadki, oraz **na podstawie próby** (*sampled*), gdy dane doświadczalne porównuje się z odpowiednio dużą próbą generowanych losowo pseudoprob.

Techniką **numerycznego próbkowania** (*numerical resampling*) najczęściej wykorzystywaną przy rekonstrukcji filogenezy jest *bootstrap*, wprowadzona przez Efrona (1979) do statystyki, a przez Felsensteina (1985) do taksonomii numerycznej. W sytuacji, gdy brak jakiegokolwiek innej informacji o badanej populacji, wartości w próbie losowej są najlepszym źródłem wiedzy o rozkładzie w tej populacji, a numeryczne próbkowanie pobranej próby stanowi najlepszą wskazówkę, czego można by oczekiwać, mając kolejne rzeczywiste próby pobrane z badanej populacji. Próbkowanie numeryczne, oprócz dostarczania danych o parametrach nieznanego rozkładu, pozwala też oszczędzić czas i wysiłek związany z pobieraniem i mierzaniem kolejnych prób z badanej populacji, jak też umożliwia wnioskowanie o parametrach w sytuacji, gdy zmierzono wszystkie dostępne obiekty. *Bootstrap* polega na tym, że z próby o liczebności n losuje się z powtórzeniami m -krotnie po n elementów i dla utworzonych w ten sposób m pseudoprob oblicza badany parametr. Losowanie z powtórzeniami oznacza, że niemal zawsze w niypróbie zabraknie pewnych elementów próby obserwowanej, podczas gdy niektóre inne wejdą do niej więcej niż jeden raz. Oczywiście liczba m pseudoprob powinna być duża, im większa, tym lepiej. W praktyce dla osiągnięcia poziomu istotności 0,05 konieczne jest nie mniej niż 1000 pseudoprob, zaś dla poziomu istotności 0,01 – co najmniej 5000 (Manly 1998).

Wykazano, że średnia z prób otrzymanych techniką *bootstrap* zbliża się do średniej z populacji, z której pochodzi próba poddana *bootstrap*, a odchylenie standardowe średniej z *bootstrap* zbliża się do błędu standardowego obliczonego dla kolejnych rzeczywistych prób, pobieranych bez powtórzeń z badanej, nieznannej populacji (Sokal i Rohlf 1995). *Bootstrap* pozwala więc obliczyć błędy standardowe i przedziały ufności (te ostatnie jednakże przy założeniu co najmniej w przybliżeniu normalnego rozkładu dla badanej rzeczywistej populacji) większości parametrów, choć nie wszystkich (np. mediany). Wykazano też, że dla dużych prób estymaty uzyskane metodami *jackknife* i *bootstrap* są sobie w przybliżeniu równe. Dla małych prób także *bootstrap* bywa zawodne, choć ogólnie jest techniką lepszą, o większej mocy niż *jackknife*. Przez moc testu rozumiemy stosunek liczby statystycznie istotnych wyników testu do liczby przypadków, gdy hipoteza zerowa powinna być odrzucona: idealny ma moc 1,0.

W wielu dziedzinach biologii, także w taksonomii, zachodzi potrzeba zbadania związku pomiędzy dwiema powstałymi niezależnie macierzami różnic lub odległości pomiędzy obiektami. Obiektami mogą być np. osobniki, populacje czy gatunki, dla których pragniemy choćby określić, czy odległości morfologiczne są odbiciem genetycznych, a jedne i drugie czy wykazują związek z odległościami geograficznymi pomiędzy obiektami. Badane macierze muszą dotyczyć tych samych obiektów, a więc

zawierać tyle samo odległości, wszystkie odległości muszą być znane. Jak już wspomnieliśmy, macierze muszą być niezależne – oczywiście nie idzie tu o rzeczywistą, biologiczną niezależność, którą przecież chcemy właśnie sprawdzić, a jedynie o niezależność formalną, czyli jedna macierz nie może być wynikiem przekształcenia, choćby najbardziej złożonego, macierzy drugiej. Macierze wykażą związek wówczas, gdy wartościom większym w jednej odpowiadać będą większe w drugiej, a mniejszym – mniejsze, bardziej niż miałyoby to miejsce w następstwie przypadku. Metodę badania związku pomiędzy takimi macierzami zaproponował Mantel (1967).

Jak wspominaliśmy, dla n obiektów obliczyć można symetryczną macierz różnic bądź odległości o rozmiarach $n \times n$, w której wzdłuż przekątnej będą leżały zera, a lewy dolny trójkąt będzie zawierał te same wartości co prawy górny. Wystarczy więc porównanie macierzy trójkątnych o rozmiarach $n(n-1)/2$. Dla pary macierzy X , Y obliczyć możemy tzw. iloczyn Hadamarda:

$$Z = \sum_{i=1}^{n-1} \sum_{j=i+1}^n X_{ij} Y_{ij}$$

Jak widać ze wzoru, jest to suma iloczynów odpowiadających sobie elementów macierzy – oczywiście im częściej większym wartościom jednej macierzy odpowiadać będą większe wartości drugiej, tym wyższa będzie wartość statystyki Z , odzwierciedlająca pozytywną korelację pomiędzy macierzami. W tym miejscu nasuwać się może pytanie: dlaczego nie oblicza się po prostu Pearsonowskiego współczynnika korelacji pomiędzy macierzami? Otóż oblicza się, tym bardziej że skalowany od 0 do 1 współczynnik łatwiej interpretować niż pozbawioną takiego skalowania wartość statystyki Z . Podkreślić jednak trzeba, że rozkłady zarówno Z , jak i obliczonego w takim przypadku współczynnika korelacji nie są znane, a więc nie znamy też poziomów istotności uzyskanych współczynników związku między macierzami. Założeniem teorii korelacji jest badanie związku pomiędzy kolejnymi niezależnymi obserwacjami dwuwymiarowymi, tymczasem odległości bądź różnice w obrębie macierzy niezależne nie są – zmiana odległości jednego obiektu w stosunku do drugiego musi pociągać za sobą zmianę jego odległości w stosunku do wszystkich pozostałych obiektów.

Bez względu więc na to, czy obliczamy wartość statystyki Z , czy współczynnika korelacji r , poziom istotności obliczyć musimy metodą próbkowania numerycznego. Dla danej pary macierzy przeprowadzamy je identycznie dla Z i r . W tym celu jedną z macierzy – obojętne którą – pozostawiamy niezmienną, a elementy drugiej poddajemy permutacji, obliczamy wartość Z , znów poddajemy permutacji i obliczamy Z , i znów, zwykle kilkaset lub kilka tysięcy razy. Liczba możliwych permutacji równa jest $n!$, a więc dla macierzy obliczonej dla 10 obiektów wynosi 3.628.800. Tak więc już dla niewielkich macierzy obliczanie wszystkich permutacji byłoby nawet dla szybkiego komputera niezmiernie czasochłonne, toteż oblicza się część losowo wybranych permutacji i to okazało się wystarczające. W praktyce dla uzyskania poziomu istotności 0,05 wystarcza 1000, a dla 0,01 potrzeba 5000 permutacji (Manly 1998). Następnie porównujemy nasz wynik z uzyskanymi z permutacji. Jeżeli na przykład okaże się, że z 999 permutacji jedynie dla czterech Z było większe od Z obliczonego dla oryginalnej pary macierzy, to oznacza to, że prawdopodobieństwo uzyskania tak wysokiej wartości

Z dla macierzy, pomiędzy którymi jakiegokolwiek związki są wyłącznie następstwem przypadku, wynosi 0,005, a więc macierze wykazują statystycznie istotną zależność.

2.9. Odległości dla cech ilościowych ciągłych

Obiekty, reprezentowane przez wielowymiarowe obserwacje, w interpretacji geometrycznej traktować można jako punkty w n -wymiarowej **hiperprzestrzeni** cech diagnostycznych. Naturalne jest więc określanie odległości pomiędzy obiektami: im obiekty są podobniejsze tym odległość pomiędzy nimi jest mniejsza. Istnieje też szereg technik, zarówno fenetycznych jak i filogenetycznych, dla których punktem wyjścia jest macierz odległości, a nie dane o kolejnych obiektach. Wygodnie jest przyjąć, że hiperprzestrzeń jest hiperprzestrzenią **metryczną**, czyli że jej topologia ma być określona przez funkcję metryczną. Dlatego miary braku podobieństwa taksonomicznego, jak również funkcje, które przekształcają miary podobieństwa w miary braku podobieństwa, powinny mieć własności **metryki**. Funkcja φ jest metryką, gdy spełnia następujące cztery warunki:

$$(1) \varphi(a, b) \geq 0 \text{ i } \varphi(a, a) = \varphi(b, b) = 0,$$

$$(2) \varphi(a, b) = \varphi(b, a),$$

$$(3) \varphi(a, c) \leq \varphi(a, b) + \varphi(b, c),$$

$$(4) a \neq b \Rightarrow \varphi(a, b) > 0.$$

Warunek (1) oznacza, że identyczne obiekty są nieodróżnialne, gdy nieidentyczne są odróżnialne lub nie. Warunek (2) mówi, że odległości są symetryczne. Warunek (3) to tzw. **warunek trójkąta**, czyli nierówność trójkąta. Warunek (4) stanowi, że gdy a , b różnią się stanami cech, to odległość między nimi musi być większa niż 0. Warunki te są oczywiste w przestrzeni Euklidesowej, ale nie jest to jedyna możliwa przestrzeń. Niektóre odległości nie spełniają warunku (4), określamy je jako **pseudometryczne** lub **semimetryczne**. Szereg użytecznych odległości to semimetryki. Aksjomat (3) może być słabszy:

$$(3') \varphi(a, c) \leq \max [\varphi(a, b), \varphi(b, c)],$$

wówczas mówimy o **ultrametryce**, czyli ultrametrycznej odległości – ultrametrycznych odległości wymagają np. fenogramy, obliczone metodą analizy skupisk (*clustering*). Ultrametryczność (*ultrametricity*) oznacza takie samo tempo ewolucji w obrębie całego drzewa, czyli działanie zegara molekularnego, co jest założeniem mocnym, lecz na ogół mało realistycznym. Aby odzwierciedlać ewolucyjne zmiany także i wtedy, gdy tempo ewolucji nie jest stałe w obrębie drzewa, metryka powinna też spełniać warunek **addytywności** (*additivity*), dany tzw. warunkiem **czteropunktowym** (*four-point condition*):

$$(5) \varphi(a, b) + \varphi(c, d) \leq \max [\varphi(a, c) + \varphi(b, d), \varphi(a, d) + \varphi(b, c)],$$

co oznacza, że spośród trzech sum: $\varphi(a, b) + \varphi(c, d)$, $\varphi(a, c) + \varphi(b, d)$ i $\varphi(a, d) + \varphi(b, c)$ dwie większe są sobie równe.

Najogólniejszą metryką jest **metryka Minkowskiego**, dla i -tego oraz k -tego obiektu dana wzorem:

$$d_{ik} = \left[\sum_{j=1}^m |x_{ij} - x_{kj}|^p \right]^{1/p},$$

gdzie p – liczba naturalna, określająca rodzaj metryki; m – liczba cech; x_{ij} , x_{kj} – realizacja j -tej cechy w obiekcie i -tym oraz k -tym (Pociecha i inni 1988). Metryka Minkowskiego opiera się na następujących założeniach: realizacje poszczególnych zmiennych są liczbami mianowanymi, różnice te są kumulowane w sposób addytywny, liniowy, wreszcie zmienne są od siebie niezależne. Założenia te istotnie ograniczają użycie metryk Minkowskiego. Aby używać ich sensownie, konieczne jest doprowadzenie do porównywalności różnych zmiennych, czyli wszystkie zmienne powinny być mierzone w takich samych jednostkach lub być niemianowane. Ponadto rzędy wielkości kolejnych zmiennych powinny być podobne, by te o wyższych wartościach nie ważyły znacznie więcej w obliczanej odległości. W praktyce więc cechy, na podstawie których oblicza się odległości Minkowskiego, powinny być wcześniej standaryzowane. Choć takie same wagi dla wszystkich cech budzą wątpliwości, to brak jakichś sensownych metod ważenia powoduje, że dla wszystkich cech przyjmuje się takie same wagi dla obliczania odległości. Warunek niezależności cech w praktyce, jak pisaliśmy, zwykle nie jest spełniony, staramy się jednak unikać cech niewątpliwie silnie skorelowanych. Odmianą metryki Minkowskiego dla $p = 1$ jest **odległość miejska** (*Manhattan distance, city block*):

$$d_{ik} = \sum_{j=1}^m |x_{ij} - x_{kj}|,$$

a dla $p = 2$ **odległość Euklidesowa** (*Euclidean distance*), najczęściej stosowana w analizach wielowymiarowych, jako najbardziej „naturalna”:

$$\Delta_{ik} = \left[\sum_{j=1}^m (x_{ij} - x_{kj})^2 \right]^{1/2}$$

Wartości metryk Minkowskiego rosną z liczbą zmiennych, co nie jest dogodne, a zupełnie wyklucza jakiegokolwiek porównania odległości dla różnych zestawów danych. Aby choć częściowo te niedogodności ominąć, oblicza się wartości średnie. **Średnia różnica**, dana wzorem:

$$d_{ik} = \frac{1}{m} \sum_{j=1}^m (x_{ij} - x_{kj}),$$

nie jest użyteczna, bowiem wartości różnicy będą raz dodatnie, raz ujemne. Niedogodność tę wyeliminowano poprzez dodawanie wartości absolutnych (bezwzględnych) w **średniej różnicy cech** (*mean character difference*) wprowadzonej przez Czekanowskiego (1909, 1932), a później Caina i Harrisona (1958), czyli **odległości miejskiej średniej**:

$$d_{ik} = \frac{1}{m} \sum_{j=1}^m |x_{ij} - x_{kj}|.$$

Zaletą tej odległości jest prostota i metryczność. Wadą to, że zawsze zaniża estymaty odległości Euklidesowej, co może mieć niekorzystne następstwa, choć niekoniecznie; nie ma też niektórych dogodnych własności odległości taksonomicznej – w sumie tak się ma do odległości taksonomicznej jak średnie odchylenie do standardowego odchylenia. **Średnia odległość taksonomiczna** (*average taxonomic distance*) to **średnia odległość Euklidesowa**, zapewne najczęściej używana w taksonomii:

$$E_{ik} = \left[\frac{1}{m} \sum_{j=1}^m (x_{ij} - x_{kj})^2 \right]^{1/2}.$$

Dla wartości standaryzowanych, zakładając, że każda z m cech jest niezależna od pozostałych i ma rozkład normalny ze średnią zero i wariancją równą jeden, dla dużych wartości m wartość oczekiwana E zbliża się do $\sqrt{2}$, a oczekiwana wariancja w przybliżeniu równa się $1/m$, a więc osiąga zero gdy m nieskończoność; po przekroczeniu $m \approx 75$ wariancja maleje bardzo wolno (Sneath i Sokal 1973). Niekiedy używa się także kwadratów odległości Euklidesowej lub średniej odległości taksonomicznej. Warto wspomnieć o dość często wykorzystywanej odmianie odległości miejskiej, znanej jako **Canberra metric** (Lance i Williams 1967a, Sneath i Sokal 1973):

$$d_{CANB(i,k)} = \sum_{j=1}^m \left(\frac{|x_{ij} - x_{kj}|}{(x_{ij} + x_{kj})} \right).$$

Odległość ta (używana też w wersji średniej, czyli dzielonej przez m), może być stosowana jedynie dla dodatnich wartości cech, a więc nie można jej obliczać np. dla cech standaryzowanych. Jej zaletą jest zależność jedynie od osobników i grup porównywalnych, a nie zakresów zmienności cech, jest też czuła raczej na proporcjonalne, a nie absolutne różnice. Zbliżony do **Canberra metric** jest **współczynnik odmienności** (*coefficient of divergence*), także obliczany dla niestandaryzowanych danych i dogodnie skalowany, od zera do jedności (Clark 1952, Sneath i Sokal 1973):

$$CD_{ik} = \left[\frac{1}{m} \sum_{j=1}^m \left(\frac{x_{ij} - x_{kj}}{x_{ij} + x_{kj}} \right)^2 \right]^{1/2}$$

Z innych częściej stosowanych odległości wspomnieć też można o *Cattell's coefficient of pattern similarity*, niemal liniowo skorelowanym ze średnią odległością taksonomiczną, którą więc zawsze można go zastąpić (Sneath i Sokal 1973); więcej odległości zestawiają Sokal i Sneath (1963) oraz Rohlf (1994). Warto też wspomnieć, że większość **dopelnień do jedności** ($1 - \theta$) rozmaitych **współczynników związku (asocjacji, podobieństwa)** ma własności metryk, teoretycznie więc można ich także używać jako metryk. Takim najczęściej używanym współczynnikiem jest Pearsonowski **współczynnik korelacji r** .

W ujęciu stochastycznym, gdy obiekty traktowane są jako obserwacje z próby, używa się innych miar odległości: współczynnika podobieństwa ras Pearsona (*Pearson's coefficient of racial likeness*) lub uogólnionej odległości (*generalized distance*) Mahalanobisa i Rao. Obiektami są tu populacje, a więc oznaczamy je indeksami dolnymi I oraz K , dla odróżnienia od indeksów i oraz k dla obiektów będących osobnikami. **Współczynnik podobieństwa ras** (Pearson 1926) dany jest wzorem:

$$C.R.L. = \left[\frac{1}{n} \sum_{j=1}^m \left(\frac{(\bar{X}_{jI} - \bar{X}_{jK})^2}{(s_{jI}^2 / t_I) + (s_{jK}^2 / t_K)} \right) \right]^{1/2} - \frac{2}{n},$$

gdzie \bar{X}_{jI} to średnia próby I dla cechy j , s_{jI} to wariancja próby I dla cechy j , a t_I to liczebność próby I . Uogólniona odległość (Mahalanobis 1936, Rao 1948) w taksonomii stosowana jest stosunkowo niedawno, bowiem jej obliczanie wymaga komputera. Teoretycznie wielowymiarowa normalność rozkładów, jednorodność prób i ich losowe pobieranie są warunkami jej stosowania, a warunki te zazwyczaj nie są spełnione. Z drugiej strony, szereg studiów wykazało znaczną odporność uogólnionej odległości na niespełnianie tych warunków (Sneath i Sokal 1973). Uogólniona odległość obliczana jest tak, aby dla tych liniowych kombinacji cech, które mają największą wariancję pomiędzy parami grup, maksymalizować różnice między parami średnich proporcjonalnie do łącznej wariancji wewnątrzgrupowej. Obliczyć ją można ze wzoru:

$$D_{IK}^2 = \delta_{IK}^* \mathbf{W}^{-1} \delta_{IK},$$

gdzie \mathbf{W}^{-1} jest odwrotnością macierzy łącznej wariancji-kowariancji (rozproszenia) wewnątrz prób (o wymiarach $n \times n$), a δ_{IK} jest wektorem różnicy pomiędzy średnimi dla prób I i K dla wszystkich cech. Istnieje transformacja Mahalanobisa, która przekształca obserwacje w taki sposób, że odległości Euklidesowe pomiędzy obserwacjami transformowanymi równe są odległościom Mahalanobisa pomiędzy oryginalnymi obserwacjami (Jajuga 1993): pierwiastek kwadratowy D_{IK}^2 to odległość Euklidesowa w D -hiperprzestrzeni. Uogólniona odległość znajduje zastosowanie w wielu technikach analizy wielowymiarowej, np. dyskryminacyjnej. Jej użycie w technikach rekonstrukcji filogenezy budzi jednak zastrzeżenia. Określając ze swej natury odległości jako funkcję zachodzenia na siebie przedziałów zmienności dla par populacji i transformując oryginalne odległości tak, aby zmaksymalizować różnice pomiędzy obiektami należącymi do różnych, określonych *a priori* populacji, może znacząco zmieniać oryginalne

odległości, co na ogół nie jest pożądane. Stosując uogólnioną odległość, trudno jest też badać odrębność wyróżnianych *a priori* taksonów. Ponadto – gdy uwzględnia się dużą liczbę cech – zachodzi niebezpieczeństwo, że niewielkie wartości własne, których prawdziwe wartości mogą być silnie zafałszowane błędami zaokrąglania, mogą wnieść duży losowy błąd do wartości D_{IK}^2 . Również odporności D_{IK}^2 na niespełnianie warunku wielowymiarowej normalności rozkładów (zwykle niespełnianego dla taksonów biologicznych, zwłaszcza powyżej poziomu populacji) nie należy przeceniać. Jeżeli korelacje pomiędzy cechami zaznaczają się słabo bądź są równe, to D_{IK}^2 jest w przybliżeniu równy obliczonemu dla cech standaryzowanych; zazwyczaj też bezpieczniejsze będzie użycie średniej odległości taksonomicznej, czyli średniej odległości Euklidesowej, właśnie dla cech standaryzowanych.

2.10. Odległości dla cech jakościowych wielostanowych i binarnych

W taksonomii często dysponujemy cechami jakościowymi lub ilościowymi nieciągłymi, czyli takimi, które kodować musimy jako dyskretne wielostanowe lub binarne. Dla cech ilościowych dyskretnych stosować można (i należy) opisane w poprzednim rozdziale odległości dla cech ilościowych ciągłych. Jeżeli znamy serię transformacyjną dla cech jakościowych wielostanowych, wiemy, że przejście np. ze stanu 1 do 3 musi pociągać za sobą większe zróżnicowanie niż pomiędzy stanami 1 i 2 bądź 2 i 3; również odległości dla cech ciągłych są odpowiedniejsze. Gdy jednak takich przesłanek brak, konieczne jest użycie innych współczynników, stworzonych dla cech wielostanowych lub binarnych. Jak pamiętamy, wielostanowe można addytywnie przekodowywać na binarne, dla których istnieje więcej współczynników niż dla wielostanowych. Wykorzystuje się wspomnianą już metryczność dopełnień do jedności większości współczynników asocjacji (związku, podobieństwa). Istnieje znaczna liczba takich współczynników (Sokal i Sneath 1963, Sneath i Sokal 1973, Rohlf 1994), lecz większości z nich użyto co najwyżej parę razy i trudno ocenić ich przydatność. Ograniczymy się do kilku. Dla łącznej liczby cech n oznaczamy przez m (*match*) liczbę cech, których stany były identyczne w obu obiektach, dla których obliczamy współczynnik asocjacji, a u (*unmatch*) liczbę cech, których stany były różne. Dla cech wielostanowych lub binarnych:

$$S_{SM} = \frac{m}{m + u} = \frac{m}{n}$$

jest to **współczynnik pokrywania się stanów cech** (*simple matching coefficient*), intuicyjnie oczywisty, jeden z najprostszych i najstarszych, wprowadzony do taksonomii numerycznej przez Sokala i Michenera (1958). Oczywiście współczynnik mieści się w przedziale $\langle 0, 1 \rangle$. Pomimo prostoty często okazuje się najlepszy. W dodatku dopełnienie do jedności $1 - S_{SM}$ jest metryczne i równe kwadratowi odległości Euklidesowej obliczonej dla niestandaryzowanych wartości stanów cech, przyjmujących wartości 0 lub 1, czyli $E = \sqrt{1 - S_{SM}}$. Rozkład wartości S_{SM} jest bliski dwumianowemu, więc jego

wysokie wartości charakteryzuje niska wariancja (Sneath i Sokal 1973). Choć współczynnik nadaje się zarówno dla cech wielostanowych, jak i binarnych, dla tych pierwszych ma potencjalnie istotną niedoskonałość. Intuicyjnie można sądzić, że pokrywanie się stanu cechy, powiedzmy trójstanowej, ważyć powinno mniej niż pokrywanie się stanu cechy, powiedzmy siedmiostanowej, a takich różnic współczynnik ten nie uwzględnia. Zastrzeżenie wydaje się znaczące, choć z drugiej strony brak jakiegokolwiek sensownej metody różnicowania wag (kosztów, prawdopodobieństw). Pozostaje więc uznać zasadność traktowania wszystkich cech tak samo, tym bardziej że skomplikowane współczynniki uwzględniające zróżnicowane wagi cech dały w wyniku wartości podobne do otrzymanych z zastosowaniem współczynników prostych. Używany dość często **współczynnik Rogersa i Tanimoto** (1960):

$$S_{RT} = \frac{m}{n+u} = \frac{m}{m+2u}$$

jest monotoniczny w stosunku do S_{SM} i ma podobne własności, więc może być zastąpiony przez S_{SM} . Inaczej skalowany, w przedziale $\langle -1, 1 \rangle$, a osiągający wartość 0 wtedy, gdy pokrywanie się stanów zachodzi dla połowy cech, **jest współczynnik Hamana** (1961), poza tym mający własności podobne jak dwa wcześniej omówione:

$$S_H = \frac{m-u}{n} = \frac{m-u}{m+u}$$

Dla cech binarnych porównanie dwóch taksonów dla danej cechy może dać jeden z czterech rezultatów, które oznaczymy kolejnymi symbolami $a\dots c$: 1,1 (a); 1,0 (b); 0,1 (c) i 0,0 (d). Gdy – jak jest często – 1 oznacza obecność jakiejś struktury czy atrybutu, a 0 brak, to wspólne występowanie powinno być traktowane jako ważniejsze dla taksonomii niż wspólny brak, czyli a waży więcej niż d , co intuicyjnie wydaje się oczywiste. Jednym z najstarszych i szeroko stosowanych, zwłaszcza w ekologii, jest **współczynnik Jackarda** (1908):

$$S_J = \frac{a}{a+u} = \frac{a}{a+b+c} = \frac{a}{n-d}$$

Prosty i przybierający wartości w przedziale $\langle 0, 1 \rangle$, zupełnie pomija „wspólny brak”, podobnie jak jego modyfikacja, monotoniczna z S_J , znana jako **współczynnik Dice’a** (1945), a przypisująca współwystępowaniu wyższe wagi: $S_D = 2a/(2a+u)$. Dla obu współczynników dopełnienia do jedności nie są metrykami, a jedynie pseudometrykami i nie zawsze spełniają warunek trójkąta. Zarazem wątpliwości budzi zupełne pomijanie „wspólnego braku”; to ostatnie zupełnie wyklucza ten współczynnik w przypadkach, gdy kodowanie binarne dotyczy obecności jednej lub drugiej struktury (stanu cechy, jak np. barwa oczu), a nie występowania lub braku określonej struktury. Niedogodności te omija **współczynnik Yule’a** (1911; za Sneath i Sokal 1973), dogodnie skalowany od -1 (zupełny brak współwystępowania stanów cech) po 1 (doskonałe pokrywanie się stanów cech):

$$S_y = \frac{ad - bc}{ad + bc}.$$

Warto też wspomnieć o **odległości Russela i Rao** (1940; za Rohlf 1994): $S_{RR} = a/n$, jak też **Kulczyńskiego** (1927; za Rohlf 1994): $S_{KI} = a/u$; ten ostatni jest prosty, jednak jego wartość nie istnieje dla obiektów o takich samych stanach wszystkich cech, co może być kłopotliwe. Użyteczny bywa też Pearsonowski **współczynnik korelacji r** w postaci **dla danych binarnych** (Pociecha i inni 1988):

$$S_\varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Istnieją też współczynniki asocjacji, umożliwiające jednoczesne użycie cech binarnych, wielostanowych i ilościowych ciągłych, takie jak współczynnik Gowera. Omawiają je Sneath i Sokal (1973). Alternatywnie możemy cechy ciągle przekodować na wielostanowe, po czym wielostanowe na binarne addytywne i ostatecznie użyć współczynników dla cech binarnych. Warto jeszcze wrócić do przekształcania wartości asocjacji, czyli związku bądź podobieństwa, w odległości. Wspomniana formuła: $d = 1 - S$ nie zawsze jest wystarczająca, zawodzi zresztą przy współczynnikach skalowanych poza przedziałem $\langle 0, 1 \rangle$. Wówczas należy rozważyć dwie inne możliwości: $d = -\ln S$ oraz $d = 1/S - 1$ (warto pamiętać, że np. $1/r$ nie jest metryką). Wybór pomiędzy powyższymi wariantami powinien być każdorazowo oparty na biologicznych, ewolucyjnych bądź fizykochemicznych (dla danych molekularnych) przesłankach. Przy niewielkich różnicach wszystkie trzy formuły dadzą taki sam wynik, przy większych – bardzo różny.

2.11. Częstości elektromorf i odległości genetyczne

Wynikiem elektroforezy białek enzymatycznych jest zestaw mniej lub bardziej wyraźnie zaznaczających się prążków, zwanych **elektromorfami**, które różnią się ładunkiem elektrycznym, a w przypadku elektroforezy pionowej na żelu poliakrylamidowym (PAGE) również wielkością cząsteczki. Dopiero znajomość dziedziczenia i/lub czwartorzędowej struktury danego białka (Richardson i inni 1986, Murphy i inni 1996) umożliwia genetyczną interpretację określonych elektromorf bądź ich grup jako **allozymów**, przypisując im kodowanie przez określony allel. Dla każdego *locus* mamy więc zestaw **częstości alleli**. Jedni teoretycy argumentują, że istotna jest obecność bądź brak danego allela, a nie częstość jego występowania, bowiem uzyskanie lub utrata allela – w wyniku mutacji – to ważne wydarzenie ewolucyjne, podczas gdy częstości wahają się choćby w następstwie dryfu genetycznego, a bez rzeczywiście bardzo licznej próby ich estymaty obarczone są dużym błędem. Inni uważają, że jest ważne, czy dany allel jest rzadki, czy bardzo częsty, a zresztą rzadkie allele łatwo całkowicie przemoczyć w niedostatecznie licznej próbie. Oba stanowiska nie są pozbawione słuszności i brak jakiegś ogólnie przyjętej przez wszystkich metody analizy taksonomicznej danych allozymatycznych.

Dla analizy kładystycznej (*parsimony*) najdogodniejsze jest kodowanie jakościowe, ignorujące częstości alleli. Pozornie najprościej byłoby uznać każdy z alleli za cechę,

ze stanami: obecny i brak. Takie kodowanie łamie jednak zasadę niezależności cech – w każdym *locus* częstości muszą się przecież sumować do jedności, a więc w obrębie danego *locus* należałoby przynajmniej jeden z alleli pominąć – a ponadto często prowadzi do biologicznie nieakceptowalnych rekonstrukcji stanów cech u przodków, czyli braku jakichkolwiek alleli w danym *locus* na którymś z etapów zrekonstruowanej filogenezy (Buth 1984, Swofford i Olsen 1990, Murphy 1993, Swofford i inni 1996). Nie narusza niezależności cech i może prawidłowo odzwierciedlać homologie uznanie *locus* za cechę, ale wtedy napotykałyśmy opisane we wcześniejszych rozdziałach problemy z polimorfizmem. Należy więc kodować jako odrębny stan każdą spotykaną kombinację alleli (np. $a - 0$, $b - 1$, $c - 2$, $ab - 3$, $bc - 4$, $abc - 5$). Mickevich i Mitter (1981, 1983; Emberton 1994) przedstawili **metodę minimalizacji zmian** (*minimum turnover*). Każdy *locus* traktuje się jako odrębną cechę. Eliminuje się allele obecne jedynie u jednego taksonu (autapomorfie). Następnie uznaje się każdą z kombinacji alleli za odrębny stan cechy. Kolejny krok to uporządkowanie kolejnych stanów danej cechy, czyli kolejnych kombinacji alleli w danym *locus*, w postaci drzewa stanów cech minimalizującego liczbę zmian cechy, a po uzyskaniu takich drzew dla wszystkich *loci* – znalezienie na ich podstawie drzewa dla wszystkich cech łącznie.

Powyższa metoda jest dość uciążliwa w stosowaniu, nie zawsze też prawidłowo odzwierciedla homologie i może nie zapewniać uzyskania drzewa najkrótszego, czyli najbardziej zgodnego z kryterium *parsimony* (Rozdział 4.5), bowiem często uniemożliwia prawidłową rekonstrukcję stanów cech przodków. Jest tak w wyniku ograniczenia stanów cech do obecnych u taksonów terminalnych, a przecież tak być nie musiało: występowanie np. alleli *a* i *b* wyłącznie razem, jako *ab*, nie wyklucza obecności u przodka jedynie *a* albo *b*. Dla kodowania jakościowego (występowanie/brak alleli) najodpowiedniejsza wydaje się technika zaproponowana przez Mabee i Humphriesa (1993), uzupełniona przez Mardulyn i Pasteels (1994). Tu również każdej stwierdzonej kombinacji alleli przypisujemy odrębny kod, czyli uznajemy za odrębny stan cechy. Następnie uzupełniamy zestaw stanów cech o inne, czyli kombinacje alleli możliwe dla alleli stwierdzonych w danym *locus*, a więc być może obecne u przodków. Wówczas jednak macierz szybko może okazać się monstrualnie wielka: np. dla pięciu alleli w *locus* możliwe jest 31 kombinacji, dla siedmiu ponad 100, a dla dziesięciu – ponad 1000. Musimy więc ograniczyć się jedynie do kombinacji, które mogą się pojawić w zrekonstruowanej filogenezie. Sposób ich wyboru omawiają Mardulyn i Pasteels (1994). Jeżeli jakaś grupa alleli pojawia się więcej niż raz u taksonów terminalnych, to należy dla niej utworzyć kolejny, odrębny stan. Gdy grupa alleli obecna jest raz, a każdy z alleli tej grupy niezależnie obecny gdzie indziej, to dla tej grupy należy też utworzyć kolejny stan. Jeżeli dla trzech danych zestawów alleli mamy wszystkie możliwe pary (np. dla *a*, *b* i *cd* mamy: *ab*, *acd* i *bcd*), to należy utworzyć kolejny stan, zawierający te wszystkie allele (*abcd*). Ostatecznie warto się przyjrzeć rozkładowi stanów cech na drzewie, aby upewnić się, czy jakaś inna kombinacja alleli w którymś z *loci*, u któregoś ze zrekonstruowanych przodków, nie pozwoliłaby uzyskać lepszej rekonstrukcji. Następnym etapem jest sporządzenie macierzy kroków, czyli kosztów zmian (patrz serie transformacyjne) pomiędzy wszystkimi stanami cech – umożliwiającą to takie programy, jak MacClade (Maddison i Maddison 1992) czy Paup (Swofford 1996). Wydaje się oczywiste, że polimorfizm *ab* powinien być stadium pośrednim pomiędzy *a* i *b*. Zwykle też uznaje się utratę allelu za jeden krok, tak samo jak uzyskanie, czyli przejście $a \rightarrow b$ kosztuje dwa kroki. Tu znów brak zgodności: szereg badaczy

uważa, że uzyskanie allela jest kosztowniejsze niż utrata i powinno ważyć dwa. Choć jest w tym pewna racja, przyjęcie takiego założenia powodować musi obciążenie rekonstrukcji w kierunku utraty alleli i tendencję do rekonstruowania przodków mających więcej alleli niż taksony współczesne, a to ewolucyjnie nie ma uzasadnienia. Z drugiej strony, jest oczywiste, że wspólne występowanie allelu znaczy więcej niż wspólny brak, a niezależne, równoległe uzyskanie allelu jest mniej prawdopodobne od równoległej jego utraty. Uwzględnić to można przez przypisanie kolejnym *loci* różnym wag, proporcjonalnych do liczby alleli występujących u więcej niż jednego taksonu.

Jeżeli nie chcemy pomijać częstości alleli, teoretycznie możemy przekodować je jak opisaliśmy dla cech ilościowych ciągłych, jednak w połączeniu z polimorfizmem dałoby to bardzo skomplikowaną macierz cech. Grupowanie taksonów, lecz nie w postaci hierarchicznego drzewa, uzyskać można bezpośrednio dla częstości alleli przy użyciu **analizy odpowiadania** (*correspondence analysis*), którą zajmiemy się w Rozdziale 4.6. Inna możliwość to wspomniana już technika *frequency parsimony* (Swofford i Berlocher 1987), minimalizująca zmiany częstości alleli. Można też wykorzystać niektóre metody **maksymalizacji wiarygodności** (*maximum likelihood* – Rozdział 4.6). Wreszcie pozostaje cała gama technik opartych na odległościach genetycznych. Obliczeniowo najprostsze, przez lata były rutynowo stosowane dla częstości alleli i do dziś dominują, choć znaczenie innych technik rośnie.

Pozostaje przedstawić najważniejsze – bądź zwyczajnie najczęściej stosowane – odległości genetyczne, przy czym znów brak jakiegokolwiek odległości zdecydowanie lepszej od innych czy uznanej za standard. Teoretycznie częstości alleli traktować można tak jak jakiegokolwiek zmienne ciągłe i wówczas stosowane odległości będą podobne do opisanych dla cech ciągłych. Z drugiej strony, częstości alleli ze swej istoty łatwo poddają się interpretacji genetycznej i wówczas stosuje się odległości stworzone specjalnie dla takich danych, bacząc, aby spełnione były warunki, dla których zaproponowano te odległości. Zaczniemy od odległości niemających ściślej określonej interpretacji genetycznej. Przez L oznaczmy liczbę *loci*. Sumowane są różnice dla kolejnych *loci*, przy brakujących danych liczba *loci* l dla kolejnych par taksonów może być różna. Najprostszą odległością jest **odległość miejska** (*Manhattan distance*) skalowana liczbą *loci*, przypisana **Prevostiemu** przez **Wrighta** (1978):

$$D_{P(i,k)} = \frac{1}{2l} \sum_{j=1}^l |x_{ij} - x_{kj}|.$$

Obliczeniowo prosta, co kiedyś było zaletą, niewrażliwa na wewnątrzpopulacyjny polimorfizm, waży tak samo określone różnice częstości alleli, bez względu na to, w której części zakresu (0, 1) występują wartości. Pozbawiona jakiegokolwiek interpretacji genetycznej czy ewolucyjnej, zdecydowanie nierealistyczna jako miara odrębności/czasu, który upłynął od rozdzielenia rozważanych taksonów, słabo nadaje się do rekonstrukcji filogenezy i rzadko jest obecnie stosowana. Niewiele bardziej skomplikowana jest **odległość** zaproponowana przez **Rogersa** (1972), czyli odległość Euklidesowa (*Euclidean distance*) skalowana liczbą *loci*:

$$D_{R(i,k)} = \frac{1}{L} \sum_{j=1}^L \sqrt{\sum_{j=1}^L (x_{ij} - x_{kj})^2} / 2$$

stosowana też w **modyfikacji Wrighta (1978)**:

$$D_{R/W(i,k)} = \sqrt{\frac{1}{2L} \sum_{j=1}^L (x_{ij} - x_{kj})^2}.$$

Prostota i oczywistość geometrycznej interpretacji nie kompensują tych samych wad, co wymienione dla odległości Prevostiego; ponadto wartości obliczonych odległości Wrighta silnie zależą od heterozygotyczności, czyli polimorfizmu wewnątrz taksonów – wartości obliczone pomiędzy dwoma taksonami o *locus* utrwalonym na określonym allelu, innym u każdego z taksonów, będą znacznie wyższe niż w przypadku polimorfizmu wewnątrz taksonów, nadal jednak niemających w tym *locus* wspólnego allelu (Wright 1978, Hillis 1984, Swofford i Olsen 1990, Swofford i inni 1996).

Również geometryczne, czyli pozbawione genetycznych założeń – choć z drugiej strony dla wielu sytuacji dobrze odzwierciedlające procesy ewolucyjne – są odległości zaproponowane przez Cavalli-Sforzę i Edwardsa (1967). Pierwsza z nich, częściej stosowana i na ogół uważana za lepszą, to **odległość łukowa (Cavalli-Sforza & Edwards arc distance)**:

$$D_{ARC(i,k)} = \sqrt{\frac{1}{l} \sum_{j=1}^l \left(\frac{2\theta}{\pi} \right)^2}, \text{ gdzie } \theta = \cos^{-1} \sum_{j=1}^l \sqrt{x_{ij}x_{kj}}.$$

Odległość ta jest niewrażliwa na wewnątrzpopulacyjny polimorfizm, osiągając maksimum równe 1 przy braku wspólnych alleli, bez względu na stopień zmienności wewnątrz rozpatrywanych taksonów. Częstości alleli są transformowane kątowno, dzięki czemu wariacje transformowanych częstości są niezależne od zakresów ich wartości. Zapewnia to standaryzację odległości w stosunku do losowego dryfu, tak że tempo wzrostu odległości w następstwie dryfu genetycznego jest niemal niezależne od początkowych częstości alleli. Odległość ta będzie więc realistycznie odzwierciedlać stopień zróżnicowania w warunkach, gdy jedynym, a przynajmniej niemal jedynym źródłem zróżnicowania jest dryf. Zarazem odległość ta dość szybko i mniej więcej liniowo rośnie przy małych różnicach, natomiast przy większych dość wcześnie jest bliska maksimum. Z natury swej lepiej nadaje się dla porównań poniżej poziomu gatunku. Podobne własności ma **odległość cięciwowa (Cavalli-Sforza & Edwards chord distance)**, obliczeniowo prostsza i będąca w sumie dość dobrym przybliżeniem łukowej:

$$D_{CHORD(i,k)} = 4 \left(L - \sum_{j=1}^L \sqrt{x_{ij}x_{kj}} \right).$$

Jeżeli zajmujemy się grupą populacji, których ostatni wspólny przodek żył stosunkowo niedawno – najczęściej będą to populacje należące do tego samego gatunku – i wykluczyć możemy mutacje jako źródło zróżnicowania, wówczas wygodnie jest użyć analizy wariancji (Falniowski, Mazan i Szarowska 1999) w ujęciu zaproponowanym przez Wrighta jako *F-statistics*. Dla grupy populacji (traktowanych jako jedna, podzielona populacja) obserwowane całkowite odchylenia od równowagi Hardy’ego i Weinberga (F_{IT} albo F w notacji Weira 1990) partycjonuje się na składnik wewnątrzpopulacyjny (F_{IS} albo f w notacji Weira) i międzypopulacyjny (F_{ST} albo θ w notacji Weira). Ten drugi może służyć jako miara zróżnicowania, znana jako **współczynnik wspólnego pochodzenia** (*coancestry coefficient*). Gdy w skończonych populacjach o ewolucyjnie efektywnej liczebności N , w których osobniki kojarzą się losowo (włączając w to losowy udział inbredu i/lub samozapłodnienia) i $F = \theta$, a n to liczebność próby (taka sama dla obu populacji):

$$\theta = \frac{\sum_{j=1}^l \left\{ \frac{1}{2} \sum_{m=1}^m (x_{ijm} - x_{kjm})^2 - \frac{1}{2(2n-1)} \left[2 - \sum_{m=1}^m (x_{ijm}^2 + x_{kjm}^2) \right] \right\}}{\sum_{j=1}^l \left(1 - \sum_{m=1}^m x_{ijm} x_{kjm} \right)},$$

(Reynolds, Weir i Cockerham 1983, z korektą błędu drukarskiego za Weirem 1990), gdzie m to liczba alleli w danym *locus* l . Jeżeli brak mutacji i na żadne z *loci* nie działa selekcja, to dla czasu t , który minął od chwili rozdzielenia rozpatrywanej pary populacji:

$$-\ln(1 - \theta) = t \ln \left(1 - \frac{1}{2N} \right) \approx \frac{t}{2N},$$

θ rośnie więc z upływem czasu, gdy ewolucyjnie efektywna liczebność populacji pozostaje stała. Gdy ta się zmienia, współczynnik odzwierciedla raczej nie czas, a kumulatywną wartość $1/N$. To zrozumiałe: współczynnik stworzono dla opisu narastającego zróżnicowania międzypopulacyjnego, zachodzącego wyłącznie w następstwie kumulowania genetycznego dryfu, a ten działa tym efektywniej, im mniejsza jest populacja. W wielkich populacjach działanie dryfu można wręcz pomijać, gdy w populacjach liczących niewiele osobników dryf stosunkowo szybko doprowadza do utrwalenia kolejnych loci na pojedynczych allelach (Hartl i Clark 1997). Gdy są one różne w różnych populacjach, θ osiąga maksimum równe 1. θ dobrze nadaje się jako odległość między bliskimi sobie populacjami, lecz nie powinien być stosowany w przypadkach, gdy niespełnione są założenia (zwłaszcza braku mutacji i selekcji), co nie zawsze łatwo ustalić, a dopiero znajomość N – albo wiedza o tym, że ewolucyjnie efektywna liczebność populacji pozostawała stała – umożliwi wykorzystanie tego współczynnika jako kwantyfikatora czasu, który upłynął od rozdzielenia populacji.

Alternatywnym estymatorem θ dla sytuacji, gdy dryf to jedyne źródło zmienności, jest zaproponowany przez Balakrishnana i Sanghvi (1968) **współczynnik G^2** , dla m alleli w *locus* l :

$$G^2 = \frac{1}{\sum_{j=1}^l (m-1)} \sum_{j=1}^l \sum_{m=1}^m \left[\frac{(x_{ijm} - x_{kjm})^2}{x_{ijm} + x_{kjm}} \right].$$

Choć jest to geometryczny współczynnik, $-\ln(1 - G^2)$ będzie estymował θ dla dużych liczebności próby n (Weir 1990). Najczęściej używaną odległością genetyczną jest zaproponowana przez Nei (1972; *Nei's genetic distance D*):

$$D = -\ln \left[\frac{\sum_{j=1}^l \sum_{m=1}^m x_{ijm} x_{kjm}}{\sqrt{\sum_{j=1}^l \sum_{m=1}^m x_{ijm}^2 \sum_{j=1}^l \sum_{m=1}^m x_{kjm}^2}} \right] = -\ln(I),$$

gdzie I to miara „genetycznej identyczności”, czyli podobieństwa (*Nei's genetic identity*), również szeroko stosowana. Nei (1978) zaproponował nieobciążony estymator (*Nei's unbiased distance/identity*), odpowiedni dla prób o małych liczebnościach (n – wielkość próby):

$$D = -\ln \left[\frac{(2n-1) \sum_{j=1}^l \sum_{m=1}^m x_{ijm} x_{kjm}}{\sqrt{\sum_{j=1}^l \left(2n \sum_{m=1}^m x_{ijm}^2 - 1 \right) \sum_{m=1}^l \left(2n \sum_{m=1}^m x_{kjm}^2 - 1 \right)}} \right] = -\ln(I).$$

Odległość Nei oblicza się rutynowo w niemal wszystkich pracach zamieszczających odległości genetyczne, najłatwiej więc porównywać szerokie zestawy danych, wykorzystując ten współczynnik, jednak budzi on szereg kontrowersji, a w wielu wypadkach niewątpliwie nie powinno się go stosować (Falniowski i inni 1993, 1996). Odległość ta nie jest metryką, nie spełnia warunku trójkąta, co już wyklucza ją, zdaniem niektórych, jako podstawę obliczania addytywnych drzew. Skalowana od zera do nieskończoności, słabo oddaje niewielkie różnice. Jej wartość zależy też silnie od wewnątrzpopulacyjnego polimorfizmu. Hillis (1984) przedstawił trzy hipotetyczne taksony dzielące identyczne częstości alleli w jednym *locus* i niemające wspólnych alleli w drugim, ale różniące się stopniem polimorfizmu tego drugiego. Obliczone wartości D mieściły się w przedziale $\langle 0,41, 1,10 \rangle$, co zupełnie nie mogło odzwierciedlać zróżnicowania czasu, który minął od ostatniego wspólnego przodka, czy choćby wielkości ewolucyjnego zróżnicowania. Aby wyeliminować tę wadę, Hillis (1984) zaproponował modyfikację odległości Nei:

$$D_H = -\ln \left(\frac{1}{L} \sum_{j=1}^l \frac{\sum_{m=1}^m x_{ijm} x_{kjm}}{\sqrt{\sum_{m=1}^m x_{ijm}^2 \sum_{m=1}^m x_{kjm}^2}} \right).$$

Najważniejszym ograniczeniem jest to, że odległość Nei, także w modyfikacji Hillisa, opisuje prawidłowo stopień zróżnicowania jedynie wówczas, gdy spełniony jest szereg założeń modelu. D stworzone zostało, by szacować liczbę substytucji kodonów przypadającą na *locus*, które miały miejsce od chwili rozdzielenia badanych taksonów (populacji). Podstawą modelu jest **teoria neutralna** ewolucji molekularnej (Kimura 1968, 1971, Kimura i Ohta 1983, Li i Graur 1991), zgodnie z którą zdecydowana większość alleli ma praktycznie taką samą wartość przystosowawczą, toteż na poziomie molekularnym dryf genetyczny z wolna eliminuje, w sposób losowy, różne allele. Kolejne mutacje, zachodzące z mniej więcej stałą i taką samą dla różnych *loci* częstością, prowadzą do powstania każdorazowo nowego, wcześniej nieobecnego allelu, o takiej samej wartości przystosowawczej jak wcześniejszy. Zarazem wyjściowo w populacjach występowała równowaga pomiędzy mutacjami a dryfem, a później ewolucyjnie efektywne liczebności populacji pozostawały stałe. Wówczas i tylko wtedy odległość Nei rośnie liniowo z upływem czasu od oddzielenia się badanych populacji. W praktyce można wątpić, czy powyższe warunki są kiedykolwiek spełnione. W każdym razie odległość ta powinna być stosowana dla kwantyfikowania długoterminowej ewolucji, obejmującej zarówno mutacje, jak i dryf, a i wówczas używać jej należy rozważnie.

2.12. Odległości dla sekwencji kwasów nukleinowych i białek

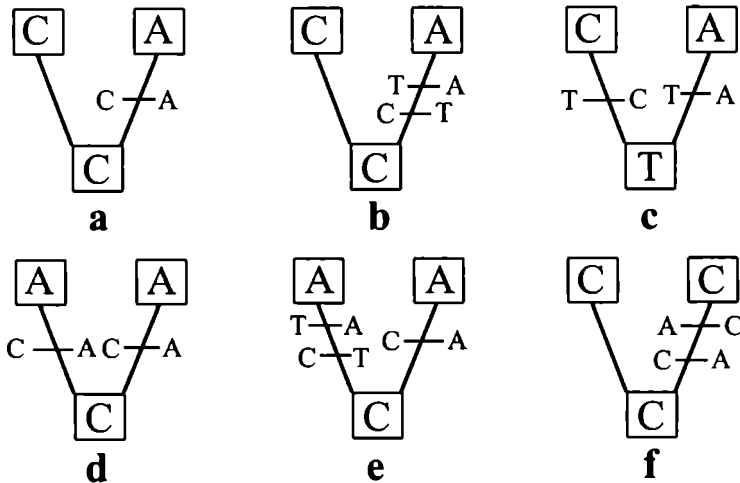
Intuicyjnie najbardziej oczywistą miarą **podobieństwa** bądź **odmienności** (tej ostatniej używać można jako odległości) pomiędzy dwiema **sekwencjami** DNA (ale również RNA czy białka) jest proporcja liczby, odpowiednio, takich samych bądź odmiennych nukleotydów (aminokwasów) do całkowitej długości zsekwencjonowanych molekuł bądź ich fragmentów. Oczywiście, jak już wcześniej wspomniano, porównywane sekwencje muszą być prawidłowo **współosiowane** (*aligned*), czego często wręcz nie da się przeprowadzić w sposób niebudzący wątpliwości. Warto też zaznaczyć, że współosiowanie omawia się jako poprzedzające dalsze etapy analizy, jak np. obliczanie odległości, jednak w rzeczywistości najlepiej przeprowadzać je jednocześnie, iteratywnie. Odległość mierzona proporcją różnych do wspólnych elementów na odpowiadających sobie pozycjach sekwencji, znana jako **odległość Hamminga** (*p-distance*), dobrze określa chemiczną odmienność molekuł, jest jednak nieodpowiednia jako miara czasu czy choćby jedynie ewolucyjnego zróżnicowania molekuł od momentu oddzielenia linii organizmów, z których pochodzą. Z jednej strony nie każda zmiana nukleotydu (w mniejszym stopniu także aminokwasu) ma takie samo znaczenie biologiczne czy zachodzi równie łatwo, z drugiej – co jest tym częstsze, im odleglejsze ewolucyjnie organizmy porównujemy – występowanie tego samego nukleotydu (aminokwasu) nie musi oznaczać, że w tej pozycji nic się nie działo: kolejna mutacja mogła przywrócić stan pierwotny.

Gdy ewolucja genu obejmowała insercje i/lub delecje – co jest regułą, gdy ostatni wspólny przodek rozpatrywanych organizmów występował dawno – współosiowanie obejmuje **luki** (*gaps*), wstawiane tak, aby homologiczne nukleotydy (aminokwasy) obu sekwencji leżały w odpowiadających sobie miejscach. Warto podkreślić, że na etapie współosiowania przyjąć musimy arbitralnie **koszt luk** (*gap penalty*), co wpływa na wyniki współosiowania. Współosiowanie to etap, będący potencjalnie największym źródłem błędów zrekonstruowanych filogenez – tym trudniejszy, im odleglejsze organizmy porównujemy. Bogatą literaturę omawiającą współosiowanie zestawia Classification Society of North America na swojej stronie internetowej (<http://www.pitt.edu/csna/>). Problemy ze współosiowaniem są jeszcze poważniejsze, gdy współosiujemy nie dwie, lecz więcej sekwencji (*multiple alignment*), co w praktyce najczęściej ma miejsce. Można wówczas posłużyć się **techniką sumy** wszystkich możliwych **par** sekwencji (przypominamy: dla n sekwencji będzie $(n - 1)/2$ par), ale ta technika (*sum of pairs alignment*) nie poddaje się sensownej biologicznej interpretacji. Dla sekwencji równie blisko (przypuszczalnie) spokrewnionych prawidłowa jest **technika gwiazdy** (*star alignment*), jednak zwykle badane organizmy są sobie bliższe lub dalsze, toteż najlepiej stosować **technikę drzewa** (*tree alignment*), czyli obliczać koszt współosiowania danej pozycji sekwencji na podstawie zrekonstruowanego drzewa. Oczywiście drzewo obliczymy na podstawie porównania już współosiowanych sekwencji, więc do współosiowania możemy dysponować co najwyżej wstępną, „roboczą” filogenezą, lecz to właśnie wskazuje na odpowiedniość iteratywnych technik współosiowania/rekonstrukcji filogenezy dla sekwencji.

Lukę można teoretycznie uznać za „piąty nukleotyd” („21. aminokwas”), jednak procesy odpowiedzialne za podstawienie nukleotydu (aminokwasu) są ewolucyjnie i chemicznie inne niż te, w wyniku których powstają insercje czy delecje, mają też miejsce częściej – zatem ich koszt jest niższy – toteż takie podejście nie wydaje się właściwe. Brak sensownego sposobu traktowania luk, więc najlepiej te rejony pominąć przy porównywaniu sekwencji (Swofford i inni 1996). Metody są dwie. Gdy luki są krótkie i rozmieszczone mniej więcej losowo, odpowiedniejsze jest **pomijanie parami** (*pairwise deletion*), czyli niebranie pod uwagę odcinków, które są lukami w jednej lub obu sekwencjach, przy obliczaniu odległości dla danej pary sekwencji (Kumar i inni 1993). **Całkowite pominięcie** (*complete deletion*) polega na nieuwzględnianiu miejsc, w których wystąpiły luki w którejkolwiek z wszystkich sekwencji, pomiędzy którymi obliczamy odległości, a więc także dla odległości pomiędzy sekwencjami, w których nie występują w danym miejscu luki. Całkowite pominięcie powoduje oczywiście pewną utratę informacji, jednak jest wskazane, gdy dotyczy regionów szczególnie zmiennych, szczególnie podatnych na insercje i delecje – pomijanie parami takich regionów spowoduje błąd systematyczny, wyrażający się zawyżaniem podobieństwa, a więc obniżaniem odległości dla niektórych par sekwencji. Jeżeli jakieś odcinki sekwencji obfitują w luki, to współosiowanie zawsze będzie wątpliwe, a dla rekonstrukcji filogenezy całe takie odcinki należy pominąć.

W toku ewolucji, w miarę upływu czasu, następuje **akumulacja mutacji**. Odległości Hamminga odzwierciedlają ten proces przy niewielkich różnicach (po krótkim czasie akumulacji mutacji). W miarę wzrostu różnic (upływu czasu) będzie to jednak coraz gorsze, coraz bardziej zaniżone przybliżenie. Gdy w danej pozycji następuje zamiana nukleotydu (aminokwasu), sekwencja w tym miejscu różni się od innych. Dalsze mutacje w tym miejscu już tej różnicy zwiększyć nie mogą, mogą jedynie ją utrzymać

bądź zmniejszyć, poprzez przywrócenie stanu pierwotnego albo równoległe mutacje w obu porównywanych molekułach. Jeżeli przykładowo w jednej sekwencji mamy A, w drugiej C, to postulujemy mutację $C \rightarrow A$ (lub $A \rightarrow C$, gdy nie znamy sekwencji przodka), czyli **pojedyncze podstawienie** (*single substitution*: Ryc. 2.2.a), gdy w rzeczywistości mogło być np. $C \rightarrow T \rightarrow A$, czyli **wielokrotne podstawienie** (*multiple substitution*: Ryc. 2.2.b), bądź $T \rightarrow C$ w jednej sekwencji, a $T \rightarrow A$ w drugiej, czyli **jednoczesne podstawienia** (*coincidental substitution*: Ryc. 2.2.c). Z kolei identyczność nukleotydu (aminokwasu) na danej pozycji porównywanych sekwencji może odzwierciedlać homologię, czyli brak zmian w tym miejscu, ale też może być następstwem dwóch zmian tworzących **równoległą substytucję** (*parallel substitution*: Ryc. 2.2.d), trzech zmian przynoszących **zbieżną, czyli konwergentną substytucję** (*convergent substitution*: Ryc. 2.2.e), bądź dwóch zmian w jednej z gałęzi dających **wsteczne podstawienie** $A \rightarrow C \rightarrow A$ (*back substitution*: Ryc. 2.2.f). Nukleotydów jest zaledwie cztery, więc prawdopodobieństwo **wstecznej mutacji** bądź mutacji równoległej musi być wysokie, a zwykle zliczenie pozycji, na których występują różnice pomiędzy sekwencjami, dać musi zawsze zaniżony estymat zmian ewolucyjnych, jakie miały miejsce od czasu występowania ostatniego wspólnego przodka organizmów, z których pochodzą porównywane molekuły.



Ryc. 2.2. Gdy na homologicznych pozycjach porównywanych sekwencji nukleotydów mamy w jednej z sekwencji C, a w drugiej A, to może to być następstwem pojedynczego podstawienia (a), lecz także wielokrotnego podstawienia (b) albo jednoczesnego podstawienia (c). Gdy nukleotyd w obu sekwencjach jest identyczny, to dowodzić to może braku jakichkolwiek zmian na tej pozycji w obu sekwencjach, lecz również wynikać może z równoległego podstawienia (d), zbieżnego, czyli konwergentnego podstawienia (e), lub wstecznego podstawienia (f)

W sumie więc, w miarę akumulacji mutacji, obserwowane różnice rosną wprawdzie szybko, później coraz wolniej, by w końcu asymptotycznie zbliżyć się do stałej, nieprzekraczalnej wartości, nawet gdy są porównywane bardzo różne odcinki czasu. Mówimy wówczas o **nasyceniu**, czyli **saturacji** (*saturation*) różnic. Często tranzycje występują znacznie częściej niż transwersje (zwłaszcza w mitochondrialnym DNA), więc

różnice pomiędzy dwiema sekwencjami będące następstwem tranzycji dość szybko osiągają stan nasycenia, zatem dalsze tranzycje nie zwiększają już różnicy. Wówczas tranzycje wnoszą niewiele „sygnału filogenetycznego”, a raczej zwiększają jedynie „szum filogenetyczny”, czyli wariancję rekonstruowanych filogenez, najlepiej więc wtedy obliczać odległości jedynie na podstawie transwersji. Co oczywiste, saturacja następuje tym szybciej, im wyższa jest częstość mutacji, różna dla różnych fragmentów DNA. Warto jednak pamiętać, że dla odcinków kodujących białka wiele podstawień będzie letalne, możliwe więc będą substytucje jedynie w pewnych pozycjach, a im więcej pozycji nie podlegających akumulowanym mutacjom, tym szybciej dochodzi do saturacji. Bywa więc tak, że odcinek DNA, charakteryzujący się bardzo niskim tempem mutacji, zachodzących jednak w (niemal) wszystkich pozycjach łańcucha, osiągnie w tym samym czasie większe zróżnicowanie między sekwencjami niż ewoluujący znacznie szybciej odcinek, dla którego, powiedzmy, połowa nukleotydów nie może podlegać substytucjom, więc szybko doszło do saturacji. Stan nasycenia występuje po różnym czasie dla różnych fragmentów DNA, toteż nie ma fragmentów „uniwersalnych”, nadających się do rekonstrukcji pokrewieństw od poziomu populacji po gromady czy typy: to, jakiego fragmentu DNA najlepiej użyć, zależy od rangi systematycznej porównywanych taksonów.

Dla celów analizy filogenetycznej konieczna jest więc **kompensacja** tych **nieobserwowalnych mutacji**, co próbuje się osiągnąć różnymi metodami (Kumar i inni 1993, Swofford 1996, Swofford i inni 1996). Zakłada się różne modele ewolucji, przy czym wychodzi się z macierzy różnic F_{XY} , zestawiającej częstości wszystkich kombinacji nukleotydów (to samo robi się dla aminokwasów, wówczas macierz jest nie 4×4 , lecz 20×20) dla danej pary sekwencji X i Y :

$$F_{XY} = \begin{pmatrix} n_{AA}/N & n_{AC}/N & n_{AG}/N & n_{AT}/N \\ n_{CA}/N & n_{CC}/N & n_{CG}/N & n_{CT}/N \\ n_{GA}/N & n_{GC}/N & n_{GG}/N & n_{GT}/N \\ n_{TA}/N & n_{TC}/N & n_{TG}/N & n_{TT}/N \end{pmatrix}, \text{ a prościej: } F_{XY} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \\ m & n & o & p \end{pmatrix}$$

gdzie n_{ik} to liczba przypadków, gdy stanowi i w sekwencji X odpowiadał stan k w sekwencji Y , $N = \sum n_{ik}$, A – adenina, C – cytozyna, G – guanina, T – tymina. Jeżeli rozpatrujemy model ewolucji sekwencji, to zamiast macierzy obserwowanych różnic rozpatrujemy macierz prawdopodobieństw określonych zmian P_{XY} . Odległość Hamminga, czyli zwyczajny, nieskorygowany współczynnik niepodobieństwa D (*dissimilarity* D), znany też jako odległość p (*p-distance*), wyrazić więc można:

$$D = p\text{-distance} = b + c + d + e + g + h + i + j + l + m + n + o = 1 - (a + f + k + p).$$

Wartość D koryguje się rozmaicie, w zależności od przyjętego modelu ewolucji, jak bowiem wiemy, w każdym przypadku wartość ta jest zaniżona. Ogólnie przyjmuje się model Markova, w którym podstawienie jednego nukleotydu innym nie zależy od wcześniejszej ewolucji tej pozycji w badanej sekwencji. Najprostszy jest jednoparametrowy model Jukesa i Cantora (1969; Li i Graur 1991, Page i Holmes 1998), zakładający, że prawdopodobieństwo podstawienia jakiegokolwiek nukleotydu którymkolwiek

wiek z pozostałych trzech jest zawsze takie samo, czyli częstość podstawień dla wszystkich par nukleotydów jest taka sama. Wówczas:

$$JC_{distance} = -\frac{3}{4} \ln \left(1 - \frac{4}{3} D \right); \text{ a macierz prawdopodobieństw: } \mathbf{P}_{XY} = \begin{matrix} & \cdot & \alpha & \alpha & \alpha \\ \alpha & \cdot & \alpha & \alpha & \\ \alpha & \alpha & \cdot & & \alpha \\ \alpha & \alpha & & & \alpha \end{matrix}$$

Największa oczekiwana wartość JC równa jest 0,75, zgodnie z modelem; jeżeli JC tę wartość przekroczy, to odległość staje się nieokreślona, jako logarytm liczby ujemnej. Model zakłada też jednakowe częstości (1/4) wszystkich czterech nukleotydów. Felsenstein (1981) zaproponował model, pozwalający na różne częstości ($\pi_A, \pi_C, \pi_G, \pi_T$) nukleotydów:

$$F81_{distance} = -B \ln \left(1 - \frac{D}{B} \right), \text{ dla: } B = 1 - (\pi_A^2 + \pi_C^2 + \pi_G^2 + \pi_T^2); \mathbf{P}_{XY} = \begin{matrix} & \cdot & \pi_C \alpha & \pi_G \alpha & \pi_T \alpha \\ \pi_A \alpha & \cdot & \pi_G \alpha & \pi_T \alpha & \\ \pi_A \alpha & \pi_C \alpha & \cdot & \pi_T \alpha & \\ \pi_A \alpha & \pi_C \alpha & \pi_G \alpha & \cdot & \end{matrix}$$

Jak łatwo zauważyć, odległość JC jest przypadkiem szczególnym $F81$ dla $B = 3/4$, co zachodzi, gdy $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$. $F81$ jest formułą nadającą się i dla sekwencji białek. Przy takich samych frekwencjach $B = 19/20$, przy zróżnicowanych

$$B = 1 - \sum_{i=1}^{20} n_i^2,$$

gdzie n_i to częstość określonego aminokwasu (Swofford i inni 1996).

Model Jukes-Cantora jest najprostszy, dla większości rzeczywistych danych zbyt prosty. Kimura (1980) zaproponował model dwuparametrowy (Li i Graur 1991, Page i Holmes 1988), zakładający inne tempo dla tranzycji, inne dla transwersji. Porównując więc dwie sekwencje, dla każdej pary sekwencji oddzielnie, osobno liczymy różnicę powstałe w wyniku tranzycji, zachodzących z prawdopodobieństwem α : $P = c + h + i + n$, osobno w następstwie transwersji zachodzących z prawdopodobieństwem β : $Q = b + d + e + g + j + l + m + o$. Wówczas:

$$K2P_{distance} = \frac{1}{2} \ln \left(\frac{1}{1 - 2P - Q} \right) + \frac{1}{4} \ln \left(\frac{1}{1 - 2Q} \right); \mathbf{P}_{XY} = \begin{matrix} & \cdot & \beta & \alpha & \beta \\ \beta & \cdot & \beta & \alpha & \\ \alpha & \beta & \cdot & \beta & \\ \beta & \alpha & \beta & \cdot & \end{matrix}$$

Model ten zakłada, jak i JC , takie same częstości nukleotydów. Założenie to – że frekwencja każdego z nukleotydów zbliża się do 0,25 – często odbiega od rzeczywistości: przykładowo łączna częstość nukleotydów G i C , zamiast spodziewanej wartości 0,5,

jest zbliżona do 0,1 w mitochondrialnym DNA *Drosophila* (Wolstenholme 1992). Dla takich przypadków Tamura (1992) sformułował odległość i model:

$$d_{Tamura} = -h \ln(1 - P/h - Q) - \frac{(1-h) \ln(1-2Q)}{2}; \mathbf{P}_{XY} = \begin{matrix} & \beta\theta_2 & \beta\theta_1 & \alpha\theta_1 \\ \beta\theta_2 & \cdot & \alpha\theta_1 & \beta\theta_1 \\ \beta\theta_2 & \alpha\theta_2 & \cdot & \beta\theta_1 \\ \alpha\theta_2 & \beta\theta_2 & \beta\theta_1 & \cdot \end{matrix}$$

gdzie: $h = 2\theta_1(1-\theta_1)$, $\theta_1 = \pi_C + \pi_G$ (łączna częstość C i G), a $\theta_2 = \pi_A + \pi_T$ (łączna częstość A i T). Uogólnieniem modelu *K2P* jest *HKY85*, wprowadzone przez Hasegawę, Kishino i Yano (1985), pozwalające na zróżnicowanie częstości nukleotydów (π_A , π_C , π_G , π_T):

$$\mathbf{P}_{XY} = \begin{matrix} & \pi_C\beta & \pi_G\alpha & \pi_T\beta \\ \pi_A\beta & \cdot & \pi_G\beta & \pi_T\alpha \\ \pi_A\alpha & \pi_C\beta & \cdot & \pi_T\beta \\ \pi_A\beta & \pi_C\alpha & \pi_G\beta & \cdot \end{matrix}$$

Brak prostego wzoru na obliczanie odległości, podobnie jak i dla szeregu innych, bardziej złożonych i dzięki temu bardziej uniwersalnych modeli (Swofford i inni 1996). Istnieje natomiast prosty wzór na odległość *F84*, zbliżoną do uogólnionej *K2P* (Tateno i inni 1994, Swofford i inni 1996):

$$F84_{distance} = -2A \ln \left(1 - \frac{P}{2A} - \frac{(A-B)Q}{2AC} \right) + 2(A-B-C) \ln \left(1 - \frac{Q}{2C} \right),$$

gdzie P i Q są tym samym co w *K2P*, $A = \pi_C\pi_T / (\pi_C + \pi_T) + \pi_A\pi_G / (\pi_A + \pi_G)$, $B = \pi_C\pi_T + \pi_A\pi_G$, a $C = (\pi_A + \pi_G)(\pi_C + \pi_T)$. Podobnie uogólnieniem modelu – i odległości – *HKY85* jest sformułowany przez Tamurę i Neiego (1993) wzór na odległość *TN* (Nei i Kumar 2000):

$$TN_{distance} = -\frac{2\pi_A\pi_G}{\pi_R} \ln \left[1 - \frac{\pi_R}{2\pi_A\pi_G} P_1 - \frac{1}{2\pi_R} Q \right] - \frac{2\pi_T\pi_C}{\pi_Y} \ln \left[1 - \frac{\pi_Y}{2\pi_T\pi_C} P_2 - \frac{1}{2\pi_Y} Q \right] - 2 \left[\pi_R\pi_Y - \frac{\pi_A\pi_G\pi_R}{\pi_R} - \frac{\pi_T\pi_C\pi_R}{\pi_Y} \right] \ln \left[1 - \frac{1}{2\pi_R\pi_Y} Q \right],$$

gdzie P_1 i P_2 to odpowiednio proporcje różnic tranzycyjnych między A i G oraz T i C, $\pi_R = \pi_A + \pi_G$, a $\pi_Y = \pi_C + \pi_T$. Rodriguez i inni (1990) sformułowali ogólny, odwracalny w czasie (*general time reversible – GTR* albo *REV*) model, dla którego odległość dana jest wzorem (Rodriguez i inni 1990, Yang i inni 1994, Swofford 1996):

$$GTR_{distance} = -tr \left[\prod \ln \left(\prod^{-1} F_{XY} \right) \right], \mathbf{P}_{XY} = \begin{matrix} & \pi_{Ca} & \pi_{Gb} & \pi_{Tc} \\ \pi_{Aa} & \cdot & \pi_{Gd} & \pi_{Te} \\ \pi_{Ab} & \pi_{Cd} & \cdot & \pi_{Tf} \\ \pi_{Ac} & \pi_{Ce} & \pi_{Gf} & \cdot \end{matrix}$$

gdzie tr to ślad macierzy, Π to diagonalna macierz średnich frekwencji nukleotydów w sekwencjach X i Y . Dokładniejsze omówienie tej odległości wymaga znajomości rachunku macierzowego, obliczeń wartości własnej i wektora własnego (pojęcia wartości własnej i wektora własnego przedstawimy w Części 3, poświęconej analizie fenetycznej); a, b, c, d, e, f to kolejne prawdopodobieństwa dla każdej z sześciu możliwych substytucji. GTR wykorzystywana jest w niektórych programach do rekonstrukcji filogenezy.

Omówione powyżej modele to oczywiście zaledwie część dotąd zaproponowanych. Im prostszy model, tym mniej dokładnie odzwierciedla rzeczywiste procesy ewolucyjne. Z drugiej strony, w miarę komplikacji modeli – dla wierniejszego odzwierciedlenia rzeczywistości – wprowadza się kolejne parametry, a dla każdego z nich należy przyjąć jakąś wartość. Zważywszy, że wartości te albo przyjmujemy „na wycucie”, albo też na podstawie posiadanych danych, czyli na podstawie wyniku procesów ewolucyjnych, których przebieg pragniemy zrekonstruować – łatwo tu więc o logikę błędnego koła – parametry te będą źródłem błędu. Mamy więc wybór pomiędzy estymatami odległości zbyt prostymi a obciążonymi. I tutaj należy zachować zdrowy rozsądek, czyli stosować model możliwie najprostszy, lecz nie za prosty. Istnieją oczywiście formalne procedury doboru parametrów: rozpoczynając od modelu najprostszego, kolejno zwiększa się jego komplikację, każdorazowo sprawdzając poprawę dobroci odwzorowania, aż poprawy nie ma lub jest bardzo nieznaczna. Wrócimy do tego jeszcze w Rozdziale 4.6, omawiającym techniki **maksymalizacji wiarygodności** (*maximum likelihood*). **Wiarygodność** (*likelihood*) dana jest wzorem:

$$L = \Pr(D|H),$$

czyli równa jest prawdopodobieństwu \Pr otrzymania danych D , gdy założymy hipotezę H . Im większa wartość L , tym bardziej zgodne są nasze dane z założoną hipotezą; alternatywne modele – jak powyższe, opisujące proces powstawania różnic pomiędzy sekwencjami – możemy wartościować, obliczając L dla naszych danych i kolejnych modeli. W praktyce wartości L bywają bardzo małe, nieraz poza zakresem kalkulatorów, toteż wygodnie jest używać ich naturalnych logarytmów (*log-likelihoods*). Gdy kolejne zdarzenia opisywane przez model traktujemy jako niezależne, prawdopodobieństwo ich łącznego zachodzenia równe jest iloczynowi prawdopodobieństw poszczególnych zdarzeń, co odpowiada sumowaniu ich logarytmów. Warto podkreślić, że techniki maksymalizacji wiarygodności pozwalają na precyzyjną i obiektywną ocenę zgodności naszych danych z proponowanym modelem, nie mówią natomiast nic o wiarygodności czy prawdopodobieństwie samego modelu (patrz Rozdział 4.6).

Przedstawione wyżej modele, nawet te najbardziej skomplikowane, przyjmują szereg upraszczających założeń: (1) zmiany na każdej pozycji następują niezależnie od pozostałych pozycji, (2) częstość substytucji jest stała w czasie i dla wszystkich linii

filogenetycznych, (3) częstości poszczególnych nukleotydów są w równowadze – niezmienne w czasie i pomiędzy liniami filogenetycznymi, (4) warunkowe prawdopodobieństwa substytucji są takie same dla wszystkich pozycji sekwencji i nie zmieniają się w czasie. Założenia te pozwalają na stosunkowo prosty opis matematyczny ewolucji DNA, ale często są nierealistyczne. Znanych jest wiele odstępstw od niezależności zmian na jednej pozycji w stosunku do zmian na pozycjach pozostałych, jak choćby w przypadku rybosomalnego RNA, którego drugorzędowa struktura stabilizowana jest parowaniami Watsona-Cricka. Jeżeli nastąpi podstawienie – zmiana zasady – na prostym odcinku cząsteczki, to struktura stanie się przestrzennie niestabilna i z dużym prawdopodobieństwem można spodziewać się podstawienia kompensacyjnego, przywracającego wiązanie parowaniem zasad. Inna sprawa, że taka kompensacja zajść może dopiero po milionach lat (Page i Holmes 1998). Częstość substytucji bynajmniej nie jest stała, nawet dla tego samego genu, w różnych liniach filogenetycznych (Avice 2000); im odleglejsze filogenetycznie grupy rozpatrujemy, tym większych różnic tempa mutacji możemy się spodziewać.

Kolejnym problemem jest filogenetyczne zróżnicowanie częstości poszczególnych nukleotydów – jak wiadomo, bardziej złożone modele ewolucji sekwencji pozwalają na zróżnicowanie częstości nukleotydów (π_A , π_C , π_G , π_T), jednak i one zakładają, że częstości te nie ulegają zmianom zależnie od taksonu, a pozostają mniej więcej stałe we wszystkich analizowanych sekwencjach. Tymczasem częstości nukleotydów podlegają znacznym wahaniom pomiędzy grupami, toteż może się zdarzyć, że użycie którejs z omówionych powyżej odległości dla rekonstrukcji filogenezy da w wyniku drzewo, na którym zgrupowane obok siebie zostaną taksony o podobnych częstościach poszczególnych zasad, choć filogenetycznie odległe. Metodą, która wprawdzie nie ocenia bezpośrednio liczby podstawień nukleotydów, ale daje estymaty odrębności sekwencji niezależne od ewentualnego zróżnicowania częstości nukleotydów pomiędzy taksonami, jest zaproponowana przez Steela (1994a) oraz Lockharta i innych (1994) technika **transformacji logarytmiczno-wyznacnikowej** (*log-determinant – LogDet*). Dla zdefiniowanej wcześniej macierzy $r \times r$ różnic F_{XY} :

$$d_{XY} = -\ln(\det F_{XY}), \text{ gdzie } \det F_{XY} \text{ to wyznacznik macierzy } F_{XY}.$$

Transformacja *LogDet* powinna prawidłowo odzwierciedlać zróżnicowanie dla każdego modelu zgodnego z modelem Markowa, jak długo częstość podstawień jest identyczna dla wszystkich pozycji, ewoluujących tak samo i niezależnie od siebie. Wszystkie 12 substytucji (wliczając brak różnic, czyli brak substytucji, a także traktując odrębnie podstawienia w obu kierunkach, a więc np. $\pi_{AC} \neq \pi_{CA}$) może zachodzić z dowolną częstością, która może się zmieniać w różnych częściach drzewa filogenetycznego – na różnych gałęziach lub w różnych częściach tej samej gałęzi. Dla ewolucji zachodzącej w ten sposób zdefiniowane powyżej odległości będą rosły z upływem czasu od oddzielenia linii filogenetycznych i można ich użyć do konstrukcji addytywnych drzew, co nie zmienia faktu, że nie są to odległości ewolucyjne, tzn. estymujące liczbę podstawień nukleotydów w przeliczeniu na pozycję. Dla **modeli stacjonarnych** (*stationery*), czyli zakładających niezmiennność częstości nukleotydów, odległości te można odpowiednio skalować (Swofford i inni 1996):

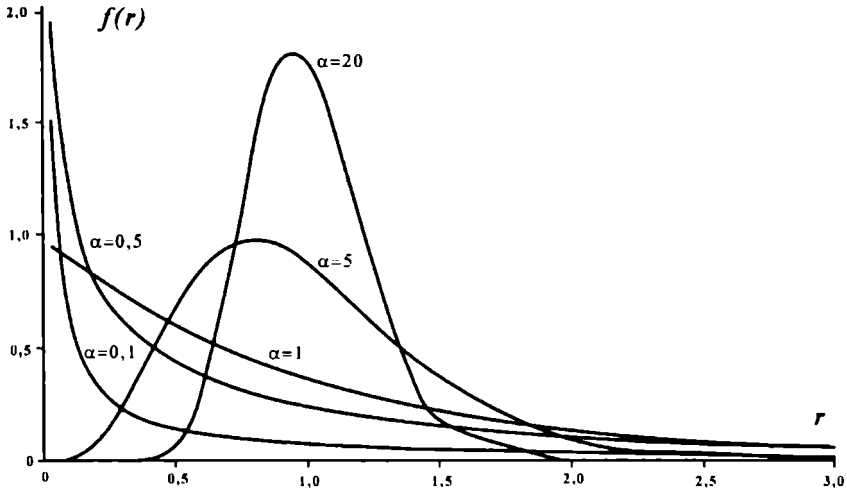
$$d_{XY_{stat}} = \frac{-\ln(\det F_{XY}) + 1/2 \ln(\det \Pi_X \Pi_Y)}{r} = -\frac{1}{r} \ln \left(\frac{\det F_{XY}}{\sqrt{\det \Pi_X \Pi_Y}} \right),$$

gdzie Π_X , Π_Y to macierze częstości stanów cech (dla DNA – nukleotydów) w sekwencjach X i Y , a r to liczba stanów cech (oczywiście dla kwasów nukleinowych $r = 4$). Im bardziej różnią się częstości nukleotydów pomiędzy sekwencjami, tym bardziej zawyżone będą wartości tak skorygowanej odległości, choć niekiedy mogą też być zaniżone. Pomimo to odległości transformowane logarytmicznie-wyznacznikowo wydają się i tak mniej wrażliwe na zróżnicowanie częstości nukleotydów niż odległości omówione wcześniej, a przynajmniej nie gorsze od nich, jak długo sekwencje nie są krótsze niż 200 par zasad (Swofford i inni 1996: tamże obszerniejsza dyskusja).

Częstość substytucji jest różna dla różnych fragmentów genomu, będąc np. bardzo wysoka dla pseudogenów czy intronów, a skrajnie niska dla fragmentów kodujących białka ważne dla procesów metabolicznych, gdzie niemal jakakolwiek mutacja jest letalna lub przynajmniej skrajnie niekorzystna. Dla sekwencji kodujących białka wiele z podstawień nukleotydów na trzecich pozycjach kodonów to mutacje ciche, niepowodujące zmiany kodowanego białka. Losowo i szybko zachodzące zmiany na trzecich pozycjach wykluczają więc ich użyteczność dla rekonstrukcji pokrewieństw odleglejszych organizmów. Zarazem częstości nukleotydów na trzecich pozycjach kodonów wykazują zróżnicowanie pomiędzy niektórymi gatunkami, co może wskazywać na oddziaływanie selekcji, zróżnicowanej między taksonami – a więc sugeruje nierzadkość konwergencji występowania określonych nukleotydów na tych pozycjach. W każdym razie częstość i rodzaj zmian nukleotydów na trzecich pozycjach są inne niż na dwóch pierwszych, a więc nie zostaje zachowany warunek takiego samego prawdopodobieństwa zmiany na każdej pozycji sekwencji. Jeżeli organizmy, z których pochodzą sekwencje, są sobie bliskie, większość różnic (bądź wszystkie) dotyczyć będzie trzeciej pozycji, toteż dla obliczeń odległości najlepiej uwzględnić całe sekwencje. Przy organizmach odległej spokrewnionych celowe jest uwzględnianie jedynie pierwszych i drugich pozycji kodonów bądź „przepisanie” sekwencji DNA na sekwencje aminokwasów i obliczanie odległości pomiędzy białkami. Można też osobno policzyć ciche mutacje, osobno pozostałe podstawienia nukleotydów. Następnie, dla eliminacji różnic częstości mutacji pomiędzy nukleotydami, użyć wyłącznie cichych mutacji – jako zasadniczo neutralnych – dla obliczeń odległości pomiędzy bliskimi taksonami, a jedynie pozostałych – dla obliczeń dla taksonów odleglejszych, eliminując w ten sposób „szum filogenetyczny”.

Prawdopodobieństwa zastąpienia jednego nukleotydu innym wykazują szeroki zakres zmienności pomiędzy kolejnymi parami zasad w obrębie danej sekwencji, a także różne zakresy i rozkłady tej zmienności dla różnych sekwencji (Yang 1996). Rozkłady częstości tych prawdopodobieństw opisuje się najczęściej, korzystając z tzw. **rozkładu gamma** (Ryc. 2.3), gdzie oś odciętych odpowiada częstości substytucji – częstości zastępowania jednego nukleotydu innym na danej pozycji sekwencji, w ciągu miliarda lat – a oś rzędnych odpowiada proporcji par zasad dla których zachodzi ta częstość. Kształt rozkładu określa parametr α , dla jego wartości poniżej 1 rozkład jest L-kształtny, powyżej 1 krzywa staje się dzwonoвата. W rzeczywistych sekwencjach jądrowych i mitochondrialnych wartości α zawierają się w zakresie 0,16–1,37, najczęściej więc są to rozkłady L-kształtne; dla trzecich pozycji kodonów α osiąga wartość

1,58 (a 0,08–0,18 dla pozycji pozostałych: Yang 1996). Niskie wartości α odpowiadają dużej zmienności tempa substytucji, gdy wyższe mniejszej, by dla $\alpha \approx \infty$ prawdopodobieństwo zmiany nukleotydu osiągało tę samą wartość dla wszystkich par zasad ($f(r) = 2/2$) w sekwencji. Jest tak, bowiem współczynnik α równy jest odwrotności kwadratu współczynnika zmienności tempa substytucji.



Ryc. 2.3. Rozkład gamma. Na osi odciętych przedstawiono częstość substytucji r (podstawienie/nukleotydy/ 10^9 lat), na osi rzędnych $f(r)$, czyli udział par zasad, dla których zachodzi dana częstość substytucji. Kształt rozkładu określa parametr α : dla $\alpha < 1$ rozkład jest L-kształtny, czyli częstość substytucji jest bardzo zróżnicowana: wysoka dla nielicznych pozycji, a niska dla większości. Wyższe wartości α charakteryzuje rozkład zbliżony do normalnego, dzwonokształtnego, gdy dla α bliskiego nieskończoności częstość substytucji jest dla wszystkich par zasad taka sama. Wg Yanga (1996)

W przypadku zróżnicowanego tempa substytucji „normalne” nieskorygowane odległości przedstawione wcześniej dadzą zaniżone oceny ewolucyjnych zmian. Aby tego uniknąć, dla szeregu modeli – jak *JC* czy *K2P* – istnieją modyfikacje, uwzględniające parametr α . Wystarczy we wzorach podanych wyżej zastąpić funkcję $\ln(x)$ wyrażeniem: $\alpha(1 - x^{-1/\alpha})$ (Swofford i inni 1996). Podobnie zmodyfikować można i uogólniony model odwracalny w czasie (*GTR*). Wartość α należy wpierv obliczyć za pomocą odpowiedniego programu lub wykorzystać dane z literatury dla podobnych fragmentów genomu u blisko spokrewnionych organizmów. Gamma-modyfikowane odległości *GTR* oraz odległości transformowane logarytmicznie-wyznacznikowo uznaje się za najbardziej uniwersalne. Uniwersalność zwykle jednak znów pociąga za sobą zwiększenie wariacji, a symulacje wykazały, że najprostsze modele często dają poprawną rekonstrukcję filogenezy – zwłaszcza dla krótkich sekwencji – pomimo niespełniania założeń modeli, na których są oparte (Swofford i inni 1996; patrz też Rozdział 4.6). W każdym razie jeżeli dwie odległości dadzą podobne wartości, należy zawsze wybrać prostszą, jako mającą mniejszą wariację (Kumar i inni 1993).

Dla sekwencji białka prostą miarą odmienności jest zaproponowana przez Kimurę (Page i Holmes 1998):

$$d = -\ln(1 - p - 0,2p^2),$$

gdzie p to proporcja pozycji, na których aminokwasy w jednej sekwencji są inne niż w drugiej. Jest to prosta, lecz bardzo niedoskonała odległość, nieoparta na ewolucyjnie realistycznych założeniach.

3. Analiza fenetyczna

3.1. Statystyczna analiza wielowymiarowa – wprowadzenie, macierze, rozkłady, jednorodność

Jak już wspominaliśmy w Rozdziałach 1.5 i 2.2, bez prawidłowo rozpoznanych homologii wynikiem analiz będzie systematyka fenetyczna, niezależnie od tego, jakiej techniki użyjemy. Wiemy też, że poprawne rozpoznanie homologii jest trudne, a często bywa zwyczajnie niemożliwe, zwłaszcza w przypadku słabiej poznanych grup, jak choćby większość bezkręgowców. Wówczas podstawą klasyfikacji będą nie pokrewieństwa, lecz **ogólne podobieństwo** (*overall similarity*), tak jak zakłada **taksonomia fenetyczna**. Odsyłając do bogatej literatury, przedstawiającej argumenty za i przeciw podejściu fenetycznemu (Sokal i Sneath 1963, Hennig 1966, Mayr 1969, Sneath i Sokal 1973, Wiley 1981, Matile i inni 1993, Kitching i inni 1998), musimy uznać, że stosowanie technik fenetycznych bywa koniecznością. Co więcej, metody te doskonale nadają się do wstępnego rozpoznania struktury danych. Wynikiem większości metod fenetycznych nie są drzewa. Wszystkie należą do **statystycznej analizy wielowymiarowej** (SAW; *multivariate statistical analysis* – MSA). Choć podstawy teoretyczne wielu technik MSA opracowano jeszcze na przełomie XIX i XX stulecia, metody te są na tyle intensywne obliczeniowo, że stosować je zaczęto dopiero wraz z rozpowszechnieniem komputerów. Łatwość ich użycia – jako że wchodzi w skład większości pakietów programów statystycznych – zaowocowała częstym, niejednokrotnie wręcz rutynowym stosowaniem, nie zawsze baczącym na spełnianie warunków zakładanych przez daną technikę bądź właściwą interpretację wyników. Metody MSA są skomplikowane i często trudne do zrozumienia dla przeciętnego biologa, nieznającego choćby rachunku macierzowego, algebry liniowej czy analitycznej geometrii wielowymiarowej. Tutaj, z konieczności, przedstawimy jedynie techniki najużyteczniejsze w taksonomii, traktując je w sposób skrótowy, uproszczony i intuicyjny. Zainteresowanych matematycznymi podstawami czy choćby bardziej formalnym wykładem, odsyłamy do licznych opracowań MSA (Hand 1981, Dunn i Everitt 1982, Bookstein i inni 1985, Krzanowski 1988, Pociecha i inni 1988, Morrison 1990, Jackson 1991, Reyment 1991, Everitt i Dunn 1992, Jajuga 1993, Rohlf 1994, Johnson i Wichern 1998), a szerszym, koncepcyjnym wprowadzeniem – do Kachigana (1991).

Gromadzone obserwacje dotyczą **obiektów**, którymi mogą być osobniki, populacje czy np. gatunki, a wychodząc poza taksonomię biologiczną – choćby akcje firm, pań-

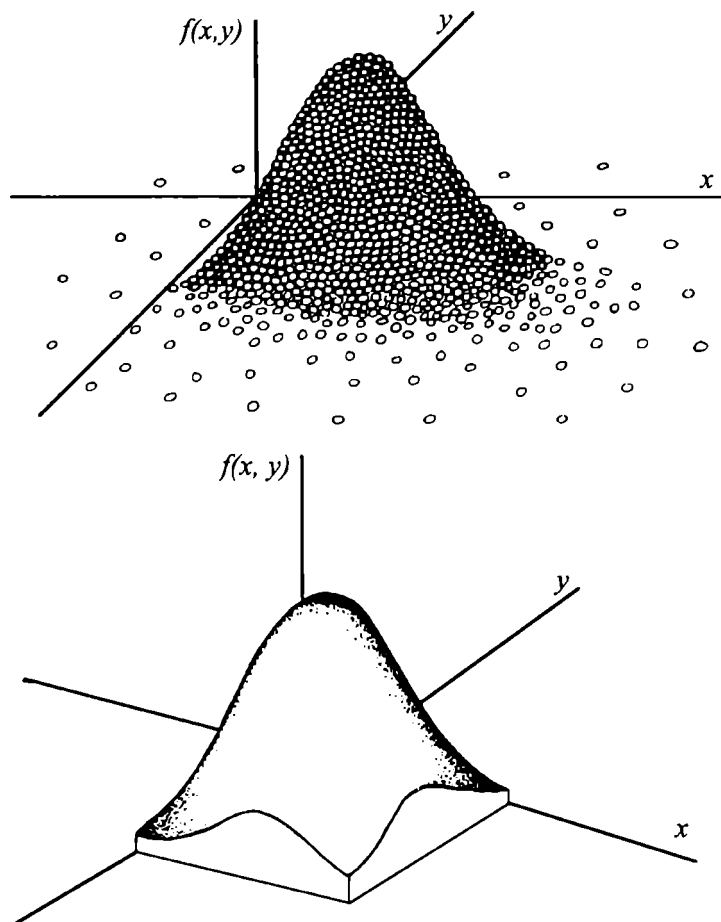
stwa świata, produkty fabryki, procesy technologiczne bądź fizjologiczne, ekosystemy. Wygodne jest stosowanie dla obiektu określenia **operacyjna jednostka taksonomiczna** (*operational taxonomic unit* – OTU) – jak wiemy, może nim być zarówno osobnik, jak i wyższa jednostka, przy czym nie musimy ani znać jej rangi, ani też wiedzieć, czy danej OTU odpowiada rzeczywiście jakaś odrębna grupa. Analizę prowadzimy dla n obiektów. Każdy z tych obiektów opisuje m **zmiennych**, czyli cech: X_1, X_2, \dots, X_m . Obiektem analizy MSA jest więc **wielowymiarowa obserwacja** dana **wektorem zmiennych**:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_m \end{bmatrix}, \text{ albo macierzą obserwacji: } \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}.$$

Obiekt można więc traktować jako punkt w m -wymiarowej **hiperprzestrzeni**. Macierz obserwacji to bezpośredni wynik ustalenia wartości m zmiennych (cech) dla n obiektów, zwykle n OTU. Niekiedy wartości jakichś zmiennych dla pewnych obiektów nie znamy – dane są niekompletne. Metody MSA nie traktują brakujących danych w sposób sensowny, mamy więc do wyboru dwie możliwości. Jeżeli dla jakiegoś obiektu brak wielu danych, to obiekt ten usuwamy z macierzy, podobnie czynimy ze zmienną, której wartości są nieznanne dla wielu obiektów. Natomiast gdy brakujących danych jest niewiele, to uzupełniamy je, podstawiając w ich miejsce wartości oszacowane, najczęściej korzystając z regresji liniowej. Teoretycznie macierz danych analizować można na dwa sposoby: badając związki pomiędzy zmiennymi (cechami) w przestrzeni obiektów (OTU's), co znane jest jako **technika R** (*R technique*), albo śledząc zależności pomiędzy obiektami (OTU's) w przestrzeni zmiennych (cech), co określane bywa jako **technika Q** (*Q technique*). Jakkolwiek oba podejścia bywają przydatne, w taksonomii zdecydowanie częściej stosuje się technikę *Q*; rozróżnienie pomiędzy tymi podejściami nie zawsze jest jednoznaczne (Sneath i Sokal 1973).

Choć zasadniczo techniki wielowymiarowej analizy są uogólnieniami odpowiednich technik jednowymiarowych, to ich zrozumienie i prawidłowe wykorzystanie jest znacznie trudniejsze. Znajomość podstawowych technik statystyki jednowymiarowej jest konieczna, a zarazem pomocna. Jeżeli jednak techniki jednowymiarowe łatwo ilustrować graficznie, danych jest na tyle niewiele, że już „gołym okiem” daje się dostrzec jakieś prawidłowości, a interpretacja biologiczna zarówno wyjściowych danych, jak i wyników analizy na ogół jest względnie prosta, to w przypadku MSA jest inaczej i bardzo łatwo niebezpiecznie zbliżyć się do modelu **GIGO** (*garbage in, garbage out*), dosłownie: „odpady wprowadzamy jako dane i odpady uzyskujemy jako wynik”. Model GIGO, szczególnie prawdopodobny w przypadku **analizy czynnikowej** (*factor analysis*) – nieprzydatnej w taksonomii numerycznej, więc pominiętej w tej książce – daje się uogólnić na praktycznie wszystkie techniki MSA, a niestety również na szereg publikacji wykorzystujących MSA w badaniach przyrodniczych. Aby się doń nie zbliżyć, należy starannie zbierać dane, dobierając takie, które mogą mieć znaczenie biologiczne, następnie starannie sprawdzić zbiór obserwacji, zwłaszcza pod kątem **jednorodności** i **eliptyczności rozkładu**. Należy też rozstrzygnąć, czy zastosujemy podej-

ście stochastyczne, czy jedynie opisowe (to drugie, jak pamiętamy, bezpieczniejsze). Dopiero następnym krokiem jest dobór metody, oczywiście uwzględniając też pytanie, jakiemu celowi analiza ma służyć. Ostatni krok to staranna analiza wyników, zwłaszcza pod kątem prawdopodobnych błędów, a następnie interpretacja wyników, zgodna zarówno z wiedzą statystyczną, jak i biologiczną oraz zdrowym rozsądkiem.



Ryc. 3.1. Funkcja gęstości prawdopodobieństwa $f(x,y)$ dwuwymiarowego (zmiennie x i y) rozkładu normalnego. U góry widoczne poszczególne obserwacje, u dołu powierzchnia jednakowej gęstości

MSA może być stosowana dla zmiennych dyskretnych, lecz aparat matematyczny opracowany jest pełniej dla cech ciągłych, częściej zresztą spotykanych, i do nich ograniczymy się tutaj. Najpełniejszą charakterystyką wektora zmiennych \mathbf{X} w populacji jest rozkład wielowymiarowy, będący uogólnieniem znanego ze statystyki jednowymiarowej pojęcia rozkładu zmiennej losowej. Wielowymiarowy rozkład dla zmiennych ciągłych określony jest funkcją gęstości, czyli nieskończenie wielu par postaci: $(x, f(x))$, gdzie: x – m -wymiarowy wektor, będący wartością wektora zmiennych \mathbf{X} , $f(x)$ – wartość funkcji gęstości dla wektora x , przy tym: $f(x) \geq 0$ i $\int [f(x)dx] = 1$ (gdzie

całka obliczona jest dla zbioru wszystkich możliwych wartości wektora \mathbf{X} : Jajuga 1993). Najpowszechniej znanym wykresem funkcji gęstości dla rozkładu jednowymiarowego jest charakterystyczna, dzwonowata krzywa Gaussa, charakteryzująca rozkład normalny. Dla $m = 2$, czyli rozkładu dwuwymiarowego, funkcja gęstości rozkładu normalnego ma podobną postać (Ryc. 3.1). Wartość maksymalną $f(x)$ nazywamy modalną rozkładu wielowymiarowego. Zarazem – podobnie jak dla rozkładów jednowymiarowych – im wyższa wartość funkcji gęstości, tym wyższe prawdopodobieństwo wystąpienia wartości z niewielkiego otoczenia punktu odpowiadającego tej wartości. Najważniejszymi charakterystykami wielowymiarowego rozkładu są: **wektor średnich** rozkładu, **macierz kowariancji** rozkładu, **macierz korelacji** rozkładu i **powierzchnie jednakowej gęstości** rozkładu. Wektor średnich rozkładu wektora \mathbf{X} to:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_m \end{bmatrix}, \text{ gdzie } \mu_j = E(X_j), \text{ a macierz kowariancji: } \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2m} \\ \dots & \dots & \dots & \dots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_{mm} \end{bmatrix},$$

gdzie: $\sigma_{jk} = E\{[X_j - E(X_j)][X_k - E(X_k)]\}$. Jak widać, elementy wektora średnich rozkładu to **wartości oczekiwane** jednowymiarowych rozkładów poszczególnych zmiennych, wchodzących w skład wektora \mathbf{X} . Elementy głównej przekątnej macierzy kowariancji to **wariancje** tych rozkładów jednowymiarowych, czyli kwadraty odchyleń standardowych kolejnych zmiennych, zaś pozostałe elementy macierzy to **kowariancje** par zmiennych, tworzących wektor \mathbf{X} . Macierz korelacji rozkładu wektora \mathbf{X} jest określona jako:

$$\mathbf{P} = \begin{bmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1m} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2m} \\ \dots & \dots & \dots & \dots \\ \rho_{m1} & \rho_{m2} & \dots & \rho_{mm} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1m} \\ \rho_{21} & 1 & \dots & \rho_{2m} \\ \dots & \dots & 1 & \dots \\ \rho_{m1} & \rho_{m2} & \dots & 1 \end{bmatrix}, \text{ gdzie: } \rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}.$$

Powyższe formuły dotyczą oczywiście ujęcia parametrycznego, czyli mogą być stosowane w modelu stochastycznym. W ujęciu opisowym będziemy mówić odpowiednio o **średniej zbioru** m (albo \bar{x}), **kowariancji zbioru** s_{jk} i **korelacji zbioru** r_{jk} . Powierzchnie jednakowej gęstości rozkładu wektora \mathbf{X} dane są równaniem: $f(x) = c$, gdzie c to stała dodatnia, określająca wartość funkcji gęstości. Jeżeli powierzchnię gęstości rozkładu dwuwymiarowego przetniemy płaszczyzną prostopadłą do osi gęstości, na poziomie odpowiadającym wartości stałej c , to otrzymamy **krzywą jednakowej gęstości** c .

Uogólnieniem powszechnie znanego jednowymiarowego rozkładu normalnego, o charakterystycznej, dzwonowatej krzywej gęstości, jest **wielowymiarowy rozkład normalny**. Odsyłając zainteresowanych matematycznym zdefiniowaniem tego rozkładu i jego właściwości do wymienionych wyżej podręczników statystyki, poprzestaniemy na sformułowaniu kilku ważnych uwag. Wielowymiarowy rozkład normalny jest naj-

lepiej opracowanym rozkładem wielowymiarowym i ma szereg dogodnych właściwości. Krzywe jednakowej gęstości m -wymiarowego rozkładu normalnego są **hiperelipsoidami**, których **centroidem** jest μ (spodziewana wartość średnia wektora X), a określone wartości gęstości c opisują części hiperprzestrzeni, w których zawierają się określone części badanej populacji (np. 50 czy 90%), czyli odpowiadają określonym prawdopodobieństwom znalezienia się obiektu wewnątrz danego obszaru hiperprzestrzeni. Jest to analogiczne do własności rozkładu normalnego jednowymiarowego, gdzie wielkość odchylenia standardowego określa nam, w jakiej odległości od wartości średniej znajdzie się dana część populacji. Ponadto transformacja liniowa wektora zmiennych o rozkładzie normalnym wielowymiarowym będzie też mieć rozkład normalny wielowymiarowy, każdy zestaw wektorów jednowymiarowych wchodzących w skład wektora wielowymiarowego o rozkładzie normalnym będzie mieć rozkład normalny, rozkłady brzegowe i warunkowe również będą wielowymiarowo normalne. Wreszcie zerowa kowariancja oznacza, że odpowiednie zmienne mają niezależne rozkłady.

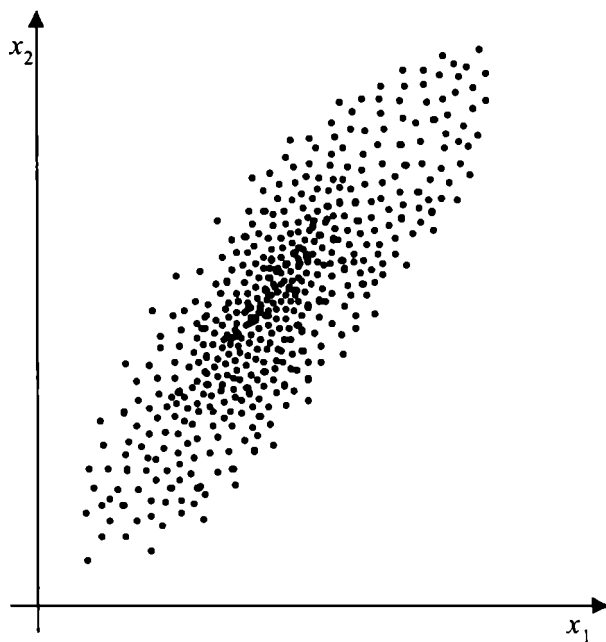
Wielowymiarowy rozkład normalny uważa się za dobry model wielu populacji badanych statystycznie (choć tu warto wyraźnie rozróżnić między populacją w rozumieniu statystyki a populacją biologiczną – dla tej ostatniej rozkład normalny najczęściej będzie zupełnie nieodpowiedni), a poza tym próby o dużej liczebności (w praktyce wystarcza zwykle $n > 30$) i pobrane losowo mają zwykle rozkład zbliżony do normalnego, bez względu na to, jaki rozkład miała populacja, z której je pobrano. W praktyce dane nigdy nie mają dokładnie rozkładu normalnego, co najwyżej zbliżają się doń. Warto przypomnieć wcześniejsze rozważania o rozkładach i obserwacjach nietypowych, niemal zawsze konieczna jest transformacja, odpowiednia do rodzaju zmiennej, aby rozkłady jednowymiarowe zbliżyły się do normalności. Często to jednak zawodzi. Rozkład w rzeczywistych populacjach niejednokrotnie odbiega od normalnego, mając „grube ogony” z dala od wartości średniej. Dla takich populacji stworzono szersze pojęcie **rozkładu eliptycznego**, którego szczególnym przypadkiem jest rozkład normalny. Jest jednak wiele rozkładów niebędących rozkładami eliptycznymi (choćby „łamanego pręta” – *broken stick* – patrz dalej), nierzadkich w przyrodzie. Nieraz też będziemy mieć do czynienia z mieszanką rozkładów: np. dwiema grupami, o różnych parametrach – a niekiedy i typie – rozkładu.

Warto podkreślić, że większość technik MSA opiera się na mocnym założeniu wielowymiarowej normalności bądź co najmniej eliptyczności badanej populacji. Założenie to jest szczególnie ważne, gdy stosujemy ujęcie stochastyczne. W MSA w ujęciu stochastycznym będziemy mówić o „rozkładzie wielowymiarowym wektora zmiennych”, podczas gdy w ujęciu opisowym o „rozkładzie wielowymiarowym zbioru obserwacji”. Zastosowanie wielu technik MSA dla danych, które nie spełniają warunku wielowymiarowej eliptyczności, jest nieuprawnione: w miarę rosnących odchyień od normalności czułość metod będzie malała, maleć więc będzie wiarygodność wyników i dość szybko osiągniemy stan, w którym opisywać będziemy nieistniejące byty oraz wykazywać niezachodzące zależności. Zanim więc zaczniemy analizy MSA, należy zbadać rozkład danych. Istnieją testy na wielowymiarową normalność (Jajuga 1993), jednak ich czułość zależy od wielkości próby, przy czym n powinno być większe od $10m$ – warunek zwykle nierealny, bowiem dla analizy prowadzonej na, powiedzmy, $m = 20$ zmiennych – wcale nie tak wielu w praktycznych zastosowaniach MSA – próba musiałaby liczyć $n = 10 \cdot 20 = 200$, a zmierzenie 20 wielkości u 200 okazów to już duży wysiłek; w dodatku w taksonomii zwykle jest więcej niż jedna próba – coś

z czymś się porównuje – więc należałoby te 4000 pomiarów wykonać co najmniej dwukrotnie. Pozostają więc inne techniki. Są to jednak – co częste w statystyce – testy jednostronne: możemy za ich pomocą wykluczyć normalność (eliptyczność) wielowymiarowego rozkładu, natomiast niewykluczenie nie gwarantuje normalności (eliptyczności), choć oczywiście zwiększa ufność dla przeprowadzanych analiz.

Wiadomo, że jeżeli choć jeden z m jednowymiarowych rozkładów nie będzie rozkładem normalnym, to wielowymiarowy rozkład, w którego skład wchodzi ten jednowymiarowy, również normalny nie będzie. Przeprowadzenie więc testów na normalność wszystkich rozkładów jednowymiarowych pozwoli nam na wykluczenie normalności rozkładu wielowymiarowego bądź na odrzucenie tych zmiennych, których rozkłady wykluczają normalność wielowymiarowego rozkładu. Testami tymi sprawdzimy też efektywność wcześniejszych transformacji danych, a także jednowymiarową jednorodność zbioru obserwacji. Dla zbioru niejednorodnego, wyraźnie podzielonego na, powiedzmy, dwie grupy, w przypadku wielu technik najlepiej prowadzić analizy dla każdego ze zbiorów osobno. Obserwacje nietypowe należy starannie rozważyć, jak już była o tym mowa w Części 2; w każdym przypadku warto przynajmniej sprawdzić, czy nietypowość obserwacji nie jest wynikiem błędu. Oczywiście dwa rozkłady nawet idealnie normalne dać mogą rozkład dwuwymiarowy niebędący rozkładem normalnym, podobnie rozkład dwuwymiarowy może mieć obserwacje nietypowe, które nie są nietypowymi w którymkolwiek z tworzących go dwóch rozkładów jednowymiarowych. Zarazem rozkład dwuwymiarowy łatwo przedstawić graficznie, w Kartezjańskim układzie współrzędnych (x_1, x_2) .

Rozrzut punktów na płaszczyźnie, gdy mamy co najmniej 20–30 obserwacji, dostarcza szeregu informacji (Ryc. 3.2). Krzywa, którą można otoczyć zbiór punktów, jest przybliżeniem krzywej jednakowej gęstości zawierającej „prawie” cały rozkład, a gdy ma kształt eliptyczny, to sugeruje, że jest to rozkład eliptyczny, być może nawet normalny. Wówczas proporcja osi wielkiej i małej elipsy wskazuje na wartość korelacji obu zmiennych, tym wyższą, im proporcjonalnie mniejsza jest oś mała. Zagęszczenie punktów w jakimś rejonie wskazuje, że tam mniej więcej leży modalna funkcji gęstości, natomiast obecność wyraźnie oddzielonych od siebie podzbiorów punktów wskazuje, że rozkład jest mieszkanką tylu rozkładów składowych, ile jest tych podzbiorów. Choć nie jest to gwarancją dwuwymiarowej normalności, to jednak pozbawiony obserwacji nietypowych i eliptyczny w zarysie zbiór punktów, zagęszczający się wzdłuż środkowych części obu osi elipsy, w dodatku uzyskany z dwóch rozkładów normalnych, zwykle będzie dwuwymiarowym rozkładem normalnym: przypadki „patologicznych” rozkładów dwuwymiarowych, które, pomimo spełniania powyższych warunków, dwuwymiarowo normalne nie są, występują rzadko (Johnson i Wichern 1998). Musimy jednak sprawdzić każdą kombinację dwuwymiarową zmiennych, a takich kombinacji jest $m(m-1)/2$, czyli 45 dla $m = 10$, 190 dla $m = 20$, a 1225 dla $m = 50$. Przy większej liczbie zmiennych jest to więc uciążliwe, choć przy wykorzystaniu modułów graficznych wielu z pakietów statystycznych nie wykracza poza możliwości. Zwykle zresztą uwzględnianych cech będzie mniej niż w wyjściowych danych, właśnie z uwagi na niemożność uzyskania normalności rozkładów wielu zmiennych.



Ryc. 3.2. Rozrzut punktów przedstawiający dwuwymiarowe obserwacje. Eliptyczny kształt chmury punktów wskazuje, że rozkład jest eliptyczny, być może nawet normalny, a proporcje osi wielkiej i małej elipsy wskazują na dość wysoką wartość korelacji między x_1 i x_2 . Zagęszczenie punktów wskazuje na rejon modalnej funkcji gęstości, rozkład wygląda na jednorodny. Choć taki obraz nie gwarantuje dwuwymiarowej normalności, to jednak można się jej zwykle spodziewać

Często zdarzy się, że pomimo transformacji zbiór danych pozostanie niejednorodny. Wiele technik, choćby te oparte na metodzie najmniejszych kwadratów (np. regresja czy średnia), jest bardzo wrażliwych na niejednorodność zbioru i nawet pojedyncza obserwacja nietypowa może powodować zupełne załamanie techniki, czyli uzyskanie wyników całkowicie nieodzwierciedlających badanej rzeczywistości. W MSA brak rutynowych metod postępowania, gdy zbiór obserwacji nie jest jednorodny (w dodatku większość technik wymaga jednorodności eliptycznej). W ujęciu stochastycznym za jednorodny uważa się zbiór losowo pobrany z populacji, charakteryzującej się rozkładem eliptycznym. W ujęciu opisowym jednorodny jest zbiór, którego wszystkie elementy zawierają się wewnątrz hiperelipsoidy. Gdy zbiór nie jest jednorodny, możemy mieć do czynienia z mieszkanką dwóch lub więcej jednorodnych zbiorów albo z jednym zbiorem, którego niejednorodność jest następstwem obecności obserwacji nietypowych. Niekiedy wyraźnie widać, że mamy do czynienia z, powiedzmy, dwoma zbiorami i analizy prowadzić wówczas należy oddzielnie dla każdego z nich. W innych przypadkach możemy użyć funkcji klasyfikacyjnych, grupujących obiekty w, kolejno, 2, 3, ..., j zbiorach, aż $j + 1$ da jeden zbiór pusty. Funkcje klasyfikacyjne, do których wrócimy w dalszych rozdziałach, iteracyjnie minimalizują odległości pomiędzy

obiektami należącymi do tej samej grupy, a maksymalizują pomiędzy obiektami należącymi do różnych grup. Po sklasyfikowaniu obiektów w j grup (a w ujęciu stochastycznym zdefiniowaniu j rozkładów) przeprowadzimy oczywiście analizy dla każdej z nich oddzielnie.

Inna strategia to zastosowanie zbiorów rozmytych (Zadeh 1975, Jajuga 1993) – upraszczając, im bardziej nietypowa obserwacja, tym mniej należy do zbioru, a więc mniej waży; obserwacja może też należeć w, powiedzmy, 60% do jednego zbioru, a w 40% do innego. Często zbiór obserwacji podzielić się da na zwarty „rdzeń”, zawierający zdecydowaną większość obserwacji, i pewną liczbę obserwacji poza tym zbiorem. Metody odporne zakładają, że właściwości badanego zbioru (populacji) dostatecznie dobrze odzwierciedlają elementy tego „rdzenia”, toteż pominięcie bądź też zdecydowanie niższe ważenie obiektów spoza „rdzenia” znacznie mniej obniży wartość wyników niż przekłamanie, będące następstwem uwzględnienia obserwacji nietypowych. W praktyce warto więc zastanowić się, czy pewnych obserwacji zwyczajnie nie pominąć.

3.2. Analiza głównych składowych

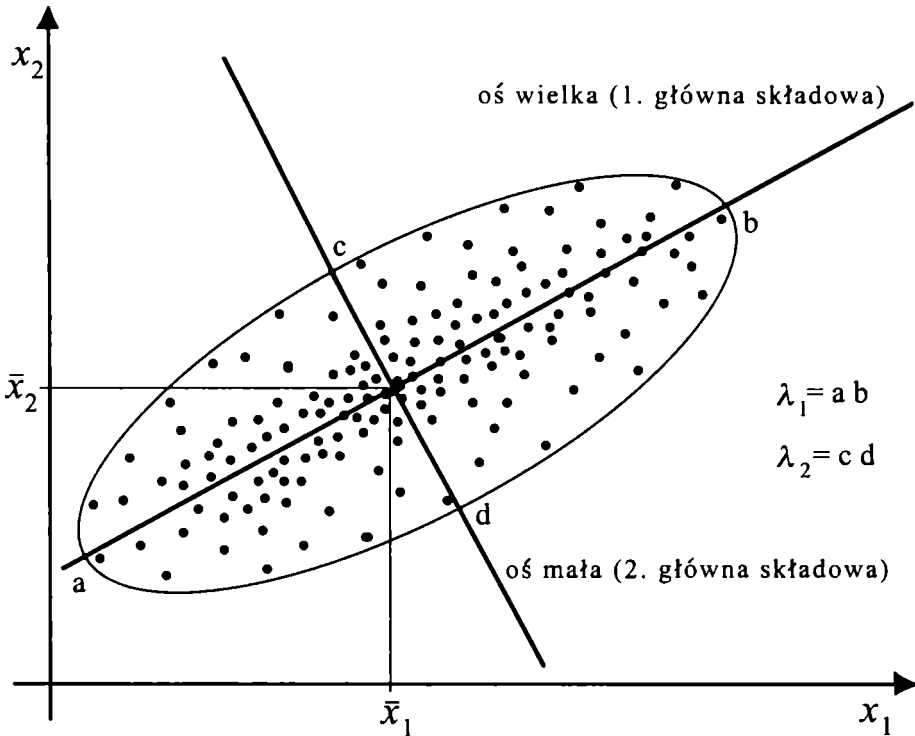
Analiza głównych składowych (*principal component analysis* – PCA) to bodaj czy nie najużyteczniejsza dla taksonomii technika MSA. Wprowadzona przez Hotellinga (1933), pozwala na redukcję danych i przejrzystą interpretację zarówno zależności pomiędzy zmiennymi, jak i struktury zbioru obserwacji. Umożliwia transformację obserwacji m -wymiarowych w dwuwymiarową przestrzeń, a więc geometryczną interpretację struktury zbioru obiektów, wskazując istnienie ewentualnych skupisk – potencjalnych taksonów – bez jakichkolwiek założeń *a priori*. PCA, jako technika graficznej prezentacji zbioru obserwacji wielowymiarowych, powinna być jednym z pierwszych etapów wielowymiarowej analizy danych. Technika ta pozwala na dostrzeżenie zależności, które inaczej pozostałyby nieuchwyte.

Zasadę PCA najłatwiej zrozumieć, rozważając geometryczne własności zbioru dwuwymiarowych, skorelowanych obserwacji (Ryc. 3.3). Jak pamiętamy, związek o charakterze korelacji zakłada niezależność obu zmiennych, nieuprawnione byłoby tu więc stosowanie regresji dla scharakteryzowania tendencji zaznaczających się w zbiorze. Dwuwymiarowość obserwacji oznacza też, że zamiast przedziałów ufności powinniśmy operować powierzchnią opisaną elipsoidalną krzywą jednakowej gęstości o zadanej wartości c . Dla dwuwymiarowego rozkładu normalnego musimy znać **średnie** z próby (\bar{x}_1, \bar{x}_2) , **odchylenia standardowe** (s_1, s_2) i **kowariancję** $(s_{12} = \sum x_1 x_2 / (n - 1))$. Z geometrii pamiętamy, że elipsę opisują dwie prostopadłe do siebie **główne osie** (*principal axes*): wielka i mała, przecinające się w środku elipsy. Środkiem elipsy będzie tu więc wartość (\bar{x}_1, \bar{x}_2) , a oś wielka przebiegnie tak, aby najlepiej zobrazować rozkład punktów. Osiągnięte to zostanie, jak w wielu technikach statystycznych, metodą najmniejszych kwadratów. Jak wiadomo, rzeczywiste wartości badanych zmiennych będą raz większe, innym razem mniejsze od wartości średnich – z definicji średniej arytmetycznej wynika, że suma ich odchyłeń równa będzie 0, niezależnie od tego, czy odchylenia były małe, czy duże. Jeżeli jednak odchylenia od średniej podniesiemy do kwadratu, to wartości ujemne znikną i suma kwadratów odchyłeń, odpowiednio większa

lub mniejsza, będzie miarą dobroci odwzorowania. Dobieramy więc przebieg wielkiej osi elipsy tak, aby suma kwadratów odchyłeń od osi głównej była najmniejsza. Równanie wielkiej osi będzie miało postać: $x_1 = \bar{x}_1 + b_1(x_2 - \bar{x}_2)$, gdzie b_1 to współczynnik nachylenia wielkiej osi, analogicznie jak w równaniu regresji. Obliczamy go z równania:

$$b_1 = s_{12} / (\lambda_1 - s_1^2), \text{ gdzie: } \lambda_1 = \frac{1}{2} \left[s_1^2 + s_2^2 + \sqrt{(s_1^2 + s_2^2)^2 - 4(s_1^2 s_2^2 - s_{12}^2)} \right] \text{ (Sokal i Rohlf 1995).}$$

Współczynnik nachylenia osi małej, prostopadłej do wielkiej, to: $b_2 = -1/b_1$. Aby opisać kształt elipsy, konieczna jest jeszcze znajomość proporcji długości wielkiej i małej osi, dana wielkością: $\sqrt{\lambda_1 / \lambda_2}$, gdzie: $\lambda_2 = s_1^2 + s_2^2 - \lambda_1$ (Sokal i Rohlf 1995).



Ryc. 3.3. Zbiór dwuwymiarowych obserwacji (x_1 i x_2) przedstawiony na płaszczyźnie otoczyć można elipsą. Oś wielka i mała tej elipsy krzyżują się w punkcie o współrzędnych \bar{x}_1, \bar{x}_2 . Geometrycznie osie te reprezentują odpowiednio pierwszą i drugą główną składową

Powyżej zdefiniowaliśmy w formalny sposób wartości λ_1 i λ_2 , ważne w MSA. Są to **wartości własne** (*eigenvalues*, zwane też *latent roots* lub *characteristic roots*) macierzy wariancji/kowariancji $\mathbf{X}_1\mathbf{X}_2$. Jak widać, są one miarą zmienności wyrażającej się wzdłuż tych osi, czyli częścią całkowitej wariancji, wyjaśnianej przez daną oś. Same zaś osie – wielka i mała – to **wektory własne** (*eigenvectors*, ale i *latent vectors*, *characteristic vectors*) też macierzy. Im wyższa wartość korelacji, tym oś wielka jest większa, proporcjonalnie do małej. W rozpatrywanym przypadku w miejsce dwóch wyjściowych zmiennych x_1, x_2 otrzymaliśmy dwie **główne składowe** (*principal components*), a więc o redukcji danych mówić tu trudno. Z drugiej strony, jak widać, pierwsza główna składowa, czyli oś wielka elipsy, wyjaśnia zdecydowaną większość zmienności i – być może – opis zmienności byłby dostateczny, gdyby drugą składową pominąć. Osie są z definicji **prostopadłe** (*ortogonalne*) do siebie, a więc formalnie niezależne (kąąt prosty odpowiada korelacji $r = 0$). Warto jednak już tutaj przypomnieć, że jest to niezależność formalna, w dodatku opisująca jedynie brak zależności liniowej – w rzeczywistości uzyskane w ten sposób zmienne mogą nie być biologicznie niezależne. Uogólniając rozważania na m -wymiarową chmurę punktów, zawartą w obrębie hiperelipsoidalnego fragmentu hiperprzestrzeni, transformacja koordynatów tych punktów w dwuwymiarową przestrzeń dwóch spośród głównych składowych pozwala na wielką redukcję danych, które w dodatku stają się interpretowalne: w miejsce n obiektów opisanych przez m zmiennych uzyskujemy n obiektów charakteryzowanych przez dwie zmienne, będące głównymi składowymi. Główne składowe są m -wymiarowymi hiperpłaszczyznami, przechodzącymi przez punkt będący wektorem średnich zbioru obserwacji i dobranymi tak, aby suma kwadratów odchyłeń punktów, będących obserwacjami, była najmniejsza.

PCA przeprowadza się na macierzy wariancji/kowariancji lub korelacji – którą z nich wybrać, rozważymy dalej. Jednorodność zbioru obserwacji umożliwi zastosowanie PCA do badania wewnętrznych zależności pomiędzy zmiennymi, a wielowymiarowo normalny rozkład pozwoli na interpretację wyników w kontekście elipsoid stałych gęstości i na wnioskowanie o parametrach rozkładu. Sama PCA natomiast nie wymaga założenia jednorodności zbioru czy wielowymiarowej normalności rozkładu, aby prawidłowo znaleźć główne składowe i wiernie rzutować obiekty w przestrzeń głównych składowych. Zresztą gdyby tak nie było, to technika ta nie mogłaby być wykorzystywana do wskazywania zróżnicowania badanej populacji na grupy wcześniej niezdefiniowane.

Główne składowe są liniowymi kombinacjami współrzędnych wektora \mathbf{X} , są unormowane (suma kwadratów współczynników kombinacji jest równa 1) i nieskorelowane z sobą. Pierwszą główną składową (PC1) wyznacza się w ten sposób, że tworzy się „nową” zmienną, która jest kombinacją liniową zmiennych oryginalnych i ma wariancję największą spośród wszystkich możliwych unormowanych kombinacji liniowych tych zmiennych – analogicznie jak w omówionym przykładzie dwuwymiarowym znajdujemy oś wielką m -wymiarowej hiperelipsy, odzwierciedlającą („wyjaśniającą”) największą możliwą część ogólnej wariancji, czyli całkowitej zmienności we wszystkich m wymiarach. Będzie to największy wektor własny macierzy i będzie można mu przypisać określoną wielkość własną, równą wariancji pierwszej składowej, a więc im wyższa będzie ta wartość, tym więcej całkowitej zmienności wyjaśni ta główna składowa. Główną składową interpretować można jako wynik przesunięcia oryginalnego układu współrzędnych, tak aby jego początek znalazł się w położeniu wielowymiaro-

wej średniej, a następnie obracania osi, aż jedna z nich bieć będzie w kierunku największej wariancji. Drugą główną składową (PC2) wyznacza się, tworząc kolejną „nową” zmienną, o największej wariancji spośród wszystkich kombinacji liniowych nieskorelowanych z PC1 (czyli ortogonalnych do PC1). Trzecia główna składowa jest kombinacją liniową tłumaczącą największą część zmienności spośród wszystkich ortogonalnych do pierwszej i drugiej, itd. Formalnie dla macierzy o m zmiennych wyznaczyć można m głównych składowych. Oczywiście wartości własne kolejnych składowych będą coraz niższe, czyli tłumaczyć będą coraz mniej ogólnej zmienności.

Normalizacja oznacza, że suma wartości własnych wszystkich m głównych składowych, równa sumie wariancji wszystkich zmiennych wchodzących w skład wektora X , równa jest też m . Jest to wygodne, bowiem gdy np. wartość własna dla, powiedzmy, drugiej głównej składowej wynosi 3,52, to intuicyjnie możemy to rozumieć jako tłumaczenie przez tę składową tyle samo zmienności, co 3,52 wyjściowej zmiennej (choć oczywiście udział tych wyjściowych zmiennych w ogólnej zmienności będzie różny). Zarazem wielkość wartości własnej, podzielona przez m , wskaże nam część wariancji (czyli całkowitej zmienności) tłumaczoną przez daną główną składową (co najczęściej podaje się procentowo).

Wreszcie dla każdej składowej można obliczyć, co wiele pakietów statystycznych zapewnia, jakie części zmienności kolejnych zmiennych wyjaśnia ta składowa. Jak wiemy, każdą z głównych składowych traktować można jako **ważoną kombinację** wszystkich zmiennych oryginalnych, i te właśnie **wagi** (*weights*), gdy są wysokie, wskazują, które zmienne oryginalne mają największe udziały w danej składowej. Można też obliczać korelację kolejnych zmiennych z daną składową (*character loadings*). Ta ostatnia miara bywa uważana za gorszą, bowiem odzwierciedla jedynie jednowymiarowy związek danej zmiennej ze składową, nie uwzględniając zupełnie oddziaływania pozostałych zmiennych, gdy waga jest miarą wielowymiarowej zależności (Johnson i Wichern 1998). Wydaje się to słuszne, choć w praktyce zazwyczaj wyższym wagom odpowiadają wyższe wartości korelacji. Ponadto oba te parametry nie wydają się zbyt użyteczne w taksonomii numerycznej, bowiem interpretacja określonej składowej jako odzwierciedlającej właśnie te a nie inne zmienne jest mocno ryzykowna, a w dodatku wymaga jednorodności, a jeszcze lepiej wielowymiarowej eliptyczności rozkładu. W niejednokrotnie rutynowym obliczaniu udziałów poszczególnych zmiennych w kolejnych składowych znajduje zapewne odbicie częste we wcześniejszej literaturze statystycznej mylenie PCA z **analizą czynnikową** (*factor analysis*) – choć tę ostatnią przeprowadza się niekiedy techniką PCA, to jej założenia są zupełnie inne i tam właśnie poszukuje się interpretacji merytorycznej składowych, co jest ryzykowne i czego tutaj czynić nie będziemy, choć niektórzy badacze uznają analizę czynnikową za odpowiednią bądź wręcz najwłaściwszą dla badań morfometrycznych (np. Bookstein i inni 1985).

Jak wspominaliśmy, PCA oblicza się dla macierzy wariancji/kowariancji albo korelacji. Wyniki będą zupełnie różne. Jak wiemy, wariancja to wielkość mianowana, a kowariancja może mieć różne rzędy wielkości, a więc dla jednych par cech być równa, powiedzmy, 1250, gdy dla innej 0,35. Tak więc w PCA prowadzonej na wariancjach/kowariancjach cechy o większej wariancji ważyć będą więcej, toteż wynik zależęć będzie niemal wyłącznie od ich wartości. Nie musi to być wadą, choć najczęściej jest. PCA oparta na wariancji/kowariancji jest odpowiednia jedynie wówczas, gdy mamy merytoryczne przesłanki, aby uważać, że cechy bardziej zmienne mają większe

znaczenie w rzeczywistości odwzorowywanej przez naszą analizę. W przypadku – dla nas najczęstszym – badań morfometrycznych PCA oparta na macierzy wariancji/kowariancji jest odpowiednia wówczas, gdy możemy założyć model wzrostu osobników jako zwiększanie rozmiarów, proporcjonalne do absolutnej wariancji każdej ze zmiennych. Jeżeli natomiast możemy założyć, że wzrost osobników obejmuje powiększanie wszystkich mierzonych zmiennych proporcjonalnie do ich współczynników zmienności, odpowiedniejsze jest użycie macierzy korelacji (Dillon 1984). W praktyce zwykle nie potrafimy określić, który z modeli wzrostu lepiej opisuje nasze dane. Najczęściej więc nie będzie błędem użycie macierzy korelacji. Wyniki PCA będą identyczne dla macierzy korelacji obliczonych dla danych oryginalnych, a w każdym razie nienormalizowanych, i macierzy wariancji/kowariancji wyliczonych dla danych normalizowanych. W praktyce jednak, gdy obliczymy korelacje dla danych niestandardyzowanych, wynikiem będzie tłumaczenie mniej więcej 90% zmienności przez pierwszą składową, co na ogół nie jest pożądane. Polecieć więc warto obliczanie macierzy korelacji na danych wcześniej standaryzowanych, co da mniej zmienności tłumaczonej przez pierwszą składową; taką procedurę zresztą wskazują Sneath i Sokal (1973).

Warto rozważyć pewne skrajne przypadki. Może się zdarzyć, że wszystkie korelacje (lub kowariancje) są równe 0. Wówczas zestaw głównych składowych będzie identyczny z zestawem oryginalnych, nieskorelowanych zmiennych; jakkolwiek analiza zależności między zmiennymi ani też redukcja danych nie będzie możliwa dla takiej m -wymiarowej sferoidalnej chmury punktów. Kolejne składowe będą miały takie same wartości własne. Dlatego też testuje się statystyczną istotność różnic wartości własnych dla kolejnych składowych (Rohlf 1994, Johnson i Wichern 1998). Inny przypadek, z którym spotkać się można niejednokrotnie w biometrii, to takie same (lub niemal takie same) wartości wszystkich korelacji r w macierzy. Dzieje się tak w wyniku zależności wszystkich mierzonych wartości od jednego czynnika – zróżnicowania rozmiarów w następstwie różnego wieku badanych osobników. Wówczas wartość własna pierwszej składowej: $\lambda_1 = 1 + (m - 1)r$, a pozostałych: $\lambda_2 = \lambda_3 = \dots = \lambda_m = 1 - r$. Dla wielowymiarowego rozkładu normalnego hiperelipsoida jest wówczas cygarowata. Do tego modelu zbliżają się często wyniki pomiarów różnych wielkości u kolejnych osobników w populacji. Inny przypadek to sytuacja, gdy ostatnia (lub parę ostatnich) wartość własna równa jest 0 albo niemal równa 0. Choć wyjaśnianie niemal całej zmienności przez co najwyżej parę „pierwszych” składowych jest wygodne, pozwalając na zdecydowaną redukcję danych i ich łatwą interpretację, to jednak zerowa wartość własna ostatniej składowej jest niepokojąca. Oznacza, że w danych występują jakieś niedostrzeżone wcześniej liniowe zależności i co najmniej którąś ze zmiennych należy wyeliminować jako niepotrzebną albo wręcz powodującą błędne wyniki wielu analiz (choć dla samej PCA nie jest to specjalnie obciążające).

Mając już obliczone m głównych składowych, warto obejrzeć projekcje oryginalnych obserwacji w ich przestrzeń: eliptyczność grupy punktów w przestrzeni kilku pierwszych składowych pozwoli na zorientowanie się, czy rozkład jest wielowymiarowo mniej więcej normalny, zaś obraz w przestrzeni paru „ostatnich” składowych wskaże ewentualne podejrzane czy wręcz nietypowe obserwacje. Aby takie wnioskowanie miało sens, zbiór punktów powinien być duży (co najmniej 30 obiektów), ponadto $n \gg m$, czyli obiektów, powinno być znacznie więcej niż zmiennych je opisujących; PCA, w której np. dla 20 obiektów mierzono 40 czy choćby 20 zmiennych, może dać zupełnie błędne wyniki i należy, w miarę możliwości, takiej strategii badań unikać.

Rzutowanie obiektów w przestrzeń głównych składowych może być interesujące, jednak zamiast m zmiennych oryginalnych mamy m głównych składowych, a więc oryginalne, „znaczące” zmienne zastąpiły zmienne abstrakcyjne, pozbawione pewniejszej merytorycznej interpretacji. Pamiętamy jednak, że wartości własne są dla kolejnych głównych składowych coraz mniejsze. Każdy pakiet statystyczny podaje nam wartości własne i procent wyjaśnianej zmienności, dla każdej PC i kumulatywnie ($\lambda_{k_{kumulacji}} = \lambda_1 + \lambda_2 + \dots + \lambda_k$). Przykładowo:

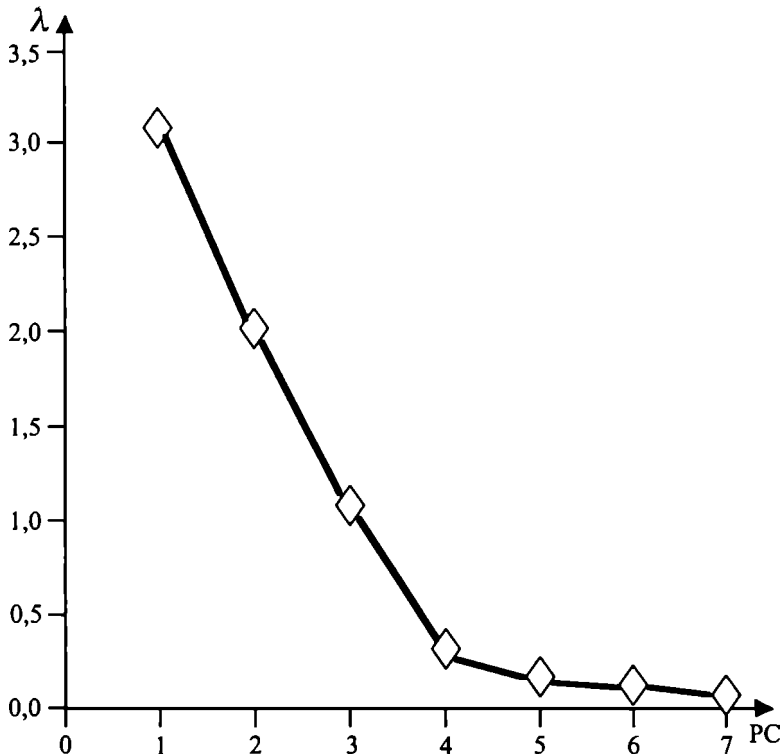
Główne składowe	Wartości własne		% ogólnej wariancji wyjaśnianej	
	składowej	kumulatywnie	składowej	kumulatywnie
PC1	3,16	3,16	45,2	45,2
PC2	2,02	5,18	28,9	74,1
PC3	1,06	6,24	15,1	89,2
PC4	0,33	6,57	4,7	93,9
PC5	0,21	6,78	3,0	96,9
PC6	0,14	6,92	2,0	98,9
PC7	0,08	7,00	1,1	100,0

Jak widać, już trzy główne składowe wyjaśniają niemal 90% zmienności siedmiu wyjściowych zmiennych, niewątpliwie więc dostatecznie dokładny opis zależności osiągnąć można, pomijając część składowych. Pozostaje rozstrzygnąć, ile. Jednoznacznej odpowiedzi nie ma. Kryteria to wielkość wartości własnych oraz procent tłumaczonej zmiennej – oczywiście znaczące dokładnie to samo, jako że to drugie oblicza się z pierwszego. Często zakłada się, że określony procent – np. 75 bywa dostateczny – wyjaśnianej zmienności wystarcza, by inne zmienne pominąć. Wówczas w powyższym przykładzie jedynie PC1 i PC2 wymagałyby uwzględnienia. Albo też zakłada się, że skoro wartość własna jednej zmiennej wyjściowej „odpowiada” jedności, to uwzględnianie składowych o wartości własnej niższej niż 1 nie ma sensu; tak jak wyjaśnianie mniej niż λ_k/m ogólnej wariancji przez k -tą składową (oczywiście tutaj dla $\lambda = 1$ część wyjaśnianej wariancji byłaby równa $1/7 = 14,3\%$). Zgodnie z tym kryterium PC3 formalnie należałoby jeszcze uwzględnić, jednak niekoniecznie. Kryteria powyższe nie mają teoretycznego uzasadnienia, praktycznie są wygodne, lecz nie wolno ich stosować ślepo.

Wartości własne kolejnych składowych nanieść można na wykres (Ryc. 3.4) ilustrujący, jak maleją. Jest to tzw. *scree plot*. *Scree* to piarg czy ospisko – wykres przedstawia stromy klif lub omywany prądem brzeg rzeki, poniżej którego występuje sedymentacja, czyli gromadzenie się rumoszu. Sedymentacja możliwa jest oczywiście jedynie wówczas, gdy stok nie jest zbyt stromy, i o to właśnie tutaj chodzi: wykres wskaże nam, od której głównej składowej wartości własne zaczną spadać powoli: jest to ta składowa, której już nie warto uwzględniać. I to kryterium wskaże na pierwsze 3 składowe, choć często różne kryteria dają odmienne wyniki.

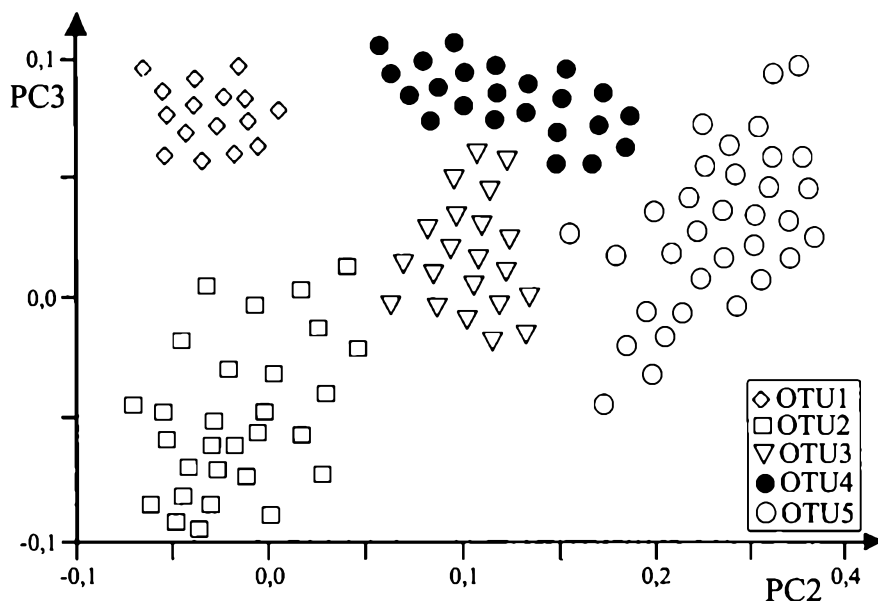
Jak wspominaliśmy, PCA nie zakłada jakiegokolwiek struktury zbioru obserwacji, ani jego jednorodności, ani dających się wyróżnić podzbiorów, i to jest jej ogromną zaletą. Nie znaczy to jednak, aby jakiegokolwiek założenia były obce tej technice. Analiza głównych składowych bada zależności pomiędzy zmiennymi, czy one są, czy też ich nie ma. Stąd wspomniana wcześniej konieczność testowania różnic pomiędzy wartościami własnymi. Technika ta dzieli ogólną wariancję pomiędzy kolejne składowe,

wybierając kolejno malejące części tej zmienności. Byłoby tak również wówczas, gdyby rzeczywistych zależności nie było. Oczywiście całkowity brak zależności to sytuacja niezwykle i raczej czysto teoretyczna – zwłaszcza w złożonym układzie jakiejś zależności zawsze występują, choć może to być dziełem przypadku i nie odzwierciedlać rzeczywistych związków. Gdyby tak było – tzn. wszystkie zależności były dziełem przypadku – podział ogólnej wariancji byłby zgodny z modelem *broken stick*, czyli łamanego pręta (Rohlf 1994). Rozkład taki, bynajmniej nierzadki w przyrodzie, odpowiada modelowi pręta, którym uderzamy raz po raz – wielokrotnie odłamają się krótkie kawałki, rzadko dłuższe. Kształt rozkładu *broken stick* przypomina *scree plot* z Ryc. 3.3. Warto więc porównać procenty tłumaczonej wariancji, obliczone dla kolejnych głównych składowych, z wartościami obliczonymi dla modelu *broken stick*. Oczywiście są to wartości krytyczne, czyli odrzucić powinniśmy te składowe, dla których wartości są niższe.



Ryc. 3.4. Wykres wartości własnych głównych składowych, tzw. *scree plot*. Jak widać, od czwartej głównej składowej wartości własne – a więc udziały tłumaczonej zmienności – spadają powoli, więc trzecia, może czwarta główna składowa są ostatnimi, które warto uwzględnić

W morfometrii wybór składowych ma jeszcze jeden ważny aspekt. Przyjmuje się, że pierwsza główna składowa (PC1) odzwierciedla głównie zróżnicowanie wielkości, a także dymorfizm płciowy czy ewentualny polimorfizm (np. Gould 1977), natomiast następnymi zmiennymi to tzw. zmienne kształtu, czyli odzwierciedlające różnice morfometryczne, inne niż wynikające z różnic ogólnych rozmiarów. Jeżeli więc nie badamy zróżnicowania wielkości, to PC1 powinniśmy odrzucać, koncentrując się na następnych zmiennych. Nie należy jednak, jak to się zdarza w literaturze, uważać pierwszej głównej składowej za odzwierciedlającą wyłącznie zróżnicowanie ogólnych rozmiarów ani tym bardziej pozostałych za wolne od wpływu zróżnicowania ogólnej wielkości („size free”): są one rzeczywiście słabo związane z rozmiarami, ale jednak związane. Zwykle więc będziemy uwzględniać składowe drugą i trzecią – tym istotniejsze jest, by pierwsza składowa nie tłumaczyła, powiedzmy, 90% wariacji, bo wówczas już trzecia – a może nawet druga – będą wyjaśniały mniej niż zgodnie z modelem *broken stick*. Graficznie zilustrować możemy co najwyżej trzy wymiary, choć precyzyjnie jedynie dwa (Ryc. 3.5). Można się więc zastanowić, czy projekcja danych w przestrzeń drugiej i trzeciej składowej wyjaśni nam dostateczną część zmienności, czy będzie dostatecznie wiernym obrazem zależności pomiędzy obiektami. Teoretycznie nie musi tak być, choć najczęściej jest.



Ryc. 3.5. Obiekty należące do pięciu różnych OTU (OTU1–OTU5), rzutowane w przestrzeń drugiej (PC2) i trzeciej (PC3) głównej składowej. Widać zarówno odrębność kolejnych OTU, jak i wzajemne położenie w przestrzeni głównych składowych, wskazujące np., że OTU4 jest bliższy OTU3 niż OTU2

Wspomniana maksymalizacja wyjaśnianej wariacji przez PCA powoduje, że brakujące dane mają niekorzystny wpływ na wynik – każdorazowy brak wartości danej cechy dla jakiegось obiektu zmniejsza wariację tej cechy, a więc i jej udział w two-

rzeniu składowych. Zarazem obiekty, dla których brakuje wartości niektórych cech, grupowane będą bliżej centroidu, niż gdyby wartości te były włączone do analizy. Ponadto PCA zakłada, że projekcja punktów w m -wymiarową przestrzeń wyjaśnia maksymalną część stwierdzonej zmienności, a to powoduje wyższe ważenie większych odległości. Analiza głównych składowych będzie więc notorycznie zaniżała odległości pomiędzy bliskimi sobie obiektami, natomiast odległości większe, zwłaszcza pomiędzy dającymi się wyodrębnić większymi grupami, odtworzy wiernie. Dla ominięcia tego mankamentu, tzn. wskazania lokalnych dystorsji w danych, zwykle rzutuje się na projekcję OTU w przestrzeń PC tzw. najkrótsze drzewo połączeń (*minimum spanning tree* – MST), które omówimy dalej (Rozdział 3.6).

3.3. Analiza głównych współrzędnych

Jak pisaliśmy, jednym z zastosowań analizy głównych składowych jest znajdowanie dwuwymiarowego, na płaszczyźnie, przybliżenia konfiguracji m -wymiarowych obserwacji, uzyskiwane na drodze minimalizacji sumy różnic kwadratów oryginalnych odległości euklidesowych i kwadratów euklidesowych odległości transformowanych, mierzonych na płaszczyźnie. Zadanie to wykonać można również korzystając z **analizy głównych współrzędnych** (*principal coordinate analysis*). Metoda ta, wprowadzona przez Gowera (1966), zaliczana jest do technik **wielowymiarowego skalowania** (*multidimensional scaling* – MDS): inaczej nazywana jest **metrycznym skalowaniem wielowymiarowym** (*metric MDS*). Podobnie jak PCA pozwala na przybliżenie kwadratów odległości m -wymiarowych kwadratami odległości dwuwymiarowych, zarazem odległości nie muszą być Euklidesowe: mogą to być jakiegokolwiek odległości, byle spełniały warunek trójkąta.

Wyjściową macierz odległości d_{ij} , o wymiarach $n \times n$, transformujemy w tzw. macierz podobieństw obserwacji C , również o $n \times n$ elementach określonych jak następuje (Sneath i Sokal 1973, Jajuga 1993): $c_{ii} = 0$, $c_{ij} = -0,5d_{ij}^2$, dla $i \neq j$. Następnie macierz C transformuje się w macierz B , tak samo o wymiarach $n \times n$, o elementach określonych równością:

$$b_{ij} = c_{ij} - c_i - c_j + c, \text{ gdzie: } c_i = \frac{1}{n} \sum_{j=1}^n c_{ij}, \quad c_j = \frac{1}{n} \sum_{i=1}^n c_{ij}, \quad c = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n c_{ij}.$$

Oznacza to, że każda z oryginalnych odległości d_{ij} jest korygowana poprzez odjęcie od niej średniej wszystkich elementów d_{ij} z jej rzędu macierzy, następnie odjęcie średniej wszystkich elementów z jej kolumny, po czym dodanie średniej wszystkich elementów całej macierzy C . Następnie wyznacza się kolejne największe wartości własne macierzy B , oznaczone odpowiednio l_1, l_2, \dots, l_k , odpowiadające im wektory własne u_1 i u_2, \dots, u_k . Ostatni krok to normalizacja wektorów własnych, tak aby suma kwadratów wektora k -tego odpowiadała wartości własnej tego wektora: macierz znormalizowanych wektorów własnych dostarcza koordynatów kolejnych OTU na głównych osiach.

Jak łatwo dostrzec, odwzorowanie m -wymiarowej konfiguracji punktów w mniejszej liczbie wymiarów jedynie wyjątkowo może być idealne. Wystarczy sobie wyobra-

zić sytuację, gdy odległości pomiędzy wierzchołkami trójkąta musimy przedstawić jednowymiarowo, czyli na prostej – jest to przypadek redukcji zaledwie jednego wymiaru, gdy w miejsce niezbędnych dla pełnego przedstawienia zależności dwóch wymiarów musimy użyć jednego. Jest to oczywiście daleko prostsze niż redukcja, powiedzmy, $m = 30$ wymiarów do dwóch, najwyżej trzech, a i tak wierne odwzorowanie trójkąta na prostej nigdy nie będzie możliwe. Rzecz jasna, odwzorowanie będzie raz lepsze, innym razem gorsze, zależnie od danych. Stąd konieczność sformułowania jakiegoś kryterium dobroci odwzorowania. Dla dwuwymiarowego przypadku analizy głównych współrzędnych będzie to współczynnik przybierający wartości w przedziale $\langle 0, 1 \rangle$, tym wyższe, im dokładniejsze przedstawienie konfiguracji OTU na płaszczyźnie (Jajuga 1993):

$$r_0 = \frac{l_1 + l_2}{\sum_{i=1}^{n-1} b_{ij}}.$$

Analiza głównych współrzędnych prowadzona na odległościach obliczonych dla wartości standaryzowanych da identyczne wyniki, jak analiza głównych składowych prowadzona na korelacjach. W przypadku, gdy dla niewielu obiektów mierzono wiele zmiennych, analiza głównych współrzędnych będzie obliczeniowo prostsza, co jednak nie ma większego znaczenia przy możliwościach dzisiejszych komputerów. Ważne natomiast jest wierniejsze niż w przypadku analizy głównych składowych odwzorowanie położenia względem innych tych właśnie obiektów, dla których opisujące je dane są niekompletne (Rohlf 1994). Najważniejsze wydaje się jednak to, że dla analizy głównych współrzędnych wystarcza sama macierz odległości, więc możemy tą techniką badać np. dane z literatury (choćby dla sprawdzenia wyników analizy skupisk), gdy opublikowano jedynie odległości, jak też spotykane niekiedy dane będące wyjściowo odległościami (jak dla reakcji immunologicznych czy hybrydyzacji DNA). W dodatku wykorzystać możemy całą gamę współczynników odległości czy asocjacji, jak długo wynikiem obliczeń nie będą duże ujemne wartości własne macierzy **B**.

3.4. Nieliniowe skalowanie wielowymiarowe

Niejednokrotnie zdarza się, że każda liniowa transformacja w przestrzeń q -wymiarową oryginalnych odległości pomiędzy obiektami przedstawia oryginalną konfigurację obiektów w sposób niezadowolający, zbyt niezgodny z wyjściową. Wówczas lepsze wyniki osiągnąć można – na ogół, choć nie zawsze – za pomocą **nieliniowego skalowania wielowymiarowego** (*nonlinear multidimensional scaling – nonlinear MDS*), wprowadzonego przez Sheparda (1962, 1966, 1980) i Kruskala (1964a, b). Technika ta zakłada, że funkcja transformująca oryginalne n obserwacji (pomiędzy którymi występuje $j = n(n-1)/2$ odległości lub podobieństw) w q -wymiarową przestrzeń nie jest liniowa, a jedynie **monotoniczna** – czyli niezminiająca uporządkowania odległości (od najmniejszej do największej): jeżeli $d_{i_1 k_1} < d_{i_2 k_2} < \dots < d_{i_j k_j}$, to po transformacji:

$d_{i_1 k_1}^{(q)} < d_{i_2 k_2}^{(q)} < \dots < d_{i_j k_j}^{(q)}$. Dlatego można stosować różne odległości lub miary podobieństwa, a nawet subiektywne rangi – wystarczy uporządkować różnice bądź podobieństwa, od najmniejszych po największe – co bywa użyteczne, zwłaszcza gdy brak obiektywnych miar odległości czy podobieństw. Nieliniowe MDS umożliwia znalezienie – gdy mamy zestaw odległości bądź podobieństw pomiędzy n obserwacjami – takiego odwzorowania w możliwie najniższej liczbie wymiarów ($q \leq n - 1$), aby wzajemny układ zależności „niemal odpowiadał” oryginalnemu. Gdy uda się ten układ odwzorować zadowalająco w dwóch bądź trzech wymiarach, to złożone częstokroć zależności można przedstawić w przejrzysty, graficzny sposób. To właśnie sprawia, że nieliniowe MDS jest szeroko stosowane w taksonomii, ale też np. ekologii czy psychologii.

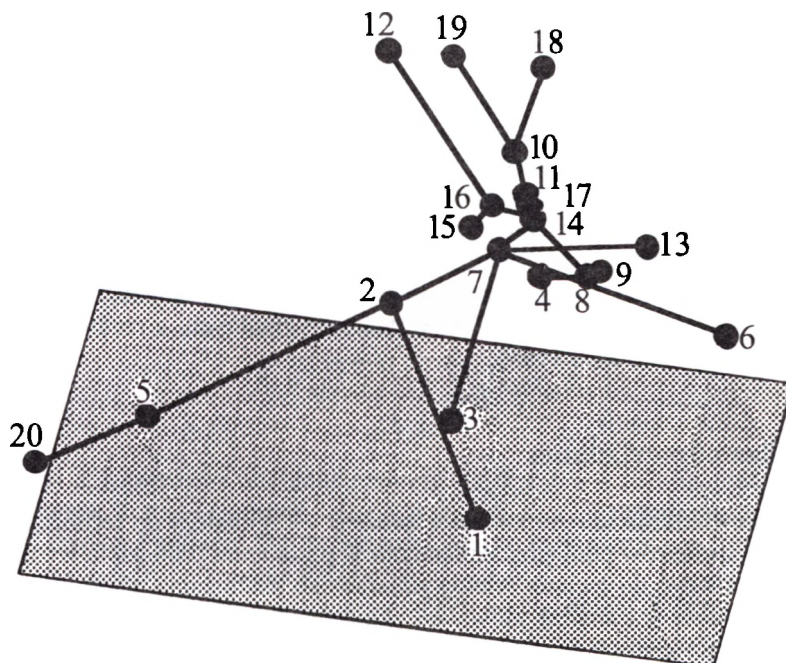
Nieliniowe MDS jest koncepcyjnie proste, należy jednak do technik „obliczeniowo intensywnych”. Oryginalne odległości lub podobieństwa zostają uszeregowane od najmniejszych po największe. Zachowując to uporządkowanie – na ile to możliwe – rozmieszcza się obiekty w q -wymiarowej przestrzeni i oblicza **współczynnik wierności odwzorowania** (*Stress*). Następnie nieco zmienia się położenie punktów i znów oblicza współczynnik: gdy ma wartość niższą, to nowy układ uznaje się za lepszy, gdy wyższą – poprzedni układ jest właściwszy. Kolejne iteracje powtarza się tak długo, aż niemożliwe jest dalsze zmniejszenie wartości współczynnika. Najczęściej używa się jednego z dwóch współczynników jakości odwzorowania: **Stress(q)** (Kruskal 1954a) i **SStress** (Takane i inni 1977):

$$\text{Stress}(q) = \left\{ \frac{\sum_{i < k} \sum [d_{ik}^{(q)} - \hat{d}_{ik}^{(q)}]^2}{\sum_{i < k} [d_{ik}^{(q)}]^2} \right\}^{1/2} \quad \text{SSStress} = \left[\frac{\sum_{i < k} \sum (d_{ik}^2 - \hat{d}_{ik}^2)^2}{\sum_{i < k} \sum d_{ik}^4} \right]^{1/2},$$

$\hat{d}_{ik}^{(q)}$ to odległości w q -wymiarowej przestrzeni, spełniające sformułowany wyżej warunek monotoniczności; nie są to odległości w tym sensie, jak opisane w Rozdziałach 2.9–2.12, a raczej kolejne numery na liście, porządkującej odległości według ich wartości. Jak łatwo zauważyć, dla wszystkich odległości spełniających ten warunek mianownik w obu przypadkach będzie równy zero, a więc doskonałe odwzorowanie odpowiadać będzie wartościom zerowym obu współczynników, gdy zupełne niezachowanie kolejności – wartości 1. Dla Stress(q) Kruskal zaproponował umowną interpretację wartości: 0,2 – odwzorowanie złe, 0,1 – odwzorowanie słabe, 0,05 – odwzorowanie dobre, 0,025 – odwzorowanie znakomite, 0,000 – doskonałe. Dla SStress wartości poniżej 0,1 uważa się za wskazujące na dobre odwzorowanie.

Obiekty rozmieszczone w wyniku nieliniowego MDS w q -wymiarowej przestrzeni traktować możemy jako q -wymiarowe obserwacje (ich koordynaty odpowiadają wartościom kolejnych „cech”), które graficznie najdogodniej przedstawić w przestrzeni ich głównych składowych. Wraz z rosnącą liczbą wymiarów wartości Stress(q) czy SStress maleją, by dla $q = n - 1$ osiągnąć zero. Brak jakiegś reguły określającej, ile wymiarów należy uwzględnić – wszystko zależy od danych. Oczywiście im mniej, tym lepiej, a dla graficznego przedstawienia najlepiej, aby to były dwa, co najwyżej trzy wymiary

(Ryc. 3.6). Zaczynamy więc od dwóch, jeżeli nie uda się uzyskać dostatecznie niskich wartości $\text{Stress}(q)$ bądź SStress , to wprowadzamy trzeci wymiar. Gdy to nadal nie wystarczy, warto sporządzić wykres najniższych wartości tych współczynników dla kolejnych wartości q . Wykres ten będzie miał postać omawianego przy analizie głównych składowych *scree plot*; podobnie jak tam, wartości „ze strefy akumulacji osadu” można już nie uwzględniać. Gdy wymiarów musi być więcej niż trzy, to warto przeprowadzić na finalnej konfiguracji analizę głównych składowych, aby w niższej liczbie wymiarów uwidocznili główne tendencje w grupowaniu obiektów.



Ryc. 3.6. Nieliniowe skalowanie wielowymiarowe, obliczone dla wartości θ między parami populacji *Bythinella* z Europy Środkowej. Grupowanie kolejnych populacji (1–20) przedstawiono w przestrzeni trzech pierwszych wymiarów, a dla wskazania lokalnych dystorsji na wykres naniesiono najkrótsze drzewo połączeń. Najbardziej odrębne są populacje 20 i 5, odrębne są też 1, 2, 3, 6 i 12. Pozostałe, poza nieco bardziej odrębnymi 13, 18 i 19, leżą skupione blisko centroidu. Warto zwrócić uwagę, że np. populacje 1 i 3 w rzeczywistości nie są tak sobie bliskie jak na rysunku – najkrótsze drzewo połączeń bezpośrednio ich nie łączy, podobnie 18 i 19, 12 i 19 czy 6 i 13. Za: Falniowski i inni (1999)

Nieliniowe MDS, jak wspominaliśmy, przeprowadzić można nawet na umownych rangach. Jest też mniej wrażliwe niż analiza głównych składowych, a nawet analiza głównych współrzędnych, na brak części danych (Rohlf 1994). Inną zaletą nieliniowego MDS jest wierne odtwarzanie zarówno większych odległości pomiędzy całymi skupiskami (podobnie jak w analizie głównych składowych), jak i niewielkich odległości pomiędzy obiektami wewnątrz samych skupisk (podobnie jak w analizie skupisk, natomiast inaczej niż w analizie głównych składowych: Sneath i Sokal 1973).

Nieliniowe MDS nie jest jednak wolne od wad. Obliczeniowo intensywne, działa na zasadzie iteracyjnej minimalizacji wartości zadanej współczynnika – jak w każdym przypadku tego typu, algorytm znajduje tzw. lokalne optimum, poza które wyjść nie może. Uzyskana najniższa wartość $\text{Stress}(q)$ bądź SStress bynajmniej nie musi być najniższą możliwą dla analizowanych danych, po prostu jakiegokolwiek dalsze niewielkie modyfikacje uzyskanego układu obiektów nie przyniosą zmniejszenia wartości współczynnika. Konieczne jest przeprowadzenie kolejnych obliczeń od początku, wychodząc od innej początkowej konfiguracji, wówczas wartość może być niższa. Podobne problemy napotkamy dalej, przy poszukiwaniu drzew najlepszych pod względem przyjętego kryterium optymalizacji. I tu i tam rozwiązaniem jest wielokrotne powtarzanie obliczeń. Dla nieliniowego MDS pięć powtórzeń to minimum, a im więcej, tym lepiej – kilkadziesiąt powinno na ogół wystarczyć. Warto też zacząć obliczenia od wstępnej konfiguracji, uzyskanej z analizy głównych składowych bądź analizy głównych współrzędnych. Wówczas mamy pewność, że wynik nieliniowego MDS przynajmniej nie będzie gorszy od konfiguracji uzyskanej tymi metodami. Wreszcie zdarzyć się może, że uzyskamy wartość $\text{Stress}(q)$ bądź SStress bliską zera, podczas gdy odwzorowanie będzie bardzo niedoskonałe. Bywa tak wówczas, gdy wyjściowe odległości mają wartości silnie skupione w paru przedziałach. Aby uniknąć takiego błędu, warto sprawdzić monotoniczność transformacji, oglądając położenie punktów w układzie współrzędnych, którego osią odciętych są wartości oryginalnych, a rzędnych – transformowanych odległości (Rohlf 1994).

3.5. Analiza odpowiadania

Analiza odpowiadania (*correspondence analysis* – CA) należy do technik graficznej reprezentacji związków pomiędzy zmiennymi w przestrzeni o niewielu wymiarach: można ją więc zaliczyć do technik wielowymiarowego skalowania, choć w pewnych aspektach przypomina też analizę głównych składowych. Analizę głównych składowych przeprowadza się dla n obiektów, dla których znamy wartości m cech, podczas gdy analizę odpowiadania prowadzi się na **tablicy wielodzielnej**, zestawiającej **częstości** lub **liczebności**. Wymienionym wcześniej przykładem takich danych mogą być częstości poszczególnych alozymów w kolejnych *loci* u różnych gatunków, ale również udziały różnych gatunków w próbach z różnych stanowisk, częstości różnych barw upierzenia u kolejnych gatunków ptaków czy udziały poszczególnych pierwiastków w strukturach zmineralizowanych różnych taksonów.

Tablica wielodzielna liczy I rzędów i J kolumn, a obiektem analizy są związki między kolejnymi rzędami – grupowanie rzędów, między kolejnymi kolumnami – grupowanie kolumn, a także związki między kolumnami a rzędami. Wynikiem będą wykresy złożone odpowiednio z I punktów odpowiadających rzędom oraz J punktów reprezentujących kolumny, a wzajemne położenie punktów będzie odzwierciedleniem związków: bliskie sobie punkty reprezentujące rzędy wskazywać będą rzędy o podobnych profilach (podobnym rozkładzie wartości w kolejnych kolumnach), a położone koło siebie punkty odpowiadające kolumnom – kolumny o podobnych profilach (podobnym rozkładzie wartości w kolejnych rzędach). Ponadto, gdy punkty odpowiadające określonym kolumnom są bliskie punktom odpowiadającym określonym rzędom, to wskazuje to na kombinację spotykaną częściej niż gdyby wartości w rzędach nie miały

związku z wartościami w kolumnach. W analizie głównych składowych odległości Euklidesowe kwantyfikują związki pomiędzy obiektami, a korelacje bądź kowariancje – pomiędzy cechami, gdy w analizie odpowiadania związki pomiędzy poszczególnymi kolumnami, a także pomiędzy poszczególnymi rzędami kwantyfikuje ta sama odległość χ^2 , określona dla kolumn (rzędów) i i j :

$$d_{ij(\chi^2)} = \sqrt{\sum_{k=1}^k \frac{\left(\frac{x_{ik}}{x_{.i}} - \frac{x_{kj}}{x_{.j}} \right)^2}{x_k}},$$

gdzie $x_{.i}$ – średnia wartość dla kolumny (rzędu) i , $x_{.j}$ – średnia wartość dla kolumny (rzędu) j . Można by więc przeprowadzić analizę głównych współrzędnych, korzystając z macierzy odległości χ^2 , obliczonych między wszystkimi rzędami i między wszystkimi kolumnami. Wyniki takiej analizy będą jednak inne niż dla analizy odpowiadania, bowiem w analizie odpowiadania obiekty są wazone, zgodnie z ich częstościami. W analizie odpowiadania dane dopasowuje się do następującego modelu (Rohlf 1994):

$$f_{ij} = \sqrt{f_{.i} f_{.j}} \left(1 + \sum_{k=1}^k \sqrt{\lambda_k} \psi_{ik} \phi_{kj} \right),$$

gdzie f_{ij} to obserwowane względne częstości ($x_{ij}/x_{.i}$), $f_{.i}$ i $f_{.j}$ to względne częstości rzędów i kolumn, λ_k to k -ta wartość własna, a ψ_{ik} i ϕ_{kj} są elementami macierzy czynników (*factors*) rzędów i kolumn. Gdyby wartości własne były równe zeru, to kolumny i rzędy macierzy byłyby całkowicie niezależne.

Analizę związków występujących w tablicy wielozdzielnej \mathbf{X} rozpoczyna się od odpowiedniego centrowania i skalowania macierzy o I rzędach i J kolumnach ($I > J$). Jeżeli sumę częstości w macierzy \mathbf{X} oznaczmy n , to nową macierz \mathbf{P} , zwaną macierzą odpowiadania (*correspondence matrix*), konstruujemy, dzieląc każdy z elementów \mathbf{X}

przez n : $p_{ij} = \frac{x_{ij}}{n}$, dla $i = 1, 2, \dots, I, j = 1, 2, \dots, J$, czyli $P = \frac{1}{n} X$. Macierz odpowiadania centrujemy. W analizie głównych współrzędnych macierz odległości bądź podobieństw centrowaliśmy, odejmując średnie dla rzędu i kolumny, po czym dodając średnią dla całej macierzy. W analizie odpowiadania natomiast dla każdej wartości z macierzy odpowiadania odejmujemy iloczyn sum rzędu i kolumny, do których należy ta wartość:

$$\tilde{p}_{ij} = p_{ij} - r_i c_j, \text{ dla } i = 1, 2, \dots, I, j = 1, 2, \dots, J, \text{ czyli } \tilde{P} = P - rc', \text{ gdzie:}$$

$$r_i = \sum_{j=1}^J p_{ij} = \sum_{j=1}^J \frac{x_{ij}}{n}, \text{ dla } i = 1, 2, \dots, I, \text{ czyli } \underset{(Ix1)}{r} = \underset{(IxJ)(Jx1)}{P} \underset{(Jx1)}{1},$$

$$c_j = \sum_{i=1}^I p_{ij} = \sum_{i=1}^I \frac{x_{ij}}{n}, \text{ dla } j = 1, 2, \dots, J, \text{ czyli } \underset{(J \times 1)}{c} = \underset{(J \times I)}{P'} \underset{(I \times 1)}{1}.$$

Następnie definiujemy macierze diagonalne: $\mathbf{D}_r = \text{diag}(r_1, r_2, \dots, r_I)$, $\mathbf{D}_c = \text{diag}(c_1, c_2, \dots, c_J)$, po czym konstruujemy macierz skalowaną: $\underset{(I \times J)}{\mathbf{P}^*} = \underset{(I \times I)}{\mathbf{D}_r}^{-1/2} \underset{(I \times J)}{\tilde{\mathbf{P}}} \underset{(J \times J)}{\mathbf{D}_c}^{-1/2}$, w której dla elementu (i, j) :

$$p_{ij}^* = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}.$$

Opis dalszych operacji wymagałby od Czytelnika znajomości rachunku macierzowego, toteż zainteresowanych odsyłam do literatury (Greenacre 1984, Lebart i inni 1984, Johnson i Wichern 1998). Dekompozycja macierzy \mathbf{P}^* pozwala obliczyć macierz $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{J-1})$, zawierającą kolejne wartości własne macierzy \mathbf{P}^* , uporządkowane od największej po najmniejszą. Dalsze operacje na macierzach prowadzą do obliczenia macierzy czynników ψ_{ij} i ϕ_k rzędów i kolumn. W ten sposób zdefiniowany zostaje układ współrzędnych, w który rzutowane będą punkty, odpowiadające kolejnym kolumnom i rzędom. Następnym krokiem jest obliczenie współrzędnych tych punktów. Pierwsze dwie kolumny macierzy – zarówno dla kolumn jak i rzędów macierzy wyjściowej – to współrzędne w dwóch wymiarach najpełniej (choć oczywiście niekompletnie) obrazujących zależności w przestrzeni dwuwymiarowej. W obrębie punktów obrazujących kolumny bądź wśród punktów reprezentujących rzędy, odległości χ^2 na wykresie odzwierciedlają liniowo różnice między kolumnami bądź między rzędami. Skalowanie jest to samo dla rzędów i kolumn, a więc można je przedstawić razem na tym samym wykresie – co też zwykle się robi i co wskazuje w przejrzysty sposób zależności między rzędami i kolumnami. Warto jednak podkreślić, że analiza odpowiadania bezpośrednio nie obrazuje bliskości między rzędami a kolumnami poprzez zróżnicowane odległości na wykresie (Johnson i Wichern 1998), bezpośrednio miary takiej odległości ta technika nie definiuje ani nie wykorzystuje. Dlatego też zależności pomiędzy kolumnami a rzędami można rozpatrywać, porównując rozmieszczenie wszystkich punktów jednej kategorii z rozmieszczeniem punktów drugiej kategorii, natomiast wnioskowanie na podstawie bliskości konkretnych punktów – reprezentującego kolumnę i reprezentującego rząd – jest ryzykowne (Lebart i inni 1984).

Podobnie jak w analizie głównych składowych, graficzna reprezentacja w przestrzeni dwuwymiarowej przedstawi jedynie część informacji zawartej w danych wyjściowych. W analizie odpowiadania część informacji, reprezentowana w kolejnych wymiarach, czyli odpowiadająca kolejnym czynnikom, nosi nazwę *inercji* (*inertia*), przy czym inercja całkowita:

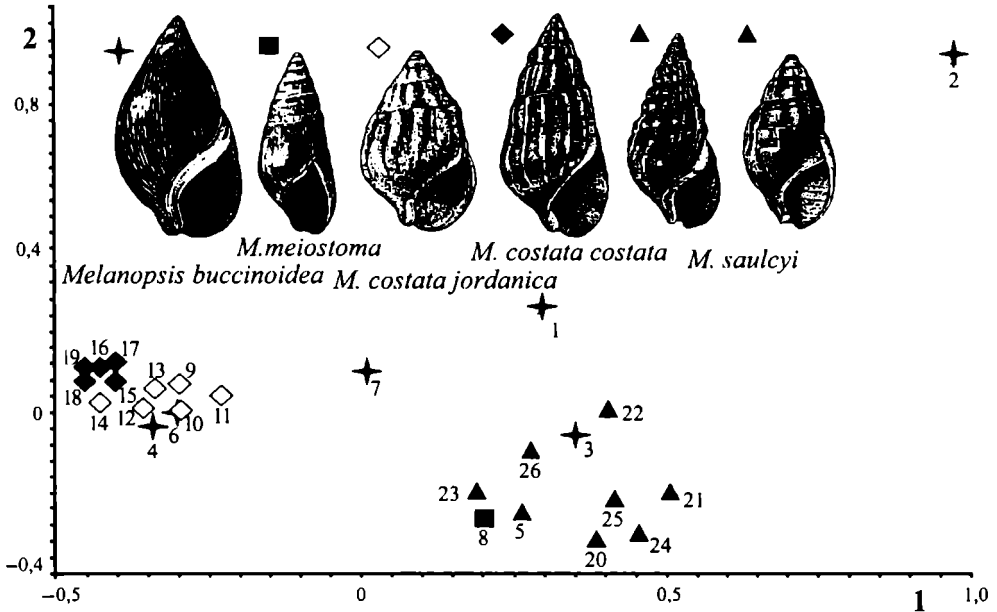
$$\text{Inercja}_c = \sum_{i=1}^K \lambda_i^2, \text{ gdzie } \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K, \text{ a } K = \min(I - 1, J - 1)$$

Całkowita inercja jest ważoną sumą kwadratów odległości od centroidów punktów reprezentujących rzędy (kolumny). Jest to więc miara ogólnej zmienności, czyli ogólnego zróżnicowania. Można wykazać, że całkowita inercja dla punktów reprezentujących rzędy jest równa całkowitej inercji dla punktów reprezentujących kolumny (Greenacre 1984). Wazona suma kwadratów koordynatów punktów przedstawiających rzędy, mierzonych na k -tej osi, która jest równa ważonej sumie kwadratów koordynatów punktów przedstawiających kolumny, mierzonych na tej samej osi, to λ_k^2 , nazywana **k -tą główną inercją** (*principal inertia*). Im wyższa wartość proporcji $\lambda_k^2 / \text{Inercja}_c$, tym więcej ogólnej zmienności tłumaczy k -ty czynnik; gdy $(\lambda_1^2 + \lambda_2^2) / \text{Inercja}_c$ przyjmuje wysoką wartość (powiedzmy, 70% czy więcej), to wskazuje to na dobrą reprezentację wyjściowych danych w tych dwóch wymiarach, na przedstawianie w nich niemal całej zmienności, której nie da się wytłumaczyć, przyjmując model niezależności wartości w rzędach od wartości w kolumnach w wyjściowej macierzy. **Ogólną inercję** można bowiem także zdefiniować:

$$\text{Inercja}_c = \sum_i \sum_j (p_{ij} - r_i c_j)^2 / r_i c_j = \chi^2 / n, \text{ przy czym: } \chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

gdzie O_{ij} – wartość obserwowana częstości i, j , $E_{ij} = nr_i c_j$ – wartość oczekiwana częstości i, j , gdyby wartości w kolumnach były niezależne od wartości w rzędach. Inercja jest więc wprost proporcjonalna do wartości χ^2 , kwantyfikującej związek – inaczej odstępstwa od niezależności – pomiędzy wartościami w kolumnach a wartościami w rzędach. Znowu więc widać, że analiza odpowiadania byłaby niemożliwa, gdyby związki między wartościami występującymi w rzędach a stwierdzanymi w kolumnach były wyłącznie następstwem przypadku. Oczywiście χ^2 nie należy mylić ze zdefiniowaną wcześniej odległością χ^2 , choć matematycznie są to wielkości pokrewne: odległość χ^2 określa wzajemne położenie punktów, reprezentujących rzędy bądź kolumny, natomiast test χ^2 kwantyfikuje związki między kolumnami i rzędami.

Wynikiem analizy odpowiadania jest wykres obrazujący grupowanie rzędów oraz kolumn, który można wykorzystać dla znalezienia grup, nieokreślonych *a priori* i często niespodziewanych (Ryc. 3.7). Wskaże też zależności pomiędzy rzędami i kolumnami – choć z zastrzeżeniami wcześniej sformułowanymi – a więc bliskość, powiedzmy, punktu „żółty” koło punktów „gatunek A” i „gatunek B” wskazuje, że bliskość tych gatunków na wykresie wynika w dużym stopniu z występowania u obu z nich żółtego upierzenia. Wspólność wektorów własnych dla rzędów i kolumn umożliwia też porównywanie zmienności, mającej dla obu kategorii ten sam kierunek na wykresie. Poza tym warto przyrzeć się wartościom χ^2 , kwantyfikującym związek dla kolejnych par rzędów i kolumn. Analogicznie jak w analizie głównych składowych mamy miary udziału kolejnych zmiennych – tutaj rzędów i kolumn, choć oczywiście merytorycznie często możemy bądź rzędy, bądź kolumny uważać bardziej za obiekty niż zmienne – w tworzeniu kolejnych czynników/wymiarów (odpowiedników głównych składowych w PCA). Będą to zarówno absolutne udziały, wskazujące, jaki procent danego czynnika tłumaczy dana zmienna, jak i kwadraty współczynnika korelacji, wskazujące dla kolejnych zmiennych istotność kolejnych czynników. Inercja dla poszczególnych czynników kwantyfikuje ich zawartość informacyjną.

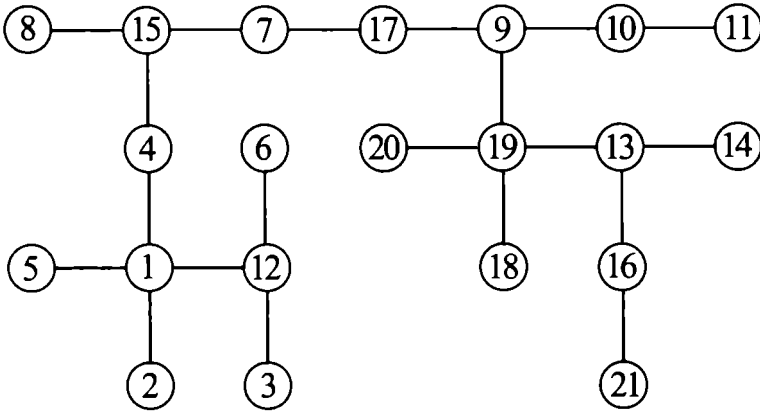


Ryc. 3.7. Analiza odpowiadania przeprowadzona dla częstości alleli u czterech gatunków (jeden reprezentowany przez dwa podgatunki) *Melanopsis* z rzeki Jordan i jeziora Genezaret. Obiekty, czyli populacje, rzutowane są w przestrzeń pierwszego i drugiego wymiaru. Widać, że zróżnicowanie molekularne *M. buccinoidea* obejmuje pełny zakres zmienności dla wszystkich gatunków: możliwe, że procesy specjacji nie zakończyły się jeszcze oraz że *M. buccinoidea* najprawdopodobniej jest gatunkiem ancestralnym dla pozostałych. *M. meiostroma* nie różni się od *M. saulcyi*, oba różnią się wyraźnie od *M. costata*. W obrębie tego ostatniego podgatunku różnią się nieznacznie, lecz nie mieszają. Za: Falniowski i inni (2002)

3.6. Najkrótsze drzewo połączeń

Jak pamiętamy, dla t obiektów (OTU), z których każdy opisuje m cech, obliczyć można macierz $t(t-1)/2$ odległości. Teoretycznie można by sobie wyobrazić graf, na którym t punktów (*vertices*) łączy $t(t-1)/2$ odcinków, zwanych **wiązadłami** (*edges*), o długościach proporcjonalnych do odpowiadających im odległości. Taki **maksymalnie połączony** (*maximally connected*) graf przedstawiłby wszystkie różnice/podobieństwa między obiektami, niekoniecznie w czytelny sposób. Tryb warunkowy powyższych zdań wynika stąd, że praktycznie nigdy nie udałoby się wiernie odtworzyć wszystkich odległości na płaszczyźnie – wystarczy wyobrazić sobie przedstawienie na płaszczyźnie choćby wzajemnego położenia wszystkich wierzchołków wielościanu, zachowujące proporcje pomiędzy wszystkimi odległościami. Można też założyć określoną wartość krytyczną odległości, poniżej której punkty łączymy – w ten sposób uzyskamy **połączony** (*connected*) **graf**, również niekoniecznie prawidłowo odwzorowujący wyjściowe odległości, ale informujący o związkach/różnicach przekraczających wartość krytyczną. Ostatnim wariantem będzie **graf minimalnie połączony** (*minimally*

connected), czyli drzewkowaty **dendryt**, z $t - 1$ wiązań. Dendryt jest linią łamaną rozgałęziającą się, lecz niezawierającą łamanych zamkniętych, łączącą wszystkie elementy klasyfikowanego zbioru obiektów. Długość wiązań odpowiada odległościom pomiędzy parami obiektów. Dendryt umożliwia nieliniowe uporządkowanie obiektów na płaszczyźnie.



Ryc. 3.8. Najkrótsze drzewo połączeń dla 21 OTU. Choć w wielu miejscach politomiczny i nienadający się do interpretacji filogenetycznej, dendryt dobrze ilustruje złożone związki między obiektami

Najkrótsze drzewo połączeń (*minimum spanning tree* – MST) w literaturze często nazywane jest „siecią Prima” (*Prim network*), choć nie jest to nazwa właściwa: dendryt, choć najczęściej nie ma formy filogenetycznych drzew, jest jednak zgodnie z matematyczną terminologią drzewem, nie siecią. Prim (1957) nie był też tym, który stworzył tę technikę, a jedynie jednym z pierwszych, którzy sformułowali algorytm obliczania takiego dendrytu (podany np. w: Pocięcha i inni 1988). Twórcami metody byli statystycy tzw. szkoły wrocławskiej (Florek i inni 1952a, b). Obserwacje są wierzchołkami grafu, a łączy się je wtedy, gdy pomiędzy danymi obserwacjami istnieje relacja **najbliższego sąsiedztwa**, czyli odległość między daną obserwacją a obserwacją, z którą została połączona, jest najmniejszą z odległości danej obserwacji do wszystkich pozostałych. Najlepszy dendryt to dendryt o najmniejszej długości. W macierzy odległości wybieramy najmniejszą wartość i punkty odpowiadające parze obiektów, pomiędzy którymi wystąpiła ta wartość, łączymy wiązadłem o długości odpowiadającej (proporcjonalnej do) tej odległości. Następnie poszukujemy najmniejszych odległości dla obu obiektów i analogicznie dołączamy kolejne obiekty wiązadłami odpowiednich długości (Ryc. 3.8).

Najkrótsze drzewo połączeń, jako niezakładające dychotomicznego grupowania, dobrze ilustruje złożoność całego układu podobieństw pomiędzy klasyfikowanymi obiektami, niewidaczniającą się w **analizie skupisk** (*clustering*). Warto więc wykorzystać je dla sprawdzenia wyników analizy skupisk i lepszego zrozumienia zależności pomiędzy badanymi obiektami. Ponadto, jak już wspominaliśmy, analiza głównych składowych (ale, w mniejszym stopniu, także analiza głównych współrzędnych i nieli-

niowe skalowanie wielowymiarowe) często nie przedstawia wiernie wzajemnego położenia bliskich sobie obiektów. Ponadto obiekty bliskie sobie w danych dwóch-trzech wymiarach niekoniecznie muszą być sobie bliskie w przestrzeni innych głównych składowych. Dlatego warto rzutować najkrótsze drzewo połączeń na projekcję obiektów w przestrzeni głównych składowych (Ryc. 3.6). Umożliwia to wykrycie lokalnych dystorsji, czyli wskazanie przypadków, gdy obiekty bliskie sobie w przestrzeni danych składowych w rzeczywistości bliskie sobie nie są, a przynajmniej być nie muszą. Stosując najkrótsze drzewo połączeń, pamiętać musimy, że topologia dendrytu może być bardzo różna, zależnie od tego, jakiej odległości użyjemy. Nie wolno więc wybierać odległości mechanicznie, konieczne jest staranne rozważenie, która odległość będzie najodpowiedniejsza dla danych, którymi dysponujemy – w tym miejscu odsyłamy do rozważań zawartych w Części 2.

3.7. Analiza skupisk, odległości kofenetyczne

Analiza skupisk (*clustering*) to najczęściej stosowana i najszerzej znana technika taksonomii fenetycznej. W najpowszechniej stosowanej odmianie wynikiem jest dychotomiczne drzewo, znakomicie nadające się do interpretacji w kategoriach biologicznej klasyfikacji, choć jako odzwierciedlające pokrewieństwa jedynie wówczas, gdy dane są ultrametryczne (co – jak wiemy – bywa rzadko). Niewątpliwie nadużywana przez dziesięciolecia, obecnie technika ta zaczyna tracić na znaczeniu, niemniej wciąż słuszne wydaje się omówienie jej nieco dokładniej, tym bardziej że jest prosta, zwłaszcza w porównaniu z większością technik MSA. W odróżnieniu od nich jest też możliwa bez komputera, a nawet i bez kalkulatora, jeżeli zestaw danych nie jest za duży. Oczywiście i tak przeprowadza się ją obecnie przy użyciu komputera, warto jednak wiedzieć dokładniej, jak działa algorytm.

Najogólniej ujmując, analiza skupisk zakłada, że zbiór n obiektów, czyli OTU zaklasyfikować można do k grup (*clusters*), przy czym $1 \leq k \leq n$. Wartość k nie jest wyjściowo znana. Teoretycznie analizę skupisk przeprowadzić można techniką **podziałową** (*divisive*) bądź **aglomeracyjną** (*agglomerative*). Techniki aglomeracyjne polegają na dołączaniu, jeden po drugim, kolejnych obiektów, natomiast podziałowe na kolejnym dzieleniu zbioru obiektów. Ogólną zasadą metod podziałowych jest maksymalizacja wariancji pomiędzy wydzielonymi skupiskami (grupami), przy jednoczesnej minimalizacji wariancji w obrębie tych skupisk (grup). Aglomeracyjne są znacznie częściej stosowane, bowiem algorytmy ich obliczania są prostsze, ponadto dużym mankamentem technik podziałowych jest to, że raz błędnie oddzielone od siebie obiekty pozostaną już w odrębnych grupach (Sneath i Sokal 1973); ograniczymy się więc do omawiania aglomeracyjnych, tym bardziej że wyniki analiz prawidłowo przeprowadzonych obu technikami będą – teoretycznie – takie same. Grupowanie może być prowadzone techniką **sekwencyjną** (*sequential*), czyli **krokową** (Pociecha i inni 1988) lub **jednoczesną** (*simultaneous*), a więc dołączane mogą być kolejne obiekty, bądź wszystkie skupiska będą powstawały jednocześnie. Techniki sekwencyjne są prostsze i powszechnie stosowane (Anderberg 1973, Everitt 1993), z drugiej strony projekcję obiektów w przestrzeni głównych składowych czy wielowymiarowe skalowanie oznaczają za swego rodzaju analizę skupisk, prowadzoną techniką jednoczesną.

Grupowanie może być **hierarchiczne** (*hierarchical*) lub **niehierarchiczne** (*nonhierarchical*). Hierarchiczne zakłada, że dwa elementy klasyfikowanego zbioru łączą się z sobą na określonym poziomie, odzwierciedlającym odległość pomiędzy nimi; elementy zbioru klasyfikuje się w zestaw różnicznych grup, te grupy w odpowiednio mniej liczny zestaw grup nadrzędnych w stosunku do tych pierwszych itd., aż cały zbiór zostanie sklasyfikowany. Klasyfikacja tego typu jest dla nas pożądana, bowiem podobnie hierarchiczna jest klasyfikacja biologiczna: populacje grupujemy w gatunki, gatunki w rodzaje, rodzaje w rodziny itd., aż po królestwa świata żywego. Metody niehierarchiczne nie zakładają rankingu grup, porzeczając na wskazaniu skupisk, dających się wyróżnić w klasyfikowanym zbiorze. Przykładem takiej techniki jest metoda **k-średnich** (*K-means*) wprowadzona przez MacQueena (1967). Zbiór dzieli się wstępnie na k grup, arbitralnie wliczając obiekty do tych grup. Dla każdej z grup obliczamy średnią (położenie centroidu). Następnie obliczamy zestaw odległości (zwykle Euklidesowych) pomiędzy poszczególnymi elementami zbioru a kolejnymi centroidami. Przemieszczamy elementy pomiędzy grupami, tak aby każdy z elementów był w grupie, której centroid jest mu najbliższy. Obliczamy nowe średnie dla grup, nowe odległości i dokonujemy kolejnych przemieszczeń, aż dalsze przemieszczenia nie poprawiają odległości obiektów od centroidu. Zarazem jeżeli jakaś para centroidów leży zbyt blisko siebie, to grupy łączymy w jedną, natomiast jeżeli grupa okazuje się niejednorodna, to musimy ją podzielić. Istotne, aby wartość k była ostatecznie wynikiem analizy, a nie dana *a priori* (Johnson i Wichern 1998).

Analiza skupisk może zakładać **zachodzenie** na siebie (*overlapping*) bądź **niezachodzenie** (*nonoverlapping*) wyróżnionych grup. Niezachodzenie oznacza wykluczenie, tzn. jakikolwiek takson nie może należeć do więcej niż jednego taksonu określonej, wyższej rangi. Znow jest to zgodne z zasadami klasyfikacji biologicznej, choć niekoniecznie musi odzwierciedlać prawidłowości dające się zaobserwować w analizowanych danych. Niewątpliwie fenogram, jak każde drzewo, nie jest w stanie przedstawić wszystkich związków między obiektami i z konieczności jest zawsze uproszczeniem, niekiedy bardzo dużym. Dopiero jednak takie uproszczenie nadaje się bezpośrednio do interpretacji w kategoriach klasyfikacji biologicznej. Zarazem bogactwo zależności między obiektami dobrze ilustrują inne, omówione wcześniej techniki, takie jak analiza głównych składowych, analiza głównych współrzędnych czy nieliniowe skalowanie wielowymiarowe. W taksonomii numerycznej używa się więc niemal wyłącznie metod **SAHN clustering** (Sneath i Sokal 1973, Rohlf 1994), co oznacza *sequential agglomerative hierarchical nonoverlapping clustering* – sekwencyjną (krokową) aglomeracyjną hierarchiczną analizę niezachodzących na siebie skupisk.

SAHN clustering przeprowadza się na **macierzy odległości** (choć oczywiście można i na macierzy podobieństw) pomiędzy obiektami. Jest więc oczywiste, że wyniki zależą od rodzaju użytej odległości (podobieństwa) i należy starannie rozważyć, która z odległości najlepiej odzwierciedli rzeczywiste związki w obrębie badanej grupy obiektów. Warto też zaznaczyć, że choć typowym zastosowaniem analizy skupisk jest klasyfikacja obiektów, to można jej użyć – i niekiedy używa się – do analizy związków między cechami. Zaczynamy od n obiektów i kolejno łączymy obiekty najbliższe, aż na fenogramie – połączone – znajdzie się wszystkie n obiektów. (1) Najpierw znajdujemy w macierzy najniższą wartość odległości (lub najwyższą wartość współczynnika asocjacji). (2) Obiekty (grupy) I i J , pomiędzy którymi wystąpiła ta wartość, łączymy

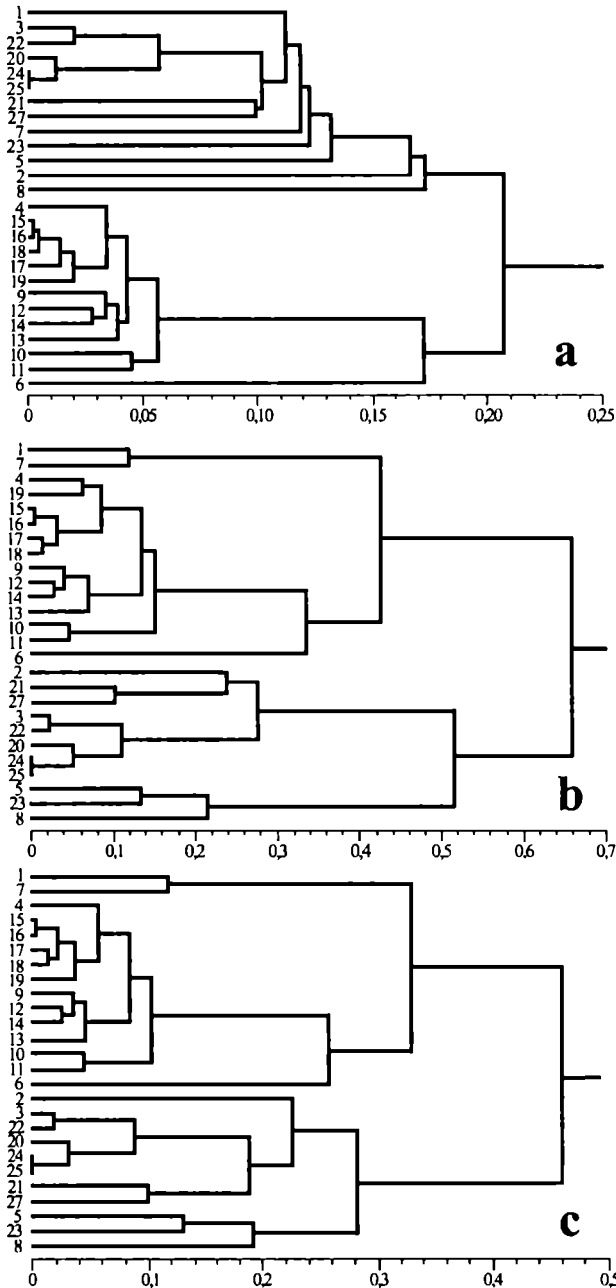
w grupę (skupisko, *cluster*), przy czym głębokość rozgałęzienia (długość gałęzi) wyznaczamy jako $l_{ij} = d_{ij}/2$.

Może się zdarzyć, że w macierzy będzie więcej niż jedna identyczna wartość, na jakimś etapie analizy skupisk najmniejsza. Bywa tak zwłaszcza wtedy, gdy analizę prowadzimy dla niewielkiej liczby cech. W dodatku wartości nie muszą być identyczne – wystarczy, by różniły się nieznacznie, powiedzmy, w granicach błędu standardowego dla odległości w danej macierzy, a nawet nieco więcej, aby należało traktować je jako identyczne. Wówczas uzyskany fenogram będzie różny, zależnie od tego, który z obiektów dołączymy najpierw. Brak właściwie jakiegóż rutynowej metody postępowania w takim wypadku (Rohlf 1994) – algorytm zwykle po prostu wybiera tę wartość, która w macierzy pojawia się pierwsza. Najwłaściwsze wydaje się jednak obliczenie fenogramów osobno dla każdej możliwej kolejności dołączania tych „równoodległych” obiektów, a następnie zestawienie wszystkich fenogramów, aby znaleźć grupy, które występują na każdym z nich. (3) Jeżeli I i J były ostatnimi OTU bądź grupami (*clusters*), to drzewo jest kompletne. Jeżeli nie, to dla I i J tworzymy nową grupę (*cluster*) U . (4) Wyznaczamy odległość od U do każdego z pozostałych OTU bądź grup K , ($K \neq I, K \neq J$). (5) Wracamy do kroku (1), przy czym mamy do sklasyfikowania o jeden mniej OTU (grupę): I i J zostały wyeliminowane, natomiast U dodane.

Algorytm ten jest wspólny dla wszystkich technik SAHN, różnice występują jedynie w kroku (4), czyli w wyznaczaniu odległości nowo utworzonej grupy od pozostałych grup (OTU). Metoda **najbliższego sąsiedztwa** (*single linkage clustering*, inaczej *nearest neighbor technique*, *minimum method*, *minimum distance method*), wprowadzona przez Florka i innych (1951a, b), przyjmuje, że $d_{KU} = \min(d_{KI}, d_{KJ})$. Oznacza to, że odległością między grupami jest najmniejsza z odległości, jakie występują pomiędzy obiektami (grupami) nienależącymi do tej samej z rozważanych grup (uwaga na hierarchiczny system grup – *clusters* – stąd grupy należące do grupy, należącej do grupy itd. – językowo jest to niejasne, ale właśnie odzwierciedla hierarchiczność).

Obliczeniowo prosta – nie wymaga przeliczania macierzy – metoda najbliższego sąsiedztwa ścieśnia przestrzeń, dając klasyfikację prostszą, o relatywnie małej liczbie poziomów (Ryc. 3.9a). Technika ta nie jest w stanie wyróżnić słabo odrębnych grup, z drugiej strony jest to jedna z niewielu odmian analizy skupisk, która jest w stanie wykazać nieeliptyczność grup tworzących skupiska, wyodrębniając długie, liniowo ułożone grupy (Johnson i Wichern 1998). Tak długo, jak różne odległości są względem siebie monotoniczne, zastąpienie jednej odległości inną nie zmienia wyniku analizy metodą najbliższego sąsiedztwa, tak samo jest w metodzie najdalszego sąsiedztwa, lecz nie w technikach opartych na średnich.

Metoda **najdalszego sąsiedztwa** (*complete linkage clustering*, inaczej *farthest neighbor technique*, *maximum method*, *maximum distance method*) przyjmuje, że $d_{KU} = \max(d_{KI}, d_{KJ})$. Odległością między grupami jest tu więc największa z odległości, jakie występują pomiędzy obiektami (grupami) nienależącymi do tej samej z rozważanych grup. Jako proste przeciwieństwo metody najbliższego sąsiedztwa, także nie wymaga przeliczania macierzy, natomiast odwrotnie niż tamta rozszerza przestrzeń, tworząc klasyfikację bardziej złożoną, bogatszą, o większej liczbie poziomów (Ryc. 3.9b). Aby ominąć obie skrajności, Lance i Williams (1967b) zaproponowali metodę **elastycznego grupowania SAHN** (*flexible clustering strategy*), opartą na ich ogólnej formule liniowej kombinatorycznej strategii:



Ryc. 3.9. Analiza skupisk przeprowadzona dla tych samych 26 populacji *Melanopsis co* na Ryc. 3.7, jako miary odległości użyto wartości θ między parami populacji. Metoda najbliższego sąsiedztwa (a) daje grupowanie płytkie, o niższej liczbie poziomów w hierarchii, podczas gdy metoda najdalszego sąsiedztwa (b) tworzy hierarchię o większej liczbie poziomów, grupowanie bardziej złożone, rozszerzając przestrzeń. Technika nieważonej średniej arytmetycznej UPGMA (c), najczęściej stosowana, nie ścieśnia ani nie rozszerza przestrzeni i ogólnie daje podobne grupowanie jak technika najdalszego sąsiedztwa

$$U_{(J,K),L} = \alpha_J U_{J,L} + \alpha_K U_{K,L} + \beta U_{J,K} + \gamma(U_{J,L} - U_{K,L}),$$

gdzie $U_{J,K}$ to ogólnie współczynnik braku podobieństwa taksonomicznego (najczęściej odległości) pomiędzy grupami (skupiskami, *clusters*) J i K . Modyfikując parametry tej formuły, opisać możemy wszystkie odmiany SAHN. Dla metody elastycznego grupowania przyjmujemy: $\alpha_J + \alpha_K + \beta = 1$, $\alpha_J = \alpha_K$, $\beta < 1$, $\gamma = 0$. Zmieniając wartość β w przedziale $\langle -1, 1 \rangle$, uzyskujemy różne własności techniki: dla $\beta = -1$ fenogram będzie przypominał obliczony techniką najbliższego sąsiedztwa, dla $\beta = 1$ fenogram będzie typowy dla techniki najdalszego sąsiedztwa, natomiast dobierając wartości pośrednie, można uzyskać takie drzewo, na którym odległości będą najbardziej zgodnie z wyjściowymi. Autorzy zaproponowali wartość $\beta = -0,25$ jako zwykle najodpowiedniejszą. Metoda ta może dać dobre wyniki, choć zachodzi niebezpieczeństwo, że badacz tak długo będzie zmieniał wartość β , aż uzyska taki fenogram, jaki pragnął otrzymać.

Sokal i Michener (1958) wprowadzili technikę **średniej** (*average linkage clustering*), przy czym średnia może być **ważona** (*weighted*) lub nie (*unweighted*), **arytmetyczna** (*arithmetic*) bądź być **środkiem ciężkości** (*centroid*). W technice średniej nieważonej waga każdego OTU jest taka sama, w ważonej – wagi są różne. Jeżeli mamy fenogram (((a,b),c),d), to dołączając OTU d, każdy z OTU a, b, c ma wagę taką samą, równą $1/3$. W technice ważonej natomiast, dla grupy (OTU) J : $w_J = 1/2^{c_j}$, gdzie c_j to liczba kroków analizy po dołączeniu grupy (OTU) J . W naszym przykładzie więc, gdy dołączamy d, $w_a = w_b = 1/2^2 = 1/4$, natomiast $w_c = 1/2^1 = 1/2$. Techniki nieważone zwykle dają lepsze odwzorowanie oryginalnych odległości na fenogramie, jednak w pewnych przypadkach techniki ważne są odpowiedniejsze, gdy uważamy za pożądane wyższe ważenie niektórych grup. Jest tak choćby wtedy, gdy różne grupy reprezentowane są przez różną liczbę OTU, a nie chcemy, aby o wyniku klasyfikacji decydowały głównie te grupy, w których OTU jest więcej.

Spośród technik opartych na średniej wyróżniamy więc **technikę nieważonej średniej arytmetycznej UPGMA** (*unweighted pair-group method using arithmetic averages*), **ważonej średniej arytmetycznej WPGMA** (*weighted pair-group method using arithmetic averages*), **nieważonego środka ciężkości UPGMC** (*unweighted pair-group centroid method*) i **ważonego środka ciężkości WPGMC** (*weighted pair-group centroid method*). W technice nieważonej średniej arytmetycznej UPGMA odległość $d_{KU} = (T_I d_{KI} + T_J d_{KJ}) / (T_I + T_J)$, gdzie T_I i T_J to, odpowiednio, liczba OTU w klasach I i J , natomiast w metodzie ważonej średniej arytmetycznej WPGMA $d_{KU} = (d_{KI} + d_{KJ}) / 2$. W technice nieważonego środka ciężkości (UPGMC, *centroid technique*) na każdym etapie grupowania oblicza się współrzędne środka ciężkości nowo utworzonej grupy, jako średnią arytmetyczną obiektów wchodzących w skład tej grupy – interpretacja geometryczna jest tu oczywista, natomiast prostej formuły obliczeniowej brak. W technice ważonego środka ciężkości, czyli mediany (WPGMC, *median method*), odległość to mediana, czyli wartość środkowa z trzech: odległości środka pierwszej podgrupy od środka nowej podgrupy, odległości środka drugiej podgrupy od środka nowej podgrupy i odległości środka pierwszej podgrupy od środka drugiej podgrupy.

UPGMA (Ryc. 3.9c) jest najczęściej stosowaną i w sumie najdogodniejszą techniką, nieścieśniającą ani nierozszerzającą przestrzeni, choć obliczone tą metodą fenogramy ogólnie przypominają uzyskane techniką najdalszego sąsiedztwa. WPGMA wazy dołączaną grupę (OTU) równo z wszystkimi razem wcześniej dołączonymi, a więc fenogram będzie ogólnie podobny jak w UPGMA, sugerując podobną strukturę taksonomiczną klasyfikowanego zbioru, jednak główne (wyższe w hierarchii) grupy będą dalej od siebie, a im później dołączany OTU (grupa), tym bardziej odrębny, kończąc dłuższą gałąź fenogramu. Jeżeli więc nie ma jakichś specjalnych powodów dla zwiększenia tej odrębności, należy stosować technikę UPGMA. UPGMC jest koncepcyjnie atrakcyjna – z uwagi na wspomnianą prostą interpretację geometryczną – jednak obliczeniowo bardziej skomplikowana, a przede wszystkim znacznie zniekształca pierwotne odległości, jest też podatna na odwrócenia (patrz niżej), więc raczej należy jej unikać, podobne wady ma też WPGMC.

Analiza skupisk – bez względu na rodzaj techniki – działa w ten sposób, że znajdzie skupiska nawet dla losowych danych czy obiektów rozmieszczonych równomiernie. Zresztą źródła błędów i zmienność nie są formalnie uwzględniane, toteż technika ta jest bardzo wrażliwa na obserwacje nietypowe. Niewłaściwa klasyfikacja obiektu na którymś etapie analizy nie zostanie już poprawiona i błąd pozostanie, brak jakiegokolwiek mechanizmu sprawdzającego czy korygującego. W dodatku, jak już wspominaliśmy, analiza skupisk – w przeciwieństwie do analizy głównych składowych czy analizy głównych współrzędnych – źle odwzorowuje różnice pomiędzy większymi, odleglejszymi od siebie grupami, gdy dla bliskich bywa lepsza niż tamte techniki. Zawsze więc warto porównać wyniki analizy skupisk i analizy głównych składowych (lub głównych współrzędnych). Jeżeli skupiska występujące na fenogramie dadzą się wyróżnić również w pierwszych 2–3 wymiarach analizy głównych składowych, to mamy większą pewność, że te skupiska realnie istnieją. Oczywiście może być tak, że skupiska te dadzą się odnaleźć dopiero w następnych wymiarach, jednak dla rzeczywistych, nie specjalnie spreparowanych danych taka sytuacja nie wydaje się zbyt prawdopodobna. Nie mniej ważne jest zastanowienie się, czy topologia otrzymanego drzewa jest merytorycznie sensowna – czy ma jakieś biologiczne znaczenie. Raz jeszcze warto przeanalizować, czy zastosowana odległość została wybrana właściwie, a także spróbować użyć innych odległości. Podobnie sensowne jest powtórzenie analizy dla, powiedzmy, UPGMA oraz metody najbliższego i najdalszego sąsiedztwa (Ryc. 3.9a–c). Grupy, które będą wspólne dla wszystkich technik (*ball clusters*), czyli o największej stałości, to grupy najbardziej wiarygodne – wewnątrz takiej grupy obiekt najbardziej odrębny i tak jest bardziej podobny do każdego z obiektów w obrębie grupy niż do jakiegokolwiek spoza niej.

Jak wiemy, fenogram jest drzewem ultrametrycznym. Tak więc odległości, dla których przeprowadza się analizę skupisk, powinny być ultrametryczne. Gdyby takie były, każda z technik SAHN dałaby identyczne wyniki (Swofford i Olsen 1990), a fenogram byłby drzewem filogenetycznym. W praktyce nigdy tak nie jest, a odległości obliczone dla danych empirycznych, czyli fenetyczne, mniej czy bardziej odbiegają od ultrametryczności. Oczywiście im bardziej odbiegają, tym mniejsza jest dobroć fenogramu, skonstruowanego na ich podstawie. Aby móc oceniać tę dobroć, czyli prawidłowość odwzorowania odległości oryginalnych na fenogramie, Sokal i Rohlf (1962) zaproponowali metodę **korelacji kofenetycznych** r_c lub r_{coph} (*cophenetic correlations*). Odległości pomiędzy poszczególnymi OTU, występujące na fenogramie, to **odległości ko-**

fenetyczne (*cophenetic distances*), C_{jk} , pomiędzy OTU j i k na drzewie. Współczynnik korelacji odległości fenetycznych z kofenetycznymi osiąga tym wyższe wartości, im lepsze odwzorowanie oryginalnych odległości na fenogramie. Można więc mówić o kryterium wyboru fenogramu o najwyższej wartości r_c (*cophenetic parsimony*).

Farris (1969) stwierdził, że wprawdzie UPGMA i pokrewne metody zwykle maksymalizują wartość r_c , jednak dla niektórych zestawów danych, choćby przy różnych liczbach OTU w kolejnych gałęziach fenogramu, współczynnik będzie najwyższy wówczas, gdy w konstrukcji fenogramu dozwolone są **odwrócenia**, czyli pomimo że $S_{A,B,C} > S_{A,B}$, to może być: $S_{A,B} > S_{A,C}$ lub $S_{B,C}$. Takie odwrócenia zwykle nie są dopuszczane w biologicznej taksonomii – nie bardzo dałoby się je sensownie interpretować – więc Farris uznał, że stosowanie r_c jako kryterium optymalizacji jest ryzykowne. Z drugiej strony, odwrócenia (*inversions*) występują jedynie wówczas, gdy dane są słabo hierarchiczne, bez wyraźniej zaznaczonych grup (czyli wówczas, gdy i tak analiza skupisk nie ma większego sensu), a w dodatku zwykle zdarzają się jedynie przy analizie techniką środka ciężkości lub mediany. Wydaje się więc, że współczynnik ten można z powodzeniem stosować dla niemal każdego empirycznego danych, pamiętając, że dla niektórych z nich nie osiągnie on tak wysokich wartości, jak dla większości macierzy. Tak czy inaczej duże różnice wartości r_c pozostają znaczące (Sneath i Sokal 1973).

Testu Mantela oczywiście użyć nie możemy (Rohlf i Fisher 1968), bowiem macierze nie są niezależne (kofenetyczna powstała na podstawie wyjściowych, fenetycznych odległości), tak jak nie można użyć tablic statystycznych dla sprawdzenia istotności korelacji, bowiem kolejne wartości w macierzy nie są niezależne. Pozostaje interpretacja subiektywna, oparta na symulacjach (Rohlf 1994): dla $r \geq 0,9$ – odwzorowanie bardzo dobre; dla $0,8 \leq r < 0,9$ – odwzorowanie dobre; dla $0,7 \leq r < 0,8$ – odwzorowanie słabe; dla $r < 0,7$ – odwzorowanie bardzo słabe. Rohlf i Fisher (1968) badali zachowanie r_c dla zestawu OTU, losowo pobieranych z populacji o rozkładzie normalnym. Rosnącej liczbie OTU włączanych do analizy skupisk odpowiadała malejąca wartość r_c , współczynnik malał też, choć wolniej, w miarę dodawania liczby cech opisujących kolejne OTU. r_c zbliżał się do wartości 0,3 dla macierzy korelacji, a 0,6 dla macierzy odległości. Zachowanie r_c nie odbiegało więc od spodziewanego i potwierdzało jego użyteczność. Własności odległości kofenetycznych omawiają też Rohlf i Sokal (1981). Warto sprawdzić, jakie wartości kofenetyczne oddzielają skupiska – czy przekraczają kilkakrotnie błąd standardowy dla odległości fenetycznych. Istotna jest też trwałość skupisk, w sensie ich odporności na dołączanie nowych OTU – czy dodawanie kolejnych obiektów nie zmienia grupowania, gdyż nowe obiekty dołączane są do któregoś z istniejących skupisk, czy też zmienia klasyfikację, tworząc nowe skupiska lub – co gorsza – zmieniając przynależność OTU wcześniej sklasyfikowanych. Dobrą strategią jest losowy podział zestawu obiektów na dwie połowy, przeprowadzenie analizy skupisk dla każdej z osobna i następnie porównanie wyników. Istnieje też cała gama technik opartych na próbkowaniu numerycznym, które omówimy przy drzewach obliczanych technikami filogenetycznymi (Rozdział 4.11), a które można zastosować i dla fenogramów.

3.8. Wielowymiarowa analiza wariancji

Wielowymiarowa analiza wariancji (*multivariate analysis of variance* – MANOVA) jest wielowymiarowym uogólnieniem jednowymiarowej analizy wariancji (ANOVA). Analizowaniem wariancji – czyli zmienności – można by właściwie nazywać całą statystykę, jednak przyjęto analizę wariancji nazywać technikę podziału, czyli partycjonowania ogólnej sumy kwadratów odchyłeń od średniej, umożliwiającą ocenę udziału różnych źródeł zmienności w zmienności ogólnej. ANOVA, wprowadzona przez Fishera, zaproponowana została dla danych doświadczalnych, stąd najczęściej stosowana terminologia. Jak pamiętamy, dla danych doświadczalnych właściwe jest podejście stochastyczne i takie podejście jest wymagane dla każdego danych analizowanych tą techniką. Badany zbiór musi też być sklasyfikowany – ANOVA ani MANOVA nie są pomocne w klasyfikacji, a jedynie mogą służyć do oceny odrębności wyróżnionych wcześniej grup.

Jeżeli, powiedzmy, 50 myszy losowo podzielono na 5 grup i każdą z grup karmiono innym pokarmem, to oczywiście możemy być pewni, że każda z 50 myszy miała inną wagę pod koniec eksperymentu. **Ogólną zmienność** tej wagi możemy jednak podzielić na **zmienność pomiędzy grupami**, będącą następstwem zróżnicowanej diety, czyli **zabiegu** (*treatment effect*) oraz **zmienność wewnątrz grup**, będącą następstwem **błędów** pomiaru i/bądź **przypadku** (*residual, error*). Oczywiście w takiej sytuacji, opisywanej przez model ANOVA **klasyfikacji pojedynczej**, zakłada się, że całkowita zmienność to suma efektu zabiegu i błędu. W **klasyfikacji podwójnej** obok zróżnicowanej diety uwzględniamy, powiedzmy, płeć myszy. Wówczas mamy model: zmienność ogólna = efekt zabiegu + efekt płci (drugiego zabiegu, bloku) + błąd, albo też – jeżeli, przykładowo, nadmiar tłuszczu w pokarmie bardziej przyspiesza wzrost wagi samic – pojawia się jeszcze jeden składnik sumy, mianowicie **interakcja** płci i diety. Może być oczywiście również klasyfikacja potrójna, z interakcją/interakcjami lub bez itd. W taksonomii dane nie pochodzą z eksperymentu, można jednak z powodzeniem zastosować ANOVA z klasyfikacją pojedynczą przykładowo w analizie zmienności wewnątrz- i między populacyjnej czy między populacjami należącymi do tego samego i różnych gatunków. Klasyfikacja podwójna może uwzględnić, obok przynależności gatunkowej, np. płeć albo populacje europejskie porównywane z azjatyckimi. Analiza wariancji pozwala też testować istotność różnic średnich między grupami.

Analogicznie MANOVA umożliwia, w najprostszym zastosowaniu, weryfikację hipotezy o równości wektorów średnich rozkładów w poszczególnych grupach. Ponownie jednak warto przypomnieć, że wielowymiarowa analiza wariancji nie znajduje zastosowania w analizie struktury danych, wskazującej na możliwą klasyfikację, a jedynie pomaga ocenić odrębność grup już wyróżnionych. Nie jest to więc technika specjalnie użyteczna w taksonomii, a zarazem opiera się na mocnych założeniach: (1) próby są pobierane losowo z każdej z badanych populacji (albo populacje są wybierane losowo w obrębie każdego gatunku) i próby z różnych populacji (gatunków) są od siebie niezależne; (2) wszystkie populacje mają tę samą, wspólną macierz kowariancji Σ ; (3) każda z populacji jest wielowymiarowo normalna lub dla każdej populacji pobrano losowo wiele (jak pamiętamy, co najmniej 30) prób. W praktyce taksonomicznej zazwyczaj założenia te nie są spełnione, toteż MANOVA stosowana bywa często w sposób nieuprawniony. Technikę omawia dokładnie szereg podręczników MSA cytowa-

nych wcześniej, tutaj ograniczymy się do paru uwag i najprostszego modelu MANOVA: klasyfikacji pojedynczej.

MANOVA przypomina ANOVA, lecz we wzorach kwadraty skalarów zastępują ich odpowiedniki wektorowe. Mamy $\ell = 1, 2, \dots, g$ populacji (gatunków, ale też np. zabiegów, najogólniej biorąc: wyróżnianych w jakiś sposób klas), a z każdej z nich pobrano $j = 1, 2, \dots, n_\ell$ osobników (populacji, powtórzeń, najogólniej: prób). Wówczas dla pojedynczej klasyfikacji:

$$x_{\ell j} = \bar{x} + (\bar{x}_\ell - \bar{x}) + (x_{\ell j} - \bar{x}_\ell), \text{ czyli } X_{\ell j} = \mu + \tau_\ell + e_{\ell j},$$

co oznacza, że obserwacja = średnia dla wszystkich obserwacji + estymowany efekt zabiegu + błąd. Tak więc, dla modelu wielowymiarowego, całkowita (skorygowana) suma kwadratów i iloczynów wektorowych równa jest sumie kwadratów i iloczynów wektorowych zabiegu (międzypopulacyjnej zmienności), danej macierzą \mathbf{B} , oraz sumie kwadratów i iloczynów wektorowych błędu (wewnątrzpopulacyjnej zmienności), danej macierzą \mathbf{W} . Macierze \mathbf{B} i \mathbf{W} definiujemy następująco:

$$\mathbf{B} = \sum_{\ell=1}^g n_\ell (\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})', \quad \mathbf{W} = \sum_{\ell=1}^g \sum_{j=1}^{n_\ell} (x_{\ell j} - \bar{x}_\ell)(x_{\ell j} - \bar{x}_\ell)', \quad \mathbf{B} + \mathbf{W} = \sum_{\ell=1}^g \sum_{j=1}^{n_\ell} (x_{\ell j} - \bar{x})(x_{\ell j} - \bar{x})'$$

Dla sprawdzenia istotności różnic między populacjami formułujemy hipotezę H_0 : $\tau_1 = \tau_2 = \dots = \tau_g = 0$, czyli że brak różnic (efektów zabiegu), a więc wektory średnich dla populacji są równe. Alternatywna hipoteza H_1 zakłada, że choć jeden $\tau_k \neq 0$, a więc choć jeden efekt zabiegu jest różny od zera albo choć jedna z populacji istotnie różni się od pozostałych. Warto o tym pamiętać – odrzucenie hipotezy H_0 nie oznacza odrębności każdej z klas, a które są odrębne, ocenić można dopiero odpowiednimi testami *post hoc*. Hipotezę H_0 możemy odrzucić wówczas, gdy wartość proporcji uogólnionych wariancji, znanej jako **lambda Wilksa** (1932) (*Wilks' lambda*), jest zbyt mała:

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} = \frac{\sum_{\ell=1}^g \sum_{j=1}^{n_\ell} (x_{\ell j} - \bar{x}_\ell)(x_{\ell j} - \bar{x}_\ell)'}{\sum_{\ell=1}^g \sum_{j=1}^{n_\ell} (x_{\ell j} - \bar{x})(x_{\ell j} - \bar{x})'} = \prod_{i=1}^s \left(\frac{1}{1 + \hat{\lambda}_i} \right),$$

gdzie $\lambda_1, \lambda_2, \dots, \lambda_s$ to wartości własne macierzy $\mathbf{W}^{-1}\mathbf{B}$, a $s = \min(p, g - 1)$, gdzie p to liczba zmiennych wchodzących w skład badanych rozkładów wielowymiarowych. Oczywiście wartość lambda Wilksa jest tym mniejsza, im mniejszy udział zmienności wewnątrzpopulacyjnej w całkowitej zmienności, czyli im mniej zmienności wewnątrzpopulacyjnej, a więcej międzypopulacyjnej (międzyklasowej). Podobna jest zasada trzech innych testów: (1) **śląd Lawleya-Hotellinga** (*Lawley-Hotelling Trace*) = $\text{tr}[\mathbf{B}\mathbf{W}^{-1}]$, (2) **śląd Pillai** (*Pillai's Trace*) = $\text{tr}[\mathbf{B}(\mathbf{B} + \mathbf{W})^{-1}]$, (3) **największa wartość własna Roya** (*Roy's largest root*) = największa wartość własna $\mathbf{W}(\mathbf{B} + \mathbf{W})^{-1}$.

Wszystkie cztery testy zachowują się niemal identycznie dla skrajnie wielkich prób (Johnson i Wichern 1988). Dla prób mniejszych największa wartość własna Roya zachowuje się różnie, jej moc jest dostateczna właściwie jedynie wówczas, gdy wyraźna różnica międzyklasowa dotyczy jednej cechy, wyróżniającej jedną z klas (wówczas jedna wartość własna jest silnie różna od 0). Pozostałe testy wydają się mieć podobną moc dla niewielkich prób; prawdopodobnie ślad Pillai jest najodporniejszy na odstępstwa od normalności rozkładów, a lambda Wilksa stosowana jest najczęściej. Gdy H_0 jest prawdziwa, a $\sum n_i = n$ ma dużą wartość, to rozkład:

$$-\left(n-1-\frac{(p+g)}{2}\right)\ln \Lambda^* = -\left(n-1-\frac{(p+g)}{2}\right)\ln\left(\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}\right)$$

zbliża się do rozkładu χ^2 o $p(g-1)$ stopniach swobody. Normalność rozkładów błędów powinna zostać sprawdzona, podobnie jak występowanie obserwacji nietypowych, wskazanych przez wyjątkowo wysokie wartości w macierzy \mathbf{W} . Omówienie bardziej złożonych modeli MANOVA znaleźć można w literaturze.

3.9. Analiza dyskryminacyjna

Analiza dyskryminacyjna (*discriminant analysis*) to rozległy dział statystyki, zajmujący się klasyfikowaniem (zaliczaniem) obiektów, będących wielowymiarowymi obserwacjami, do określonych wcześniej klas (grup). Z tym oczywiście wiązać się musi określanie charakterystyk, różniących te grupy. Analiza dyskryminacyjna znajduje szereg zastosowań, począwszy od oznaczania organizmów żywych czy diagnostyki medycznej, a kończąc na komputerowych systemach rozpoznawania obrazu. Zależnie od techniki, przyjmuje się model stochastyczny bądź opisowy, najczęściej wymagana jest wielowymiarowa normalność rozkładu i spełnianie określonych założeń dotyczących macierzy kowariancji. Zazwyczaj dysponuje się zbiorem sklasyfikowanych obserwacji, który pełni funkcję zbioru „uczącego”, umożliwiającego klasyfikację pozostałych obserwacji. Analiza dyskryminacyjna to szeroki zestaw technik, opartych na **statystycznych funkcjach decyzyjnych**, sieciach neuronowych bądź **funkcjach dyskryminacyjnych** (*discriminant functions*).

W taksonomii biologicznej najczęściej wykorzystuje się liniową funkcję dyskryminacyjną, wprowadzoną przez Fishera (1936, 1938). Zanim ją przedstawimy, zaznaczyć warto, że technika ta – jak wszystkie dyskryminacyjne – bywa w taksonomii biologicznej nadużywana, a co najmniej jej wyniki są niezupełnie prawidłowo interpretowane. Klasy muszą być dane *a priori*, gdy tak nie jest, techniki stosować nie można. Analizy dyskryminacyjnej nie da się więc użyć dla stwierdzenia, do ilu grup można sklasyfikować badany zbiór obiektów. Musimy wiedzieć, do jakich klas należy badany zbiór, musimy też mieć – w praktyce im liczniejsze, tym lepiej – podzbiory już sklasyfikowane do odpowiednich klas. Wówczas możemy zarówno sklasyfikować – z mniejszą lub większą pewnością – pozostałe obiekty, precyzując granice między klasami, jak też badać, które z cech i w jakim stopniu określają przynależność do określonej klasy. Warto przy tym pamiętać, że nasze wysiłki mające na celu skonstruowanie jak najlep-

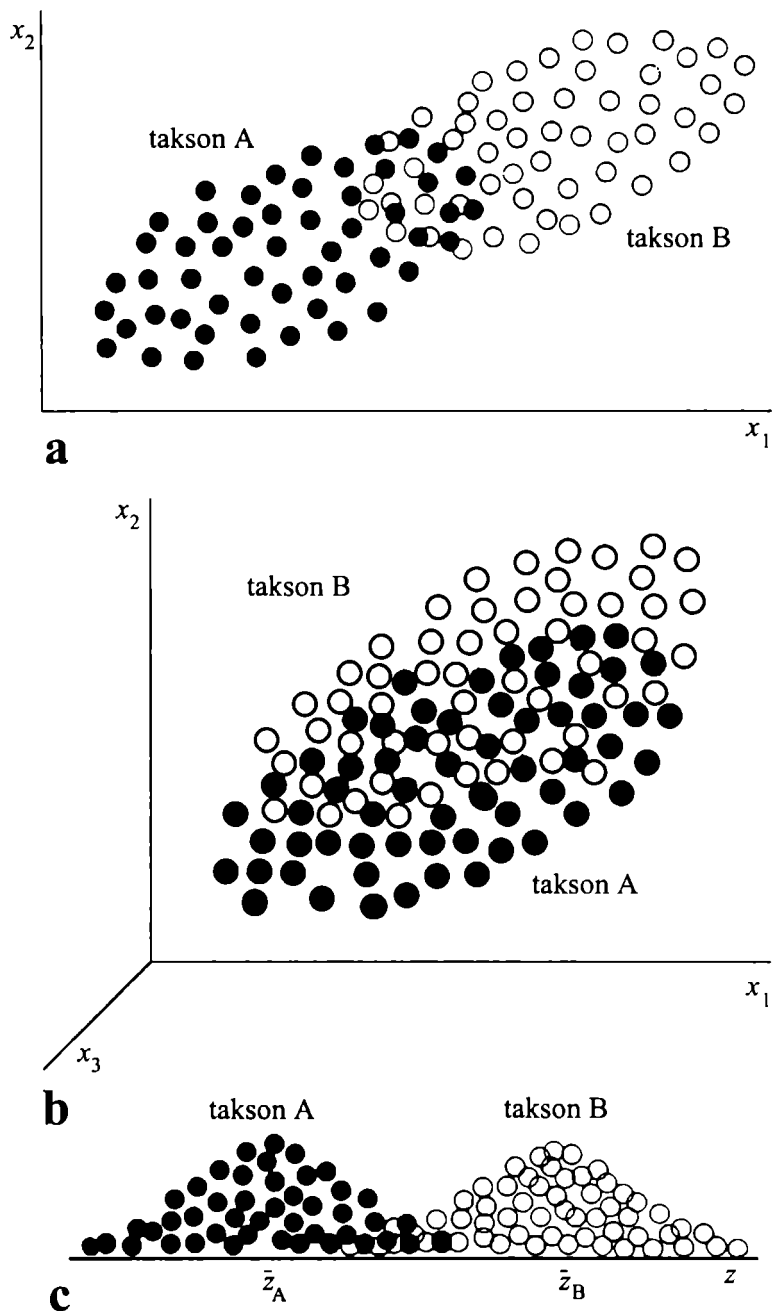
szej funkcji dyskryminacyjnej nie mają większego sensu, jeżeli wektory średnich dla grup nie różnią się statystycznie istotnie, co sprawdzić możemy za pomocą MANOVA, jeżeli rozkłady są wielowymiarowo normalne.

Oprócz zaliczania obiektów do klas – np. oznaczania osobników do określonych gatunków – analiza dyskryminacyjna będzie przydatna wówczas, gdy chcemy sprawdzić znalezioną w literaturze opinię, że, powiedzmy, dwa gatunki chrząszcza różnią się szerokością przedplecza i długością zuwaczek. Wówczas znajdujemy funkcję dyskryminacyjną na podstawie szeregu cech, w tym dwóch badanych, a następnie sprawdzamy udział tych badanych cech w funkcji dyskryminacyjnej i dobroć dyskryminacji. Analiza dyskryminacyjna pokazać też może, że dany zestaw cech różnicuje lepiej gatunki A i B, natomiast C odróżnia się słabo od A. Nie można natomiast użyć tej techniki do wnioskowania o odrębności grup określonych *a priori* – grupy te dane są jako założenie, ponadto technika maksymalizuje wariancję międzygrupową, minimalizując łączną wariancję wewnątrzgrupową. Maksymalnie więc zbliża do siebie obiekty wewnątrz grup, a oddala obiekty należące do różnych grup. Łatwo więc o stworzenie obrazu odrębności, w rzeczywistości nieistniejącej. Z drugiej strony, słabe odróżnianie obiektów przy użyciu funkcji dyskryminacyjnej może dowodzić jednorodności grupy, czyli wskazywać na bezpodstawność wyodrębnienia badanych klas, przynajmniej na podstawie badanego zestawu cech. W takim przypadku warto jednak starannie sprawdzić, czy zostały spełnione warunki dotyczące rozkładów i macierzy kowariancji, o czym niżej.

Bywa, że dwa gatunki różnią się wartościami średnimi jakiejś cechy x_1 , ale zakresy zmienności zachodzą na siebie. Następna zmienna x_2 także różnicuje, lecz tylko częściowo (Ryc. 3.10a). Podobna sytuacja ma miejsce dla trzech zmiennych (x_1, x_2, x_3), z których żadna nie umożliwi identyfikacji z jakąkolwiek dającą się określić pewnością (Ryc. 3.10b). Aby sytuację przybliżyć, załóżmy, że idzie o dwa gatunki, z których jeden jest żywicielem pośrednim groźnego pasożyta, a przynależność osobników przedstawionych na wykresach znamy na podstawie cech molekularnych, powiedzmy, sekwencji DNA dla paru genów. Oczywiście wykonanie kilku pomiarów morfometrycznych – ustalenie wartości kilku cech – jest prostsze i tańsze niż sekwencjonowanie, więc możliwość oznaczenia gatunku na podstawie pomiarów miałaby nieocenione znaczenie. Dla takich właśnie przypadków Fisher (1936) stworzył **liniową funkcję dyskryminacyjną** (*discriminant function analysis*). Polega ona na stworzeniu funkcji, której wartości są liniową kombinacją wszystkich zmiennych – w ten sposób dla każdej obserwacji m zmiennych zastępuje jedna zmienna z , określona następująco:

$$z = a_{1j}x_{i1} + a_{2j}x_{i2} + \dots + a_{mj}x_{im},$$

gdzie: $a_{1j}, a_{2j}, \dots, a_{mj}$ to współczynniki liniowej funkcji dyskryminacyjnej dla klasy C_j . Transformację wielowymiarowych obserwacji w jednowymiarową funkcję dyskryminacyjną z przeprowadza się, dobierając współczynniki $a_{1j}, a_{2j}, \dots, a_{mj}$ tak, aby funkcja z osiągała możliwie najwyższe wartości dla obserwacji należących do klasy C_j , a jak najniższe – dla nienależących do tej klasy. Dzięki temu funkcja z umożliwia różnicowanie obiektów należących do tej klasy od pozostałych, lepiej niż jakakolwiek z wyjściowych zmiennych. W ilustrowanym przykładzie mamy dwie klasy: j i k , ale oczywiście można odróżniać elementy klasy j od elementów nienależących do j , a więc technikę tę stosować można dla więcej niż dwóch klas. Wówczas musimy wyznaczyć,



Ryc. 3.10. W przestrzeni dwóch zmiennych (x_1 i x_2) zakresy zmienności dwóch taksonów A i B zachodzą na siebie (a), a po dodaniu trzeciej zmiennej (x_3) zachodzenie zaznacza się nawet jeszcze bardziej (b). Obliczona funkcja dyskryminacyjna z (c) maksymalizuje separację obiektów należących do taksonów A i B

dla k klas, $k - 1$ funkcji dyskryminacyjnych. Zwykle stosuje się analizę sekwencyjną, czyli kolejno próbuje wyznaczyć funkcje odróżniające kolejne klasy – wybiera się tę, która różnicuje najlepiej którąś z klas od pozostałych, po czym powtarza próby w obrębie $k - 1$ klas, itd. Funkcja z dobierana jest tak, aby zmaksymalizować **separację** wyróżnianych klas (Ryc. 3.10c):

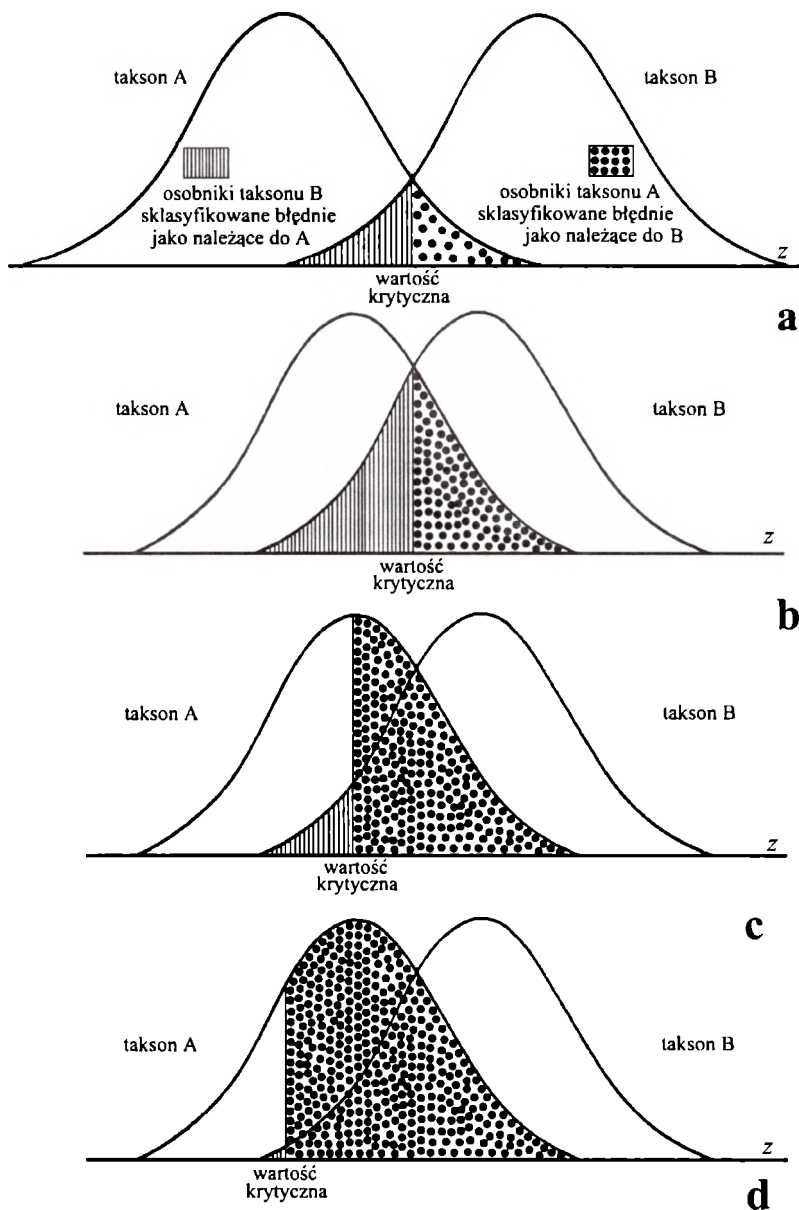
$$\text{Separacja} = \frac{|\bar{z}_j - \bar{z}_k|}{s_z}, \text{ gdzie: } s_z^2 = \frac{\sum_{i=1}^{n_j} (z_{ji} - \bar{z}_j)^2 + \sum_{i=1}^{n_k} (z_{ki} - \bar{z}_k)^2}{n_j + n_k - 2};$$

s_z^2 to łączna wariancja klas C_j i C_k , o liczebnościach n_j , n_k . **Wartość krytyczna** (*cuttoff score*) funkcji z oddziela wartości funkcji dla obiektów zaklasyfikowanych do jednej z klas od obiektów zaklasyfikowanych do drugiej klasy. Zwykle jest to średnia wartości średnich dla obu klas $[(\bar{z}_j + \bar{z}_k)/2]$, co wydaje się oczywiste dla rozkładów normalnych o identycznych macierzach kowariancji i gdy frekwencje w próbie obiektów obu klas są identyczne, ale tak być nie musi. Dla wyjściowych m zmiennych oblicza się dla każdej z grup macierz $m \times m$ wariancji-kowariancji. Następnie oblicza się macierz łącznej wariancji-kowariancji wewnątrzklasowej \mathbf{W} , jako średnią macierzy wyjściowych, po czym przelicza na macierz odwróconą \mathbf{W}^{-1} , co nie jest możliwe, gdy choć jedna ze zmiennych nie wykazuje zmienności. Następnie obliczamy wektor zmienności międzygrupowej:

$$\delta_{JK} = [(\bar{x}_{1J} - \bar{x}_{1K}), (\bar{x}_{2J} - \bar{x}_{2K}), \dots, (\bar{x}_{mJ} - \bar{x}_{mK})], \text{ a następnie: } \mathbf{z}_{JK} = \mathbf{W}^{-1} \delta_{JK},$$

czyli wektor funkcji dyskryminacyjnej. Funkcja dyskryminacyjna transformuje oryginalne obserwacje tak, że średnia wewnątrzgrupowa wariancja jest równa dla osi każdej z cech i osie są do siebie skośne tak, aby rozrzut obiektów każdej z klas możliwie najbardziej zbliżył się do hipersfery (Sneath i Sokal 1973). Odległość pomiędzy centroidami j i k mierzona w jednostkach funkcji dyskryminacyjnej jest równa pierwiastkowi kwadratowemu odległości Mahalanobisa D^2 między centroidami dla danych wyjściowych, a wartość absolutna różnicy średnich wartości funkcji dla obu klas równa jest D^2 .

Populacje reprezentujące klasy niekoniecznie muszą być wielowymiarowo normalne, ale muszą mieć identyczne macierze kowariancji, ponieważ posługujemy się łączną kowariancją klas. Oczywiście oznacza to, że wariancja każdej z cech diagnostycznych musi być identyczna w każdej z klas (gdy dla różnych cech może być oczywiście różna), tak samo identyczne we wszystkich klasach muszą być współczynniki korelacji dla danej pary cech. Technika działa też lepiej, gdy populacje są wielowymiarowo normalne, dobrze więc współczynniki liniowej funkcji dyskryminacyjnej dobierać dla danych odpowiednio transformowanych. Ponadto normalność rozkładów pozwala na ocenę prawdopodobieństwa błędnej klasyfikacji, zgodnie z kryterium położenia obiektu w stosunku do centroidu, mierzonej odchyleniem standardowym. Grupy powinny być



Ryc. 3.11. Gdy zakresy zmienności zmiennych oryginalnych zachodzą na siebie – a jedynie wówczas analiza dyskryminacyjna ma sens – to również zakresy zmienności wartości funkcji dyskryminacyjnej będą na siebie zachodzić, nawet dla najlepiej dobranej funkcji, choć zachodzić będą w mniejszym stopniu niż dla danych oryginalnych. Zachodzenie będzie mniejsze (a) lub większe (b), pociągając za sobą nieuniknione błędne klasyfikacje niektórych obiektów. Zmieniając wartość krytyczną (b, c, d), można zmniejszyć liczbę błędnie sklasyfikowanych osobników taksonu A lub B, lecz zawsze oznacza to zwiększenie liczby błędnie sklasyfikowanych osobników drugiego z taksonów

jednorodne, osobniki każdej z klas powinny tworzyć skupisko o podobnej liczebności, kształtu i orientacji w hiperprzestrzeni, a dla prawidłowego działania funkcji dyskryminacyjnej liczebność próby powinna być jak największa. Gdy kowariancje są różne, odpowiedniejsza jest **kwadratowa funkcja dyskryminacyjna**, ale funkcja taka jest jeszcze bardziej wrażliwa na odchylenia od normalności rozkładów. Liniowa funkcja dyskryminacyjna dla danych pochodzących z rzeczywistych pomiarów, gdy warunki dotyczące rozkładów i kowariancji zwykle nie są spełnione, działa lepiej lub gorzej. Może przydzielać osobniki do klas zupełnie źle, gdy korelacje dla określonych par cech są w jednej z klas wysokie, w innej niskie bądź w jednej dodatnie, a w drugiej ujemne. Oczywiście nawet gdy rozkłady są idealnie wielowymiarowo normalne i macierze kowariancji identyczne, udział błędnych oznaczeń: (źle sklasyfikowane obiekty klasy j + źle sklasyfikowane obiekty klasy k)/ $n_j + n_k$, może być wysoki. Błędnych oznaczeń nie da się uniknąć, jak długo zakresy zmienności cech diagnostycznych na siebie zachodzą (a gdyby nie zachodziły, to taka funkcja byłaby zbędna). Błędnych oznaczeń będzie tym więcej, im bardziej zachodzą na siebie zakresy zmienności funkcji diagnostycznej (Ryc. 3.11a, b). Przesuwając wartość krytyczną, możemy zmniejszać liczbę błędnych oznaczeń obiektów jednej klasy, zwiększając liczbę błędnych oznaczeń obiektów klasy drugiej (Ryc. 3.11c, d). Może to mieć znaczenie praktyczne: przykładowo mniej ryzykowne jest błędne uznanie niejadowitego węża za jadowitego niż odwrotnie, błędne rozpoznanie groźnej choroby niż jej przeoczenie.

Współczynniki liniowej kombinacji dla kolejnych cech to inaczej wagi – im wyższa wartość wagi tym wyższy udział cechy w tworzeniu funkcji dyskryminacyjnej. Jeśli wartości cech nie są standaryzowane, wagi nie muszą odzwierciedlać dokładnie istotności określonych cech dla odróżniania badanych obiektów. Odzwierciedlają je natomiast procentowe udziały cech w tworzeniu funkcji dyskryminacyjnej, dla cechy i : $100 z_i \delta_i / D^2$, gdzie z_i i δ_i to i -te elementy wektorów \mathbf{z}_{JK} i δ_{JK} . W ten sposób odróżnimy cechy najlepsze dla klasyfikowania obiektów do określonej klasy. Zazwyczaj szukać ich należy wśród zmiennych odznaczających się dużą różnicą średnich dla grup (mierzonych odchyleniami standardowymi) i słabo skorelowanych z innymi cechami. Możemy też badać dobroć klasyfikacji dla różnych zestawów cech, co wydaje się najważniejsze. W ten sposób niemal zawsze stwierdzimy, że tak samo – lub niemal tak samo – dobre odróżnianie obiektów z różnych klas możliwe jest przy użyciu znacznie mniejszej liczby cech niż wyjściowo mierzone, co pozwoli oszczędzić czas i wysiłek. Dobroć odróżniania – separacji – oceniamy na podstawie wspomnianej wcześniej proporcji: (źle sklasyfikowane obiekty klasy j + źle sklasyfikowane obiekty klasy k)/ $n_j + n_k$. Im niższa wartość tej proporcji, tym odróżnianie lepsze.

Teoretycznie najprościej jest ocenić wartość tej proporcji, czyli udział błędnych klasyfikacji, zwyczajnie licząc sklasyfikowane błędnie obiekty zbioru, dla którego obliczono funkcję dyskryminacyjną. Wartość taka będzie jednak obciążona błędem – zaniżona – bowiem funkcję obliczono dokładnie dla tych danych, może więc nie być odpowiednia dla wszystkich obiektów badanych klas. Często zaleca się więc, aby podzielić losowo zbiór na dwa podzbiory tej samej wielkości, dla pierwszego obliczyć funkcję dyskryminacyjną, a sprawdzić jej dobroć na drugim podzbiorku. Procedura taka ma sens jedynie dla rzeczywiście dużej liczebności badanego zbioru, a poza tym i tak pewne własności badanego zbioru – także odróżniania badanych klas – mogą zostać pominięte w pierwszym, „uczącym” podzbiorku. Godna polecenia jest procedura obliczeniowo intensywna: bierzemy $n_j - 1$ obiektów klasy j i n_k obiektów klasy k i ob-

liczymy funkcję dyskryminacyjną, po czym wykorzystujemy ją dla zaklasyfikowania pominiętego obiektu klasy j . Odnotowujemy, czy klasyfikacja była prawidłowa, czy nie, po czym powtarzamy procedurę, włączając ten, a pomijając inny obiekt tej klasy, i tak dla wszystkich obiektów klasy j , po czym tak samo dla wszystkich obiektów klasy k . Pamiętajmy też, że funkcja dyskryminacyjna zalicza obiekty do którejś z klas, lecz nie sprawdza, czy do którejkolwiek z nich należą. Jeżeli więc mamy funkcję odróżniającą osobniki gatunku A od osobników gatunku B, to osobnik gatunku C włączony do analizy niewątpliwie zostanie zaliczony do A lub B. Jest to zresztą typowe niebezpieczeństwo znane każdemu, kto kiedykolwiek używał jakiegoś klucza do oznaczania – zazwyczaj okaz, powiedzmy, niewielkiego afrykańskiego ślimaka bez trudu da się oznaczyć do gatunku przy użyciu klucza do polskich ślimaków.

Wszystkie sformułowane wcześniej założenia, dotyczące normalności i jednorodności rozkładów oraz ich kształtu w hiperprzestrzeni, jak też ograniczenia interpretacji wyników, obowiązują także dla wykorzystywanej niekiedy w taksonomii biologicznej dyskryminacji za pomocą **współrzędnych dyskryminacyjnych** (*discriminant coordinates*), czyli **zmiennych kanonicznych** (*canonical variates*). Tego drugiego terminu nie należy mylić ze zmiennymi kanonicznymi w analizie korelacji kanonicznej (*canonical correlation*), choć formalna zależność między tymi dwoma rodzajami zmiennych kanonicznych istnieje (Jajuga 1993). Współrzędne dyskryminacyjne pozwalają na jednoczesne przedstawienie K klas dla n obiektów, maksymalizując różnice pomiędzy klasami. Wiemy, ile jest klas i jakie są liczebności n_j obiektów należących do każdej z klas. Metoda polega na liniowej transformacji wektora zmiennych \mathbf{X} w zestaw nieskorelowanych (ortogonalnych) zmiennych Z_1, Z_2, \dots, Z_m , z których dwie pierwsze: Z_1 i Z_2 dane są zależnościami: $Z_1 = \mathbf{a}_1^T \mathbf{X}, Z_2 = \mathbf{a}_2^T \mathbf{X}$. Zmienne te dobiera się tak, aby osiągnąć maksimum funkcji określającej dobroć przedstawienia – maksymalizację różnic między klasami w przestrzeni zmiennych Z_1 i Z_2 . Należy więc zmaksymalizować: $\mathbf{a}^T \mathbf{B} \mathbf{a} / \mathbf{a}^T \mathbf{W} \mathbf{a}$, gdzie \mathbf{a} to wektor m -wymiarowy, \mathbf{B} – macierz rozrzutu międzyklasowego, \mathbf{W} – macierz rozrzutu wewnątrzklasowego (Jajuga 1993):

$$\mathbf{B} = \frac{1}{K-1} \sum_{j=1}^K n_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T, \quad \mathbf{W} = \frac{1}{n-K} \sum_{j=1}^K \sum_{i \in C_j} (\mathbf{x}_i - \bar{\mathbf{x}}_j)(\mathbf{x}_i - \bar{\mathbf{x}}_j)^T, \quad \bar{\mathbf{x}}_j = \frac{1}{n_j} \sum_{i \in C_j} \mathbf{x}_i.$$

Widać więc, że maksymalizowana jest liniowa kombinacja zmiennych wektora \mathbf{X} tak, aby proporcja zmienności międzyklasowej do wewnątrzklasowej była możliwie jak największa. Rozwiązaniem jest wektor \mathbf{a} , będący wektorem własnym i odpowiadający największej wartości własnej macierzy $\mathbf{W}^{-1} \mathbf{B}$. Jest to zarazem wektor \mathbf{a}_1 , w którego skład wchodzi współczynniki pierwszej współrzędnej dyskryminacyjnej. Drugą współrzędną dobieramy tak, aby była nieskorelowana z pierwszą i znów spełniała to kryterium, następnie kolejne zmienne – przypomina to analizę głównych składowych, podobnie jak tam kolejne współrzędne tłumaczą coraz mniejszą część międzyklasowego zróżnicowania. Tak więc kolejne wartości własne $\lambda_1, \lambda_2, \dots, \lambda_m$ macierzy $\mathbf{W}^{-1} \mathbf{B}$ będą coraz mniejsze. Współrzędnych dyskryminacyjnych używamy najczęściej dla przedstawienia różnic między klasami w przestrzeni dwuwymiarowej, zwykle pierwszych dwóch współrzędnych. Przedstawienie takie będzie oczywiście pełniejsze lub nie, zależnie od danych wyjściowych. Dobroć przedstawienia ocenić możemy

wartością proporcji: $(\lambda_1 + \lambda_2) / \sum_{i=1}^m \lambda_i$, należącą do przedziału $\langle 0, 1 \rangle$, osiągającą 1 dla odwzorowania doskonałego. Oczywiście i w tej technice, po obliczeniu zmiennych kanonicznych dla zbioru sklasyfikowanego, uczącego, wykorzystać je możemy do przeprowadzenia dyskryminacji obiektów, dla których przynależność do określonych klas musimy dopiero ustalić.

4. Analiza filogenetyczna

4.1. Zakres metod filogenetycznych, algorytm a model, kryteria optymalizacji

Terminu „systematyka (taksonomia) filogenetyczna” używa się zamiennie z terminem „kladystyka” i to zapewne jest powodem nierzadkiego uważania przymiotników filogenetyczny i kladystyczny za równoznaczne. Tak zapewne można tłumaczyć nierzadkie ograniczanie pojęcia analizy filogenetycznej do analizy kladystycznej, czyli redukcjonistycznej (*parsimony*). W takim rozumieniu do analizy filogenetycznej nie należałyby już nawet techniki oparte na maksymalizacji wiarygodności – choć przecież ich istotą jest uzyskiwanie możliwie największej zgodności z założonym modelem ewolucji – nie mówiąc już o wszelkich technikach opartych na odległościach, uznawanych wtedy za fenetyczne. Oczywiście można bronić takiego stanowiska, jednak nie wydaje się ono słuszne. Jak pamiętamy, analiza skupisk (*clustering*) jest typową metodą fenetyczną, choć nawet ona użyta do rekonstrukcji filogenezy da prawidłowy wynik dla danych ultrametrycznych (aczkolwiek wiemy, że ultrametryczne dane spotkamy jedynie wyjątkowo). Dla innych technik obliczających drzewa na podstawie odległości trudno byłoby wykazać, że rekonstruują pokrewieństwa na podstawie ogólnego podobieństwa, a nie posiadanej wiedzy o procesach ewolucyjnych. Warto też przypomnieć, że przy nieprawidłowo rozpoznanych homologiach nawet metoda kladystyczna da obraz fenetycznych podobieństw. W książce tej uznajemy więc za filogenetyczne wszystkie te techniki, które starają się uwzględnić to, co wiemy o procesach ewolucyjnych w obrębie grupy, której filogenezę badamy.

Omówiona przy analizie fenetycznej analiza skupisk to typowa metoda **algorytmiczna**, czyli taka, dla której dany jest **algorytm** konstruowania drzewa na podstawie określonych danych. Mamy więc wówczas określony zestaw kroków, których wykonanie prowadzi do skonstruowania „najlepszego” drzewa. Metody algorytmiczne spotyka się również wśród technik analizy filogenetycznej, najbardziej znana z nich to metoda dołączania sąsiada (*neighbor-joining*). Metody algorytmiczne są technikami szybkimi, nawet dla setek OTU obliczenia nie są długie i nie nastęrczają jakichkolwiek trudności współczesnym komputerom. Ta niewątpliwa zaleta nie równoważy jednak poważnej wady: wynikiem obliczeń jest pojedyncze drzewo, i to zarówno w przypadku, gdy jest ono najlepsze, a wszystkie inne zdecydowanie gorsze, jak też gdy istnieje szereg drzew jedynie nieco gorszych albo wręcz obliczone drzewo jest jednym z wielu,

nawet wieluset, rekonstruujących filogenezę jednakowo dobrze, czyli w takim samym stopniu zgodnych z danymi i tym, co wiemy o procesach ewolucyjnych w badanej grupie organizmów. Często też istnieją drzewa lepiej odtwarzające filogenezę niż to znalezione. Wadę tę jedynie do pewnego stopnia kompensować można (Swofford i inni 1996), ewaluując drzewa techniką *bootstrap* (patrz Rozdział 4.11). Dla technik algorytmicznych trudno też modyfikować założenia, trudniej dopasowywać je do określonego modelu ewolucji.

W analizie filogenetycznej najczęściej korzysta się więc z technik opartych na rozmaitych **kryteriach optymalizacji**, pozwalających na ocenę, które z dwóch drzew jest lepsze, czyli bardziej zgodne z danymi i założonym modelem ewolucji. Nie znaczy to, że techniki te nie wykorzystują algorytmów – algorytmy są podstawą wszelkich technik komputerowych i dobrze skonstruowany algorytm decyduje zarówno o szybkości obliczeń, jak i o ich prawidłowym wyniku. W technikach tych jednakże algorytmy są jedynie narzędziami; co więcej, czas życia takich algorytmów jest krótki, są stale doskonałone i właściwie pozostawić je możemy specjalistom, tworzącym coraz lepsze wersje programów do rekonstrukcji filogenezy. Zarazem w technikach opartych na kryterium optymalizacji kolejna wartość tego kryterium obliczana jest (za pomocą odpowiedniego algorytmu, najwydajniej kalkulującego wartość **obiektywnej funkcji ewaluującej** drzewo) dla kolejnych drzew, generowanych odrębnym algorytmem konstruującym drzewa. Techniki te są plastyczne – łatwo modyfikować kryterium, dopasowując je do naszej wiedzy o ewolucji – a zarazem ich wynikiem nie jest jedno drzewo, a zestaw drzew, wiemy więc, czy nasze „najlepsze” drzewo jest rzeczywiście najlepsze, a jeżeli jest, to o ile lepsze od innych. Te zalety okupione są jednak teoretycznie koniecznością zbadania wszystkich drzew, a wiemy już, że liczba możliwych drzew rośnie gwałtownie z liczbą taksonów terminalnych, już dla mniej więcej 20 przekraczając możliwości obliczeniowe komputerów. Wówczas pozostają metody przybliżone (omówimy je w Rozdziale 4.3), a i tak obliczenia stają się niezmiernie długotrwałe. Mimo to techniki te dominują i tak zapewne pozostanie.

Rekonstrukcja filogenezy to zestaw technik estymacji hipotetycznego przebiegu ewolucji w badanej grupie, rozumianego jako określona sekwencja kladogenez i procesów anagenetycznych, choć te ostatnie nie zawsze są estymowane. Obliczone estymaty powinny być jak najlepiej zgodne z danymi – rozmieszczeniem stanów cech u taksonów terminalnych – ale również z założonym modelem ewolucji, która mogła przecież przebiegać różnie. Wyjściowym założeniem jest możliwość opisu ewolucji za pomocą drzewa, a więc co najmniej rzadkość hybrydyzacji międzygatunkowej czy poziomego transferu materiału genetycznego. Zarazem „**sygnał filogenetyczny**” musi być dostatecznie silniejszy od „**szumów**”, czyli zmienność wewnątrzgatunkowa cech użytych do rekonstrukcji filogenezy musi być mniejsza niż różnice międzygatunkowe stanów tych cech, a homologie rozpoznane prawidłowo. Założenie nieodpowiedniego modelu ewolucji będzie źródłem błędu systematycznego, choć oczywiście obliczona rekonstrukcja nie musi być całkowicie bezużyteczna (Rozdział 4.11). Szereg technik zakłada określony model ewolucji, inne nie, ale nie znaczy to, aby były one wolne od założeń. Przykładem metoda redukcjonistyczna (*parsimony*), bardzo często uznawana za wolną od założeń (*assumption free*), więc jako taka wolna od następstw arbitralnych decyzji taksonoma: nie jest to prawdą, bowiem technika ta z natury swej zakłada najprostszy z możliwych przebieg ewolucji, co właściwie nigdy nie odpowiada rzeczywistemu procesowi. Brak sformułowanych założeń nie zapobiega odzwierciedlaniu przez rekon-

struowane drzewo określonego modelu ewolucji, w następstwie określonego zachowania aparatu matematycznego, użytego w danej technice – przykładem choćby omówione wcześniej własności różnych współczynników odległości czy podobieństwa.

Metody rekonstrukcji filogenezy można też podzielić na wykorzystujące odległości oraz korzystające bezpośrednio ze stanów cech u kolejnych taksonów. Te drugie uważa się zwykle za lepsze i zdecydowanie więcej technik należy do tej grupy. Nie znaczy to jednak, aby metody oparte na odległościach musiały być gorsze. Ich wynik zależy zarówno od wybranej techniki konstrukcji drzewa, jak i od własności samej odległości, użytej dla rekonstrukcji. W pewnych sytuacjach dane są wyjściowo odległościami, jak np. dla hybrydyzacji kwasów nukleinowych czy odległości immunologicznych, i wówczas nie mamy wyboru. W pozostałych przypadkach możemy zestawy stanów cech dla kolejnych OTU przeliczyć lub nie na macierz odległości pomiędzy tymi OTU. Użycie odległości umożliwia wykorzystanie obliczeniowo szybkich technik algorytmicznych. Zarazem prawidłowo dobrana odległość pozwoli na realistyczne odwzwierciedlenie wielkości zmian ewolucyjnych pomiędzy taksonami, co łatwo przedstawić na drzewie, a biologiczna interpretacja długości gałęzi takiego drzewa jest wówczas oczywista. Metody oparte na odległościach mogą więc dawać nie gorsze rekonstrukcje niż techniki wykorzystujące zestawy stanów cech dla poszczególnych taksonów. Ich wadami są natomiast redukcja informacji i pojawianie się niekiedy biologicznie nierealistycznych lub nieinterpretowalnych odległości na drzewie, pomimo ich matematycznej poprawności. Redukcja informacji, polegająca na zastąpieniu potężnego nieraz zestawu danych macierzą odległości, sama w sobie nie musi być mankamentem, niekiedy jest wręcz pożądana. Z drugiej strony, konwersja stanów cech w macierz odległości przekreśla możliwość rekonstrukcji ewolucji cech na obliczonym drzewie. Odległości na drzewie mogą być ujemne i wówczas nie dają się interpretować ewolucyjnie. Mogą też przybierać nieinterpretowalne wartości: np. dwa taksony terminalne dzielić może 52,5 substytucji. Oczywiście substytucja może zająć lub nie, nie istnieje substytucja połowiczna. Z drugiej strony odległości na drzewie traktować możemy jako estymaty, a wówczas 52,5 jest do przyjęcia, choć w rzeczywistości substytucji musiałoby być 52 lub 53.

Jest oczywiste, że jedne z technik rekonstrukcji filogenezy są lepsze, a inne gorsze w określonych warunkach. Aby porównać użyteczność różnych technik, należy je porównać przynajmniej w pięciu aspektach: **wydajności** (*efficiency*), **mocy** (*power*), **spójności** (*consistency*), **odporności** (*robustness*) oraz **falsyfikowalności** (*falsifiability*). Wydajność określa, jak szybkie są obliczenia. Oczywiście szybkie są metody algorytmiczne, gdy dla opartych na kryterium optymalizacji już około 20 taksonów to za dużo, by móc nie stosować technik niegwarantujących znalezienia najlepszego drzewa; najwolniejsze są techniki oparte na maksymalizacji wiarygodności. Moc metody jest tym większa, im mniej danych wystarcza dla osiągnięcia prawidłowej rekonstrukcji. Im więcej danych, tym pewniejsza rekonstrukcja, ale w praktyce danych zwykle jest nie za wiele; teoretycznie, nawet gdy metodę dobrano idealnie prawidłowo, dla niewielkiej liczby danych rekonstrukcja może być błędna w następstwie losowych błędów – niereprezentatywności użytego zestawu danych. Spójność oznacza, że przy dostatecznej liczbie danych rekonstrukcja będzie prawdziwa. Odporność techniki jest tym większa, w im większym stopniu można naruszyć jej założenia, wciąż uzyskując prawidłową rekonstrukcję, a falsyfikowalność oznacza, że metoda poinformuje użytkownika o złamaniu założeń, wykluczającym jej użycie. Teoretycznie dobra technika rekonstrukcji

filogenezy powinna co najmniej zadowalająco spełniać wszystkie pięć kryteriów. W praktyce takiej techniki brak i raczej trudno się spodziewać, by ją kiedyś wynaleziono, a wybierać musimy technikę najmniej niedoskonałą dla analizy naszych danych. Szerzej zajmiemy się tym w Rozdziale 4.11.

4.2. Metody algorytmiczne oparte na odległościach

Jak we wszystkich technikach opartych na odległościach, dane wyjściowe – będące wynikiem odpowiednich obliczeń dla danych empirycznych – traktujemy jako **doświadczalne przybliżenia** estymowanych **odległości rzeczywistych** (*patristic distances*), za które uznaje się sumy długości odpowiednich gałęzi zrekonstruowanego drzewa, łączących dane taksony. Gdyby więc odległości empiryczne były idealnie addytywne bądź ultrametryczne, to skonstruowanie drzewa odzwierciedlającego je idealnie wymagałoby zwyczajnego połączenia odpowiednich gałęzi. W praktyce nigdy tak nie jest i stąd konieczność wyboru którejś z metod estymacji drzewa. Estymujemy oczywiście długość gałęzi, ale też topologię – to ostatnie jest we współczesnej rekonstrukcji filogenezy ważniejsze, choć często brak formalnego związku między „prawdziwą” topologią a względną wartością danego kryterium optymalizacji: zakłada się po prostu, że dla najlepszej wartości parametru, czyli najlepszego dopasowania długości gałęzi, topologia też powinna być najlepsza. Z wielu zaproponowanych technik omówimy tu te najlepiej sprawdzone bądź najczęściej stosowane.

Metoda dołączania sąsiada

Metoda **dołączania sąsiada** (*neighbor joining* – NJ), zaproponowana przez Saitou i Nei (1987), to zapewne najlepsza z technik algorytmicznych. Konceptyjnie należy do metod poszukujących drzew, które odzwierciedlają możliwie najprostszy – obejmujący najmniej zdarzeń – proces ewolucyjny, możliwy do postulowania dla posiadanych danych. Takie właśnie jest założenie techniki kladystycznej (*parsimony*) czy minimalnej ewolucji (*minimum evolution*: jak zobaczymy, dołączanie sąsiada to algorytmiczny wariant tej techniki). Warto uświadomić sobie, że założenie rekonstruowania możliwie najprostszego przebiegu procesów ewolucyjnych nie ma jakiegokolwiek ontologicznego uzasadnienia – w rzeczywistości ewolucja mogła przebiegać w sposób nawet najbardziej skomplikowany, a ponad wszelką wątpliwość nie biegła w sposób najprostszy, w dodatku poszukiwanie najprostszego możliwego scenariusza wiąże się zawsze z pewną utratą danych. Z drugiej strony, zakładanie najprostszego z możliwych przebiegu wydarzeń ma uzasadnienie metodologiczne, opierające się na regule brzytwy Ockhama, głoszącej, aby nie tworzyć więcej bytów niż to nieodzowne. Poszukujemy więc dającej się określić dolnej granicy zbioru możliwych wydarzeń, bowiem inaczej nie mielibyśmy przesłanek określających granice sensownego komplikowania rekonstrukcji filogenezy – zawsze można by postulować kolejny scenariusz, jeszcze bardziej skomplikowany. Z drugiej strony czasem takie przesłanki są, a zawsze musimy pamiętać, że najprostszy model jedynie formalnie jest najlepszy.

Jak pamiętamy, analiza skupisk (najczęściej używana w wersji UPGMA) wymaga ultrametrycznych danych; w praktyce technika zachowuje się nieźle, jak długo odstępstwa od ultrametryczności nie są wielkie (Nei i Kumar 2000). Dołączanie sąsiada NJ

obliczeniowo przypomina UPGMA, ale technikę skonstruowano tak, aby nieultrametryczne dane korygować do mniej więcej ultrametrycznych, dopuszczając dla każdej gałęzi własne tempo ewolucji. Dane muszą być jednak choć w przybliżeniu addytywne. Warto pamiętać, że obliczone drzewo jest addytywne i nieukorzenione; ukorzenia się je dopiero techniką grupy zewnętrznej lub najdłuższej gałęzi, ukorzenie nie wynika jak w UPGMA z samego procesu konstrukcji drzewa, choć wiele programów rysuje takie drzewa jako ukorzenione. Przez **sąsiadów** (*neighbors*) rozumiemy dwa taksony połączone w obrębie nieukorzenionego drzewa jednym węzłem, w którym schodzą się gałęzie zakończone tymi terminalnymi OTU. Technika NJ oblicza położenie kolejnych węzłów, a nie OTU lub skupisk jak UPGMA.

Analizę rozpoczyna się od macierzy odległości oryginalnych i drzewa w postaci gwiazdy (*star tree*), od środka której rozchodzą się promieniście wszystkie gałęzie, zakończone taksonami terminalnymi. Macierz odległości przelicza się w taki sposób, że oddalenie pomiędzy dwoma węzłami jest wazone ich średnimi oddaleniami od wszystkich pozostałych węzłów, co koncepcyjnie odpowiada normalizacji odrębności każdego z taksonów średnim tempem ewolucji tego taksonu. W tak zmodyfikowanej macierzy znajdujemy parę najbliższych sobie węzłów (pierwsze dwa będą taksonami terminalnymi, czyli węzłami zewnętrznymi), które łączymy. Zarazem oba te taksony usuwamy z gwiazdy, na ich miejsce umieszczając łączący je węzeł – dla tych taksonów ancestralny – a więc gwiazda liczy o jeden takson mniej. Proces kontynuujemy aż do całkowitego przekształcenia gwiazdy w dychotomicznie rozgałęzione drzewo. Algorytm techniki dołączania sąsiada (Studier i Kepler 1988, Swofford i inni 1996, Nei i Kumar 2000) można przedstawić następująco:

- (1) z oryginalnej macierzy $(N - 1)N/2$ odległości d między N węzłami terminalnymi (na tym etapie N równe jest całkowitej liczbie taksonów, dla których obliczamy drzewo), zakładając że $d_{ii} = 0$, dla każdego węzła terminalnego i obliczamy jego **łączną odrębność** (*net divergence*) r_i od wszystkich pozostałych taksonów:

$$r_i = \sum_{k=1}^N d_{ik} ;$$

- (2) obliczamy macierz **M** odległości **skorygowanych** pod względem **tempa ewolucji** (*rate corrected*), której elementy definiujemy następująco:

$$M_{ij} = d_{ij} - (r_i + r_j)/(N - 2);$$

ponieważ przypadek $i = j$ oczywiście pomijamy, a macierz jest symetryczna, obliczamy elementy tej macierzy dla wszystkich i i dla $j > i$; zapisywanie całej macierzy jest zbędne, wystarczy zachowanie najmniejszej wartości M_{ij} ;

- (3) definiujemy nowy węzeł u , z którego wybiegają gałęzie do węzłów i i j oraz do reszty drzewa; wyznaczamy długość gałęzi v z węzła u do węzłów i oraz j :

$$v_{iu} = d_{ij} / 2 + (r_i - r_j) / [2(N - 2)], \quad v_{ju} = d_{ij} - v_{iu} ;$$

(4) definiujemy odległości od u do każdego z pozostałych węzłów terminalnych k (dla wszystkich $k \neq i, k \neq j$):

$$d_{ku} = (d_{ik} + d_{jk} - d_{ij})/2;$$

(5) usuwamy odległości od węzłów i oraz j z macierzy i zmniejszamy N o 1;

(6) jeżeli pozostało więcej niż dwa węzły, wracamy do kroku (1); jeżeli nie, to drzewo jest całkowicie skonstruowane, z wyjątkiem długości gałęzi łączących te ostatnie dwa węzły (i, j), którą przyjmujemy:

$$v_{ij} = d_{ij}.$$

Każdy cykl kroków (1) do (6) tworzy jeden wewnętrzny węzeł i estymuje długości gałęzi, łączących się w tym węźle; obliczone długości gałęzi i znaleziona topologia umożliwiają narysowanie nieukorzenionego, addytywnego drzewa.

Technika dołączania sąsiada, podobnie jak i inne poszukujące addytywnego drzewa, może przypisać ujemne wartości długościom niektórych gałęzi. Zdarzyć się tak może zwłaszcza w następstwie estymowania wartości, które w rzeczywistości są niewielkie. Ujemne wartości nie mają oczywiście biologicznego sensu, toteż Kuhner i Felsenstein (1994) zmodyfikowali algorytm tak, aby ujemne wartości zastępowane były zerem, a odpowiednia poprawka wprowadzana dla długości łączących się w tym węźle gałęzi tak, aby łączna odległość pomiędzy przylegającą parą węzłów terminalnych została zachowana. Jak każda z technik algorytmicznych, dołączanie sąsiada konstruuje jedno drzewo, nawet gdy jest ono jednym z wielu jednakowo uprawnionych. Kumar (1966) oraz Pearson i inni (1999) zaproponowali modyfikację algorytmu, dzięki której zamiast jednego obliczane jest więcej drzew, które można porównać; jest to więc technika pośrednia między algorytmicznym NJ a opartą na kryterium optymalizacji techniką minimalnej ewolucji. Gascuel (1997) zaproponował modyfikację algorytmu NJ, polegającą na przyjęciu różnych wag dla d_{ik} , d_{jk} i d_{ij} [krok (4)], dla zmniejszenia wariancji d_{ku} . Technika, nazwana przez autora BIONJ, ma zwiększać spójność metody w zakresie znajdowania prawidłowej topologii. Symulacje wykazały, że BIONJ zachowuje się nieco lepiej niż NJ, gdy różnice pomiędzy sekwencjami są wysokie (Gascuel 1997), ale w sumie obie techniki dają identyczne lub niemal identyczne drzewa (Nei i Kumar 2000).

Metoda dołączania sąsiada zazwyczaj błędnie zrekonstruuje topologię drzewa, gdy liczba n nukleotydów lub aminokwasów w sekwencjach jest niewielka (Rzhetsky i Nei 1993, Nei i inni 1998). Podobnie będzie zresztą, gdy użyjemy innych technik. Zachowanie NJ będzie wówczas tym gorsze, im większa będzie liczba taksonów T , dla których obliczamy drzewo; tak samo zresztą zachowują się wszystkie techniki rekonstrukcji filogenezy, choć krytyczna wartość n zależy od techniki i danych (Rozdział 4.11). Jak wykazały symulacje, dla dostatecznie wysokiego n i odpowiednio dobranego, nieobciążonego estymatu liczby substytucji nukleotydów użytego jako odległości NJ powinien prawidłowo zrekonstruować topologię drzewa (Saitou i Imanishi 1989, Rzhetsky i Nei 1992a, Gascuel 1997). O ile obliczenia prowadzone są z dostateczną dokładnością, technika ta bardzo rzadko znajduje – w kolejnych powtórzeniach – więcej niż jedno „najlepsze” drzewo (Takezaki 1998), co ją korzystnie wyróżnia. Zarazem czas obliczeń jest proporcjonalny do T^3 (Swofford i inni 1996), a więc liczba taksonów możliwa do analizy przy użyciu UPGMA jest też analizowalna dla NJ. W sumie więc

dołączanie sąsiada jest dobrą techniką rekonstrukcji filogenezy, a gdy liczba taksonów jest za wysoka dla metod stosujących kryterium optymalizacji, NJ pozwala na obliczenie przybliżonego drzewa, które następnie się optymalizuje.

Drzewo Wagnera

Drzewo Wagnera (*distance Wagner tree*) to zestaw technik algorytmicznych lub wykorzystujących kryterium optymalizacji, zaproponowanych dla danych molekularnych (Farris 1970, 1972, Swofford i inni 1996). O ile dołączanie sąsiada – a również szereg technik opartych na kryterium optymalizacji – zakłada, że odległości obliczone na podstawie danych empirycznych są raz zaniżonymi, a raz zawyżonymi estymatami odległości rzeczywistych, to obliczając drzewo Wagnera, przyjmujemy, że wszystkie empiryczne estymaty są zaniżone. Jak wiele razy wskazywaliśmy, rekonstruując przebieg ewolucji, nie potrafimy ocenić, ile zmian w rzeczywistości zaszło: to, że w miejscu, powiedzmy, guaniny znajduje się cytozyna, nie dowodzi bynajmniej, aby zajęć mogła jedynie substytucja $G \rightarrow C$, w rzeczywistości mogło być $G \rightarrow A \rightarrow C$ bądź nawet $G \rightarrow T \rightarrow G \rightarrow A \rightarrow C$. Jeżeli więc nie potrafimy korygować odległości tak, aby uwzględniły te „ukryte” zdarzenia, to sensowne jest przyjęcie, że nasze estymaty są w rzeczywistości dolną granicą przedziałów możliwych wartości. Tak działa technika: oblicza drzewo najkrótsze z możliwych przy założeniu, że odległości na drzewie nie mogą być mniejsze od empirycznych. Dla odległości addytywnych technika działa prawidłowo, lecz jeżeli dane nie są addytywne, to zachowanie techniki staje się niepewne; metoda jest obecnie rzadko stosowana.

Technika rozkładu cząstkowych rekonstrukcji

Technika **rozkładu cząstkowych rekonstrukcji** (*split decomposition*), zaproponowana przez Bandelta i Dressa (1992), pozwala na graficzne przedstawienie tendencji w macierzy odległości. Techniki konstruujące drzewa działają tak, że dla każdego danego obliczą drzewo, choć założeniem przedstawienia związków między taksonami za pomocą drzewa jest hierarchiczna struktura danych, czyli wyraźne grupowanie się analizowanego zbioru taksonów w podzbiory, jak już o tym pisaliśmy, omawiając analizę skupisk. W przypadku technik opartych na kryterium optymalizacji niehierarchiczność danych znajdzie odzwierciedlenie w szeregu drzew o całkiem różnej topologii, lecz takiej samej lub niemal takiej samej wartości kryterium. Stosując metody algorytmiczne, nie wiemy, na ile wiarygodna może być obliczona rekonstrukcja. Wówczas właśnie pomocna będzie technika rozkładu cząstkowych rekonstrukcji (Swofford i inni 1996, Page i Holmes 1998). Przez **cząstkowe rekonstrukcje** (*splits, partitions*) rozumiemy uzyskane na podstawie części danych – niekoniecznie dychotomiczne – drzewa odzwierciedlające zróżnicowanie w obrębie badanego zbioru taksonów. Jeżeli mamy więc, powiedzmy, taksony A, B, C, D i E, to część cech może wskazywać np. na cząstkową rekonstrukcję ((A,B),(C,D,E)), a część na ((A,B,C),(D,E)). Technika rozkładu cząstkowych rekonstrukcji wykaże dobrze uzasadnione grupowanie, jeśli takie występuje, a także niezgodne z sobą grupowania, jeżeli mają uzasadnienie w danych. Takie niezgodności zdarzają się nierzadko, w następstwie niedostatecznej kompensacji nieobserwowalnych zmian ewolucyjnych przy transformacji danych w odległości, a także w wyniku homoplazji czy ewolucji siatkowatej.

Metoda opiera się na czteropunktowym warunku addytywności (patrz Rozdział 2.9), zgodnie z którym jeżeli pokrewieństwa taksonów I, J, K i L (tworzących łącznie tzw. kwartet) odzwierciedla drzewo ((I,J),(K,L)) i odległości są addytywne, to:

$$d_{IJ} + d_{KL} < d_{IK} + d_{JL} = d_{IL} + d_{JK} .$$

Gdy dane nie są w pełni addytywne, powyższa równość nie będzie zachodzić, dalsze odstępstwo od addytywności pociągnie za sobą niespełnianie również i nierówności. Dopóki ta nierówność jest spełniona, można konstruować addytywne kwartety gatunków, a z nich składać całą rekonstrukcję – na tym polega **metoda sąsiedztwa** (*neighborliness*: Sattath i Tversky 1977, Fitch 1981). W praktyce często przynajmniej niektóre kwartety nie spełniają nawet nierówności. Technika rozkładu cząstkowych rekonstrukcji roboczo zakłada, że $d_{IJ} + d_{KL}$ przynajmniej nie ma najwyższej wartości spośród wszystkich trzech sum. Techniki filogenetyczne przyjmują, że jeżeli suma $d_{IJ} + d_{KL}$ jest większa od $d_{IK} + d_{JL}$ i $d_{IL} + d_{JK}$, to dane wykluczają drzewo o topologii ((I,J),(K,L)) (Swofford i inni 1996). Możemy jednak zapytać, w jakim stopniu dane są zgodne z pozostałymi możliwymi topologiami, stosując to samo kryterium. Dla każdej topologii obliczyć więc można **współczynnik odrębności** (*isolation index*):

$$S_{AB,CD} = [(d_{AD} + d_{BC}) - (d_{AB} + d_{CD})] / 2; \quad S_{AC,BD} = [(d_{AD} + d_{BC}) - (d_{AC} + d_{BD})] / 2 .$$

Zamiast wewnętrznej gałęzi, łączącej dwa wewnętrzne węzły drzewa, narysować możemy oko sieci, o bokach określonych tymi współczynnikami. Jeżeli taksonów jest więcej niż cztery, to odchylenia od warunku czteropunktowego obliczamy dla każdego możliwego do wyróżnienia kwartetu. Bandelt i Dress (1992) wykazali, że jedynie część cząstkowych rekonstrukcji może być przedstawiona na płaskim grafie (*split decomposable portion*); wyrazili też przypuszczenie, że pozostałe odzwierciedlają większość losowego szumu w danych. Gdyby tak było, metoda pozwalałaby oddzielać sygnał od szumu, lecz zapewne pozostaje to jedynie pobożnym życzeniem. Często też analizę należy prowadzić dla nie więcej niż 4–10 taksonów jednocześnie, aby wyniki były czytelne (Swofford i inni 1996).

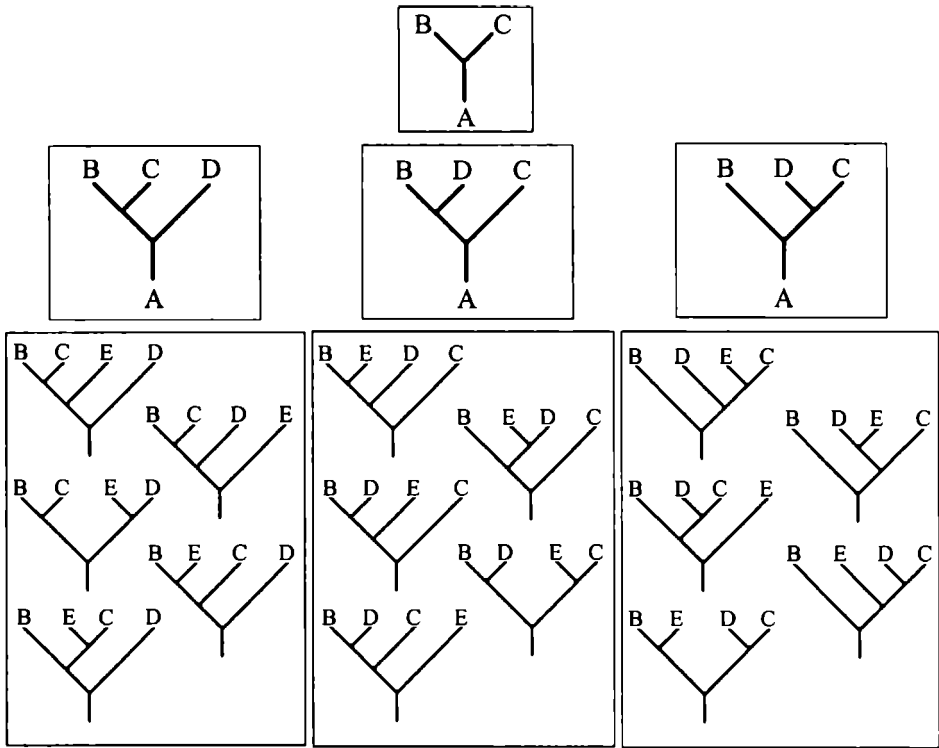
4.3. Problem znalezienia najlepszego drzewa

Jak wielokrotnie mówiliśmy, techniki oparte na kryterium optymalizacji uważa się za lepsze od algorytmicznych. Ich istotą jest porównywanie dobroci kolejnych drzew, które generowane być muszą odrębnym algorytmem. Techniki generowania drzew dzielimy na **pewne** (*exact*) i **heurystyczne** (*heuristic*), czyli częściowe i niepewne. Z rozdziału o drzewach wiemy już, dlaczego nie możemy zawsze stosować technik pewnych: jeżeli już dla 20 taksonów terminalnych liczba drzew nieukorzenionych równa jest 221 643 095 476 699 771 875, a dla 63 przekracza 10^{100} , to jest oczywiste, że już dla około 20 taksonów generowanie wszystkich drzew i obliczanie wartości kryterium optymalizacji dla każdego z nich zbliża się nieuchronnie do granic możliwości obliczeniowych najszybszych komputerów.

Techniki pewne

Najprostszą metodą jest technika **pełnego poszukiwania** (*exhaustive search, implicit enumeration*). Bierzemy dowolne trzy spośród taksonów, wchodzących w skład grupy, której filogenezę rekonstruujemy. Dla trzech taksonów możliwe jest tylko jedno nieukorzenione drzewo (Ryc. 4.1). Następnie dołączamy dowolny z pozostałych taksonów: uczynić to można na jeden z trzech sposobów, a więc mamy już trzy drzewa (Ryc. 4.1). Kolejny, piąty takson, dołączyć można na pięć sposobów do każdego z trzech drzew, więc możliwych nieukorzenionych drzew jest już 15 (Ryc. 4.1). Szósty takson dołączyć można na siedem sposobów do każdego z 15 drzew, a więc powstanie 105 nieukorzenionych drzew (jak pamiętamy, dla T taksonów drzew ukorzenionych jest tyle co nieukorzenionych dla $T+1$). Proces jest kontynuowany aż do włączenia wszystkich taksonów na wszystkie możliwe sposoby (dla każdego z kompletnych drzew obliczamy wartość optymalizowanego kryterium), jeżeli czas obliczeń nie przekroczy akceptowalnych granic, a więc jeżeli taksonów nie jest więcej niż kilkanaście. Technika pełnego poszukiwania jest najwolniejsza, lecz jako jedyna pozwala na uzyskanie rozkładu wartości optymalizowanego parametru, co ma znaczenie przy ocenie wiarygodności rekonstrukcji (Rozdział 4.11).

Rozkładu takiego nie zapewni nam technika **limitowanego (ograniczonego) dołączania** (*branch-and-bound*: Hendy i Penny 1982, Kumar i inni 1993, Swofford i Begle 1993). Także pewna, czyli gwarantująca znalezienie najlepszego drzewa, jest od pełnego poszukiwania szybsza, a więc umożliwia analizę dla większej liczby taksonów. Odbywa się to kosztem pomijania licznych topologii, a więc rezygnacji ze znajomości rozkładu wartości optymalizowanego parametru. Powiedzmy, że stosując technikę kładystyczną, poszukujemy najkrótszego drzewa. Jedną z technik heurystycznych (patrz niżej) znaleźliśmy drzewo o długości, założmy, 256. Przyjmujemy więc 256 za górną granicę długości drzewa: być może jest więcej niż jedna topologia o takiej długości, jak znaleziona techniką heurystyczną, być może też istnieją wśród możliwych rekonstrukcji drzewa krótsze. Rozpoczynamy więc obliczenia techniką limitowanego dołączania. Identycznie jak w pełnym poszukiwaniu konstruowane są kolejne drzewa, poprzez dokładanie na wszystkie możliwe sposoby kolejnych gałęzi. W trakcie dokładania kolejnych gałęzi obliczana jest długość drzewa (a ogólniej: wartość optymalizowanego parametru). Jeżeli osiągnie ona limit – w naszym przypadku długość 256 – jeszcze przed dołączeniem ostatniej gałęzi, to wiemy, że jakiegokolwiek dołączenie kolejnej gałęzi może jedynie wydłużyć drzewo, a więc dalsze dołączanie należy przerwać i przejść do dołączania gałęzi do innego z krótszych, niekompletnych drzew. W ten sposób oszczędzamy czas, nie konstruując zbytecznie części drzew, niewątpliwie dłuższych niż 256, i to jest przewagą limitowanego dołączania nad pełnym poszukiwaniem. Jeżeli na którymkolwiek etapie uzyskamy kompletne drzewo o długości mniejszej, np. 255, to tę nową długość przyjmujemy za nowy limit i zapisujemy wyłącznie drzewa o tej długości, na niej też porzucając dalsze dołączanie.



Ryc. 4.1. Kolejne kroki przy pełnym poszukiwaniu. Dla którychkolwiek trzech taksonów tworzymy jedyne możliwe trójtaksonowe drzewo, dołączamy na trzy możliwe sposoby czwarty takson. Następnie do każdego z trzech czteretaksonowych drzew dołączamy na wszystkie możliwe sposoby piąty takson, uzyskując $5 \times 5 = 25$ drzew: są to wszystkie możliwe drzewa nieukorzenione dla pięciu taksonów

Tu można by zapytać, o ile szybsza jest ta metoda? Odpowiedź zależy od charakteru analizowanych danych, a ściślej od tego, na ile są one hierarchiczne, czyli w jakim stopniu są rzeczywiście pogrupowane taksony, w świetle danych stanowiących podstawę rekonstrukcji filogenezy. Inaczej mówiąc, jak silny jest sygnał filogenetyczny w stosunku do filogenetycznego szumu. Im więc więcej homoplazji bądź symplezjomorfii, tym gorzej, zajmiemy się tym jeszcze w dalszych rozdziałach. W skrajnym przypadku „złych” danych, czyli danych niehierarchicznych, technika limitowanego dołączania może wygenerować tyle samo drzew co metoda pełnego poszukiwania, a więc zająć tyle samo czasu – będzie tak, gdy „najlepszych” drzew jest wiele (nawet setki czy tysiące), a górna granica długości osiągana jest dopiero po dołączeniu ostatniego taksonu. Inna rzecz, że analiza filogenetyczna dla takich danych skazana jest na niepowodzenie: jakiegokolwiek metody by nie użyć, wyniki będą mało wiarygodne i niejednoznaczne. Z drugiej strony zdarza się, że dane są właśnie takie.

Opisana sytuacja to jednak skrajność i zwykle technika limitowanego dołączania jest zdecydowanie szybsza niż pełne poszukiwanie. Dodać trzeba, że dla doświadczalnych (nie generowanych losowo) danych, zebranych przez taksonoma mającego pewną wiedzę o użyteczności różnych cech w analizie filogenezy w danej grupie, technika ta będzie wielokrotnie szybsza, umożliwiając analizę dla nawet ponad 20 taksonów.

Znalezienie – techniką przybliżoną – wstępnego limitu dla optymalizowanego parametru nie jest obligatoryjne, lecz przyspiesza obliczenia. Podobnie przyspieszy je dołączanie w pierwszej kolejności taksonów najbardziej się między sobą różniących, tak aby wzrost długości drzewa następował szybciej (Swofford 1996, Swofford i inni 1996). Kolejne dołączanie prowadzimy, wykorzystując **algorytm max-mini** (*max-mini algorithm*: Nei i Kumar 2000): każdorazowo dołączamy (pozostając przy przykładzie szukania najkrótszego drzewa) ten z niedołączonych wcześniej taksonów, którego dołączenie najbardziej zwiększy (max) długość drzewa, choć oczywiście dołączamy go w takim miejscu, aby wzrost długości drzewa był najmniejszy (mini).

Techniki heurystyczne

Dla większej liczby taksonów, chcąc nie chcąc, skorzystać musimy z technik heurystycznych. Z technik tych skorzystamy również dla mniejszej liczby taksonów (powiedzmy kilkunastu), jeżeli dane są niewystarczające – jak np. sekwencje liczą za mało par nukleotydów – bowiem wówczas rekonstrukcja i tak najprawdopodobniej będzie błędna, więc nie ma sensu prowadzenie długotrwałych obliczeń (Nei i Kumar 2000). Popularnie ujmując, techniki heurystyczne są niepełne i niepewne, formalnie to grupa metod, dla których nie istnieje aparat matematyczny – brak algorytmu umożliwiającego pewne znalezienie najlepszego drzewa, jeżeli nie zostaną sprawdzone wszystkie możliwe – konieczne jest więc znalezienie rozwiązania przybliżonego bądź niepewnego na drodze empirycznej, czyli w praktyce wypróbowując jedynie część topologii. W dodatku dla drzew trudno mówić o rozwiązaniach „przybliżonych” – ta właściwa i jedynie „nieco lepsza” topologia może skrajnie się różnić od tej „niemal najlepszej”.

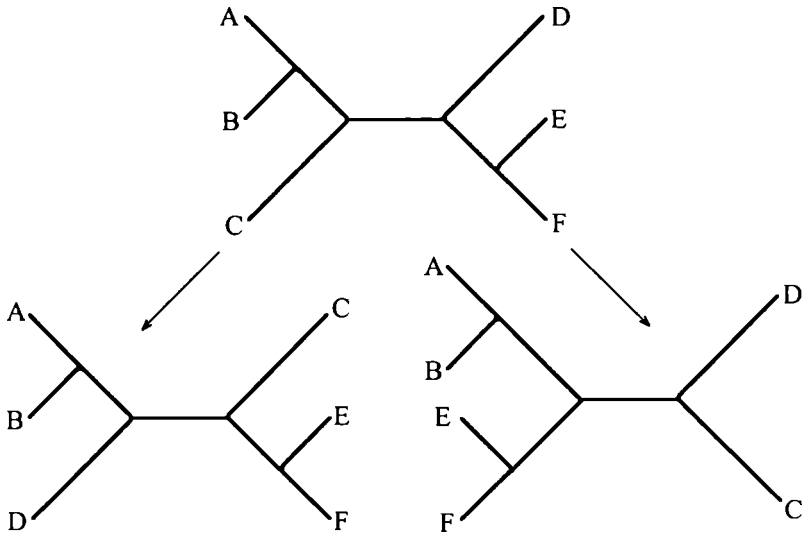
Aby zilustrować problem, używa się często niekoniecznie najbardziej realistycznego porównania. Wyobraźmy sobie komandosa, który zrzucony zostaje nocą, ciemną i mglistą, w nieznanym pagórkowatym terenie. Zadanie, które musi wykonać, wymaga dotarcia na szczyt wzgórza najwyższego w okolicy. Jedyne co może on zrobić, to iść pod górę, mając nadzieję, że znajduje się na stoku tego właśnie wzgórza. Jest oczywiste, że albo wykona w ten sposób zadanie, albo nie, ale nic innego zrobić nie może. Poszukiwanie najlepszego drzewa metodami heurystycznymi napotyka ten sam problem: każda z technik polega na iteratywnym zwiększaniu dobroci drzewa, czyli optymalizowaniu wartości danego kryterium (najczęściej omawia się to na przykładzie metody kladystycznej, czyli szukania najkrótszego drzewa). W końcu zostaje osiągnięta wartość najlepsza, ale nie wiemy, czy znaleziono rzeczywiście najlepsze drzewo, czy też raczej jedno z **lokalnych optimumów**. Chodzi o to, że rozkład wartości kryterium optymalizacji przypominać może rzeczywiście teren górzysty, z licznymi lokalnymi optimumami. Wówczas znalezienie innego optimum wymaga konstruowania drzewa od początku, bądź niemal od początku. Techniki heurystyczne nie poprzestają więc na znalezieniu jednego optimum, lecz nawet znalezienie wielu optimumów nie może gwarantować, że nie ma innego, „bardziej optymalnego”. I tu jednak sytuacja zależy od danych, użytych do rekonstrukcji. Liczne optima lokalne są charakterystyczne dla drzew generowanych losowo, a także dla drzew obliczonych na podstawie danych empirycznych, lecz z licznymi homoplazjami, słabo hierarchicznych, o niskim poziomie filogenetycznego sygnału. Im lepsze dane, tym mniejsze zagrożenie lokalnymi optimumami i tym większa szansa znalezienia najlepszego drzewa technikami heurystycznymi. Znow jednak: dane są, jakie są, trudno poza tym wskazać granicę pomiędzy dany-

mi dostatecznie i niedostatecznie dobrymi, a ponadto gwarancji znalezienia najlepszego drzewa nie ma nigdy. Jeszcze do tego wrócimy.

Stopniowe dodawanie (*stepwise addition*) polega na każdorazowej optymalizacji przy dołączaniu kolejnego taksonu. Bierzemy pierwsze trzy taksony i łączymy w nieukończone drzewo, dla którego obliczamy wartość kryterium optymalizacji (pozostaliśmy przy wygodnym przykładzie poszukiwania drzewa najkrótszego), po czym dołączamy czwarty takson tak, aby wartość kryterium była najlepsza (czyli w naszym przypadku – drzewo było najkrótsze), a więc próbujemy je dołączać w każdym z trzech możliwych miejsc, aby wybrać najlepszą topologię. Do niej dołączamy piąty takson, znów sprawdzając wszystkie pięć możliwych topologii i wybierając najlepszą. Do najlepszej w taki sam, najlepszy sposób dołączamy szósty takson, i tak dalej, aż wszystkie taksony zostaną dołączone. Jak łatwo dostrzec, uzyskane drzewo – i charakteryzująca je wartość optymalizowanego parametru – zależeć będzie od kolejności dołączania taksonów. Algorytm należy bowiem do grupy „**algorytmów maksymalizujących doraźny zysk**” (*greedy algorithms*): każdy kolejny takson dołączany jest optymalnie dla tego etapu konstruowania drzewa, choć dołączenie go w innym miejscu może prowadzić do uzyskania lepszego kompletnego drzewa. Drzewo uzyskane w wyniku jednorazowego zastosowania stopniowego dodawania nie będzie drzewem optymalnym nawet dla „dobrych” danych – chyba że przez przypadek kolejność dodawania taksonów będzie też optymalna, co jest dla większej liczby taksonów niezmiernie mało prawdopodobne.

Wynik techniki zależy zarówno od wyboru pierwszej trójki (trypletu) taksonów, jak i kolejności dołączania następnych. Kolejność może odpowiadać kolejności taksonów w macierzy danych, możemy ją zmieniać i powtarzać obliczenia, najlepiej używając generatora liczb pseudolosowych dla losowania kolejności w powtarzanych obliczeniach – tak działają programy z pakietu PHYLIP (Felsenstein 2000: opcja *Jumble*). Zwykle zaleca się wówczas co najmniej 10 powtórzeń, lecz raczej powinno ich być więcej. Inna metoda to sprawdzenie wartości kryterium optymalizacji dla wszystkich możliwych trójek taksonów i wybranie tego trypletu, dla którego ta wartość jest najlepsza. Następnie wybiera się ten z taksonów, którego dołączenie zapewni najlepszą wartość kryterium, itd., aż do umieszczenia wszystkich taksonów na konstruowanym drzewie (Swofford i Begle 1993, Swofford 1996). Technika ta, określana jako *closest* (najbliższy), jest obliczeniowo intensywna. Farris (1970) zaproponował technikę, którą nazwał prostym algorytmem (*simple algorithm*): wybieramy któryś z taksonów, zwykle pierwszy w macierzy danych (lecz może to być jakikolwiek z OTU bądź hipotetyczny przodek), i obliczamy między nim a każdym z pozostałych taksonów współczynnik, będący sumą absolutnych wartości różnic stanów cech (tzw. *advancement index*). Początkowy tryplet powstaje więc z tego „taksonu odniesienia” i dwóch innych, mających najniższe wartości tego współczynnika, a następne taksony dołączamy w kolejności rosnących wartości współczynnika. Nie istnieje wariant stopniowego dodawania, który by był najlepszy dla każdego danych. Warto więc spróbować wielu opcji. Stopniowe dodawanie stosuje się dla znalezienia wyjściowego drzewa, które poddaje się dalszej optymalizacji, natomiast porzucenie na drzewie obliczonym tą techniką jest ryzykowne.

W pewnym sensie odwrotnością stopniowego dodawania jest technika **rozkładu gwiazdy** (*star decomposition* – SD), choć jej zastosowanie ogranicza się do metod, w których wartość kryterium optymalizacji obliczyć można dla drzewa politomicznego. Technikę tę wykorzystuje omówione wcześniej dołączanie sąsiada, lecz również niektóre techniki maksymalizacji wiarogodności (Swofford i inni 1996). Początkowo łączy się wszystkie taksony terminalne w gwiazdę, o jednym wewnętrznym węźle. Następnie odłączamy kolejno każdą z możliwych par taksonów, łącząc je odrębnym, drugim węzłem wewnętrznym. Wybieramy to z uzyskanych drzew o dwóch wewnętrznym węzłach, dla którego wartość kryterium optymalizacji jest najlepsza. Następnie odłączamy z gwiazdy kolejny takson, wybierając ten, którego odłączenie da najlepsze drzewo, po czym następny, i tak aż do uzyskania w pełni dychotomicznego drzewa. I ten algorytm maksymalizuje doraźną korzyść, a więc budzi te same zastrzeżenia co stopniowe dodawanie.

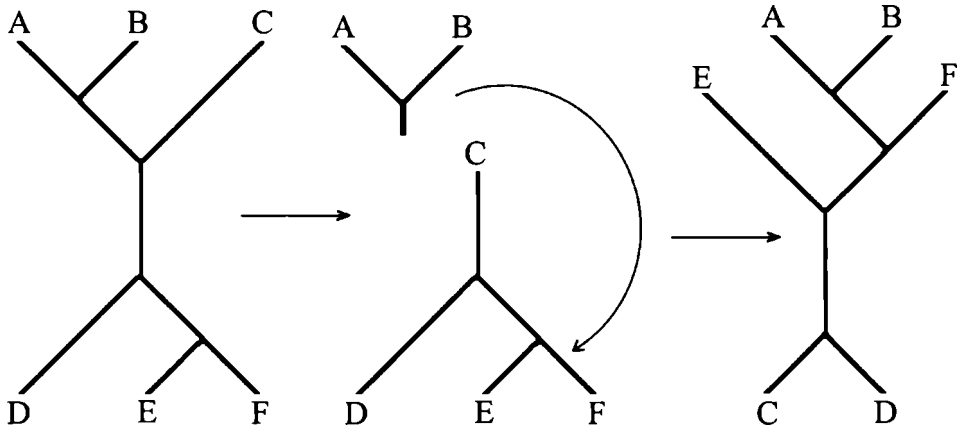


Ryc. 4.2. Wymiana najbliższego sąsiada, czyli lokalna wymiana gałęzi. Dla jednej gałęzi wewnętrznej możliwe są dwie takie zamiany

Dalsze poszukiwania optymalnego drzewa prowadzimy metodą **wymiany gałęzi** (*branch swapping*). Jego istotą jest mniejsze lub większe zmodyfikowanie topologii drzewa, aby wyjść z lokalnego optimum – nowe drzewo, nawet jeżeli jest lepsze pod względem optymalizowanego kryterium, znów nie daje gwarancji, że nie znaleźliśmy kolejnego lokalnego optimum, zamiast poszukiwanego drzewa najlepszego. Warto tu ponownie podkreślić, że dla topologii nie można mówić o „w przybliżeniu najlepszym drzewie”. Oczywiście dla każdego drzewa wskazać można inne, różniące się odeń minimalnie: długość drzewa, powiedzmy 134, jest niemal identyczna z długością 133. Problem jednak leży w tym, że niemal identyczne topologie mogą być bardzo różne pod względem optymalizowanego parametru i na odwrót – najmniejsza zmiana wartości kryterium optymalizacji może odpowiadać zupełnej przebudowie topologii. Poszu-

kujemy więc drzewa najlepszego – choć może być więcej niż jedno takie drzewo – a nie „niemal najlepszego”. W dodatku to najlepsze oddzielać może od modyfikowanego przez nas, lokalnie optymalnego, szereg innych lokalnych optimów. Warto więc kolejne optymalizacje zaczynać od różnych topologii, nawet zdecydowanie gorszych od najlepszych dotąd znalezionych. Wymianę gałęzi najczęściej prowadzi się jedną z trzech technik: **wymiany najbliższego sąsiada** (*nearest neighbor interchange* – NNI), **przemieszczania cząstkowego drzewa** (*subtree pruning and regrafting* – SPR) czy **rozcinięcia i powtórne łączenia drzewa** (*tree bisection and reconnection* – TBR).

Każda wewnętrzna gałąź w pełni dychotomicznego drzewa łączy się z czterema „najbliższymi sąsiadami”, czyli innymi gałęziami, po dwie na każdym z kończących gałęź węzłów. Wymiana najbliższego sąsiada NNI (Ryc. 4.2), zwana też **lokalną wymianą gałęzi** (*local branch swapping*), polega na zamianie miejscami tych gałęzi: jak widać, dla jednej gałęzi wewnętrznej możliwe są dwie takie zamiany (Ryc. 4.2), a pełny proces wymiany najbliższego sąsiada obejmuje takie zamiany przeprowadzone kolejno dla wszystkich wewnętrznych gałęzi – być może któraś ze zmian utworzy drzewo o lepszej wartości kryterium optymalizacji. NNI to mniej efektywna technika wymiany gałęzi (Kitching i inni 1998) niż metody **globalnej wymiany** (*global branch swapping*): SPR i TBR, polegające na odrywaniu i przemieszczaniu w inne miejsce fragmentu drzewa, oczywiście próbując różnych topologii i każdorazowo obliczając wartość kryterium optymalizacji dla nowego drzewa.

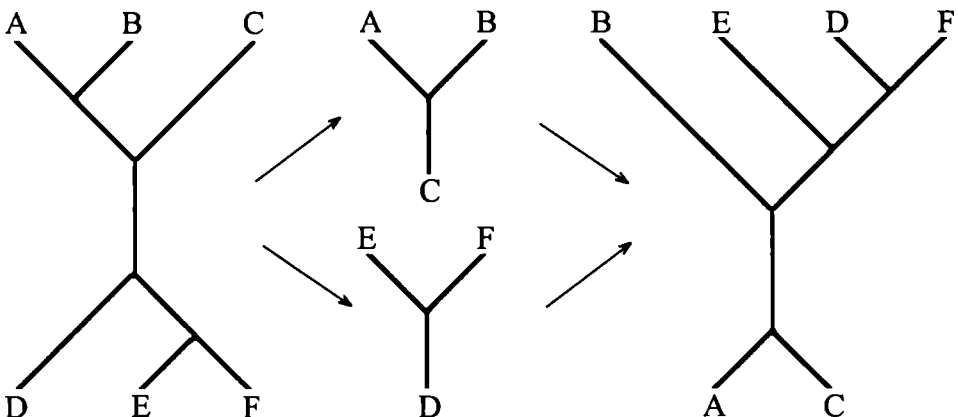


Ryc. 4.3. Technika przemieszczania cząstkowego drzewa. Ukorzeniony fragment drzewa zostaje oderwany od pozostałej części drzewa i dołączany kolejno do każdej z gałęzi tej pozostałej części

Technika przemieszczania cząstkowego drzewa SPR (Ryc. 4.3) polega na oderwaniu ukorzenionego drzewa od pozostałej części (*residual*) drzewa i łączeniu tego ukorzenionego drzewa kolejno z każdą z gałęzi pozostałej części drzewa. W skrajnym przypadku oderwanym drzewem może być jeden takson, lecz można też przemieszczać i pół drzewa, tyle że przemieszczane drzewo cząstkowe pozostaje ukorzenione, czyli dołączane do pozostałej części drzewa zawsze tą samą gałęzią. Brakiem ukorzenienia

obu łączonych, cząstkowych drzew charakteryzuje się technika rozcinania i powtórno-
go łączenia drzewa TBR (Ryc. 4.4), w której wszelkie możliwe sposoby łączenia
cząstkowych drzew są sprawdzane. TBR generuje więc najwięcej drzew, toteż jest naj-
efektywniejsza spośród technik wymiany gałęzi, zarazem jednak i najintensywniejsza
obliczeniowo. W podstawowej wersji drzewo dzielone jest na dwa, lecz niektóre pro-
gramy pozwalają nawet na podział drzewa na 10 – znów rośnie prawdopodobieństwo
znalezienia drzewa optymalnego, choć także za cenę wydłużenia obliczeń. Tak więc
dla większej liczby taksonów również bardziej rozbudowane użycie TBR przestaje być
możliwe, choć istnieją nieco szybsze algorytmy wymiany gałęzi (Goloboff 1996). Co
więcej, nawet pełne wykorzystanie TBR nie gwarantuje znalezienia najlepszego drze-
wa (Maddison 1991).

Niezłą strategią, dającą wysokie prawdopodobieństwo znalezienia optymalnego
drzewa, jest wielokrotne użycie stopniowego dodawania, po czym zastosowanie TBR
dla każdego z tak obliczonych drzew (Nei i Kumar 2000). Inna rzecz, że dla osiągnię-
cia rzeczywiście wysokiego prawdopodobieństwa stopniowe dodawanie przeprowadzić
należy też rzeczywiście wiele razy – to „wiele”, jak zwykle, z przyczyn praktycznych
będzie mniejsze od teoretycznie wymaganego, w dodatku im więcej taksonów, tym
więcej powtórzeń jest wymagane. W każdym razie dla, powiedzmy, 50 taksonów 100
powtórzeń z trudem wystarczy. Niejednokrotnie na wykresie wartości kryterium opty-
malizacji występują tzw. „płaskowyże dobroci drzew”, czyli szereg topologii o tej sa-
mej wartości kryterium, wszystkie suboptymalne. W takich przypadkach jedynie za-
chowanie większej liczby suboptymalnych topologii, a także szeregu topologii nieco
gorszych od suboptymalnych, po czym prowadzenie dalszej optymalizacji – najlepiej
TBR – dla każdej z nich może umożliwić wyjście z takiego płaskowyżu i znalezienie
rzeczywiście najlepszego drzewa (Swofford i inni 1996, Kitching i inni 1998). Jeżeli
optymalizacja drzew o odmiennych topologiach, znalezionych w wyniku stopniowego
dodawania, prowadzi każdorazowo do tej samej topologii, to zwiększa to wiarygod-
ność takiego drzewa, choć oczywiście też nie gwarantuje, że jest to globalne optimum.



Ryc. 4.4. Technika rozcinania i powtórno-
go łączenia drzewa. Charakteryzuje ją brak ukorzenia obu
łączonych cząstkowych drzew, więc wszystkie możliwe sposoby łączenia zostają sprawdzone

Kumar i inni (1993) zaproponowali technikę heurystyczną, która przypomina metodę limitowanego dołączania. Przez analogię do stosowanego tam algorytmu max-mini w technice tej wykorzystuje się algorytm **min-mini** (*min-mini algorithm*), który na każdym etapie wybiera minimum z minimalnych długości drzew, trzymając się przykładu poszukiwania najkrótszego drzewa. Po wyborze najkrótszego trypletu dołącza się czwarty z taksonów na jeden z trzech możliwych sposobów i oblicza długość tego drzewa, przyjmując tę wartość za tymczasowy limit na tym etapie. Podobnie dołącza się kolejne i zapamiętuje ich topologie wraz z kolejnymi wartościami limitów, aż do dołączenia wszystkich taksonów. Następnie – analogicznie jak w limitowanym dołączaniu – oblicza się długość kolejnego z czterotaksonowych drzew – gdy jest mniejsza, uznaje się ją za nowy limit, gdy większa, porzuca się dalsze dodawanie taksonów, itd. Algorytm max-mini wykorzystuje jednak globalne górne granice na kolejnych etapach, gdy stosowany tu algorytm min-mini korzysta z lokalnych górnych granic, nie gwarantuje więc znalezienia najlepszego drzewa (Nei i Kumar 2000). Można limit dla każdego etapu powiększyć o zadaną wartość, stałą bądź proporcjonalną do liczby taksonów na danym etapie (tzw. *search factor*), co oczywiście zwiększa szanse znalezienia najkrótszego drzewa, choć też i wydłuża obliczenia. Szacowanie wiarygodności znalezionych drzew (techniką *bootstrap*, Rozdział 4.11), obliczonych dla 48 sekwencji DNA o długości 1000 par nukleotydów każda (w technice kladystycznej raczej za krótka sekwencja dla pewniejszej rekonstrukcji), dało takie same wyniki dla drzew obliczonych metodami wymiany najbliższego sąsiada i – teoretycznie znacznie lepszą – techniką rozcinania i powtórno łączenia (Takahashi i Nei 2000). Kolejny raz potwierdza to, że dla kiepskich danych nie warto prowadzić długich obliczeń, bowiem wynik i tak będzie niepewny.

4.4. Metody stosujące kryterium optymalizacji oparte na odległościach

Podobnie jak przy opisanych wcześniej technikach algorytmicznych, odległości obliczone na podstawie danych empirycznych uznajemy za przybliżenia estymowanych odległości rzeczywistych, odzwierciedlonych sumami długości odpowiednich gałęzi na zrekonstruowanym drzewie. Konstruując drzewo albo próbuje się jak najdokładniej odzwierciedlić na nim doświadczalne odległości, co jest istotą **technik najmniejszych kwadratów** (m.in. Fitch-Margoliash), albo też poszukuje się drzewa najkrótszego, co zakłada metoda **minimalnej ewolucji**. W każdej metodzie przyjmuje się wzajemną niezależność kolejnych odległości, a warunek ten zwykle nie jest spełniony – chociażby błąd w ustaleniu stanów cech u jednego z taksonów pociąga za sobą błędne odległości między tym taksonem a wszystkimi innymi. Zakłada się też addytywność odległości. Wynikiem obliczeń jest drzewo addytywne nieukorzenione, które ukorzeniać trzeba metodą grupy zewnętrznej lub najdłuższej gałęzi. Istnieje jednak modyfikacja techniki najmniejszych kwadratów (program KITSCH w pakiecie PHYLIP), obliczająca drzewo ultrametryczne (Felsenstein 2000). Technika przypomina algorytmiczną analizę skupisk, lecz góruje nad nią obliczaniem więcej niż jednego drzewa, poddawanego optymalizacji.

Techniki najmniejszych kwadratów i pokrewne

Grupa należących tu metod za kryterium optymalizacji uznaje minimalizowanie różnicy między znanymi doświadczalnymi przybliżeniami d_{ij} odległości między taksonami a poszukiwanymi estymatami rzeczywistej odległości ewolucyjnej e_{ij} między tymi taksonami, estymatami odzwierciedlonymi na obliczonym drzewie. Minimalizowana jest więc wartość E błędu dopasowania danych doświadczalnych do danego drzewa o T taksonach terminalnych:

$$E = \sum_{i=1}^{T-1} \sum_{j=i+1}^T w_{ij} |d_{ij} - e_{ij}|^{\alpha},$$

gdzie α może mieć wartość 1 lub 2, a w_{ij} to schemat ważenia, różnicujący metody tej grupy. Dla $\alpha = 1$ jest to tzw. **statystyka f** Farris'a, minimalizowana jest suma wartości bezwzględnych błędów dopasowania, zaś dla $\alpha = 2$ jest to **kryterium najmniejszych kwadratów**, powszechnie stosowane w matematyce i statystyce. Warto podkreślić, że drzewo obliczone techniką najmniejszych kwadratów, jeśli tylko wartości błędów dopasowania mają rozkład normalny, będzie nie gorsze od uzyskanego techniką maksymalizacji wiarygodności, uznawaną często za najlepszą, choć wymagającą bez porównania dłuższych obliczeń. Zakłada się więc, że odległość pomiędzy każdą parą taksonów jest elementem rozkładu, którego wartością oczekiwaną jest suma odległości – w rozumieniu wielkości ewolucyjnych zmian – mierzona na drzewie od jednego do drugiego węzła terminalnego, a wariancja tego rozkładu jest proporcjonalna do potęgi p wartości oczekiwanej, przy czym znajomość – lub oszacowanie – wartości p decyduje o przyjętym schemacie ważenia. Jeżeli więc wiemy, że błędy dopasowania estymatów odległości są mniej więcej podobne dla wszystkich danych, stosujemy kryterium najmniejszych kwadratów ($\alpha = 2$), natomiast jeżeli mamy podstawy sądzić, że niektóre z błędów dopasowania – lecz nie wiadomo które – są szczególnie duże, to bezpieczniej będzie minimalizować sumę wartości bezwzględnych błędów, czyli przyjąć $\alpha = 1$. Podkreślmy, że zwykle przyjmujemy $\alpha = 2$, a w przypadku szczególnie błędnych wartości – które jednak potrafimy wskazać – również stosujemy kryterium najmniejszych kwadratów, natomiast tym konkretnie wartościom przypisujemy szczególnie niskie wagi w_{ij} . Właśnie schemat ważenia dopasowujemy do oceny błędów naszych danych, dokładniejszej niż świadomość, że pewne odległości są szczególnie błędne.

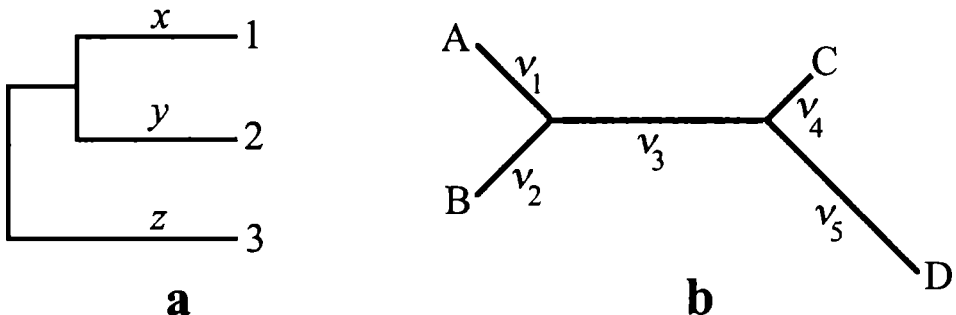
Najczęściej stosowane schematy ważenia, to: (1) $w_{ij} = 1/d_{ij}^0 = 1$, jest to więc nieważona technika najmniejszych kwadratów, zaproponowana przez **Cavalli-Sforzę i Edwardsa** (1967); $p = 0$; (2) $w_{ij} = 1/d_{ij}^2$, technika zaproponowana przez **Fitcha i Margoliasha** (1967); $p = 2$; (3) $w_{ij} = 1/d_{ij}$, technika **pośrednia** między pierwszą i drugą; $p = 1$. We wszystkich trzech przypadkach wielkości błędów nie potrafimy ocenić, możemy jedynie wskazać, jaki charakter mają te błędy. W przypadku gdy przypuszczamy, że wszystkie eksperymentalne odległości obarczone są błędem o podobnej wartości absolutnej, najwłaściwsze jest użycie techniki Cavalli-Sforzy i Edwardsa, czyli (1), a więc nieważonej techniki najmniejszych kwadratów. Będzie tak choćby wówczas, gdy w macierzy odległości obok dużych wartości mamy odległości bardzo małe, bowiem wtedy znaczącym źródłem błędu będzie zaokrąglenie, o podobnej wartości bezwzględnej zarówno dla odległości małych, jak i dużych. Częściej przypusz-

czać można, że dla wszystkich danych doświadczalnych podobne wartości będzie miał błąd procentowy i wówczas najwłaściwsza będzie metoda Fitcha i Margoliasha (2). Pośrednia technika (3) powinna być stosowana wówczas, gdy spodziewamy się, że błędy są proporcjonalne do pierwiastków kwadratowych wartości w macierzy (Felsenstein 2000). Oczywiście często trudno jest ocenić, który z tych trzech modeli najlepiej odpowiada naszym danym. Teoretycznie najwłaściwszy sposób ważenia to: $w_{ij} = 1/\sigma_{ij}^2$, gdzie σ_{ij}^2 to spodziewana wariancja d_{ij} (Swofford i inni 1996). W praktyce wielkość wariancji trudno ocenić, technika wymaga długich obliczeń, a dla małych odległości wariancja zbliża się do zera. W przypadku np. identycznych sekwencji w_{ij} staje się nieokreślona (mianownik równy zero). W sumie więc ten sposób ważenia budzi wątpliwości (Swofford i inni 1996, Nei i Kumar 2000).

Warto zaznaczyć, że zwykle wybór jednej z powyższych technik ważenia nie wpływa zbyt mocno na topologię obliczonego drzewa. Fitch i Margoliash (1967) zaproponowali tzw. **średnie procentowe odchylenie standardowe** s :

$$s = \left\{ \frac{\sum_{i=1}^{T-1} \sum_{j=i+1}^T [(d_{ij} - e_{ij})/d_{ij}]^2}{n(n-1)} \right\}^{1/2} * 100\%$$

dla porównywania dobroci odwzorowania odległości oryginalnych na różnych drzewach. Oczywiście im wartość s jest mniejsza, tym lepsze odwzorowanie. Wielkość ta, znana jako **kryterium Fitcha i Margoliasha**, nie powinna być mylona z techniką Fitcha i Margoliasha i może charakteryzować drzewo obliczone inną metodą. Kryterium jest użyteczne, zwłaszcza gdy porównujemy drzewa obliczone z różnych danych: np. interesuje nas, czy drzewo obliczone na podstawie odległości Cavalli-Sforzy i Edwardsa jest lepsze od obliczonego dla odległości Nei, czy też może jest na odwrót. Dla różnych danych porównywanie wartości E nie ma oczywiście sensu, natomiast standaryzowane kryterium s to umożliwia.



Ryc. 4.5. Zasada obliczeń addytywnego drzewa techniką Fitcha i Margoliasha (a), na drodze rozpatrywania kolejnych drzew trójtaksonowych (takson 1 + takson 2 + „takson 3” – czyli cała reszta taksonów). Na skonstruowanym – tu czterotaksonowym – drzewie (b) obliczyć można kolejne e_{ij} między taksonami, a na ich podstawie wartość kryterium E . Bliższe objaśnienia w tekście

Obliczanie drzewa klasyczną metodą Fitcha i Margoliasha (1967) przypomina w sumie analizę skupisk UPGMA, tyle że dane są nie ultrametryczne, a addytywne – właściwie analizę skupisk można by uznać za algorytmiczny odpowiednik techniki Fitcha i Margoliasha, tak jak dołączanie sąsiada to algorytmiczna wersja techniki minimalnej ewolucji. Metoda Fitcha i Margoliasha (Weir 1990, Nei i Kumar 2000) wykazuje to, że dla trzech taksonów: 1, 2 i 3 długości gałęzi łączącego je drzewa mogą być jednoznacznie obliczone (Ryc. 4.5a):

$$d_{12} = x + y, \quad d_{13} = x + z, \quad d_{23} = y + z, \quad \text{a więc:}$$

$$x = (d_{12} + d_{13} - d_{23})/2, \quad y = (d_{12} - d_{13} + d_{23})/2, \quad z = (d_{12} + d_{13} + d_{23})/2.$$

Znajdujemy więc w macierzy odległości wartość najmniejszą i taksony, pomiędzy którymi występuje ta wartość, uznajemy za 1 i 2. Taksony te łączymy, wyznaczając zakończone nimi gałęzie o długościach równych połowie tej oryginalnej odległości i łączących się w wewnętrznym węźle, natomiast wszystkie pozostałe taksony to „łączny” takson 3. Tak więc odległość między 1 a 3 to średnia odległości 1 od wszystkich innych taksonów z wyjątkiem 2, a odległość między 2 a 3 to oczywiście średnia odległości 2 od wszystkich pozostałych taksonów z wyjątkiem 1. Zamiast taksonów 1 i 2 mamy więc nowy węzeł – toteż liczba węzłów do dołączenia zmniejszyła się o jeden. Przeliczamy więc część macierzy odległości tak, aby zamiast odległości od taksonów 1 i 2 uzyskać odległości od nowego węzła wewnętrznego: będą to oczywiście odległości średnie oryginalnych odległości od 1 i 2. W tej nowej macierzy znów wyszukujemy odległość najmniejszą – oczywiście niekoniecznie od nowego wewnętrznego węzła – i znów za taksony 1 i 2 uznajemy te taksony (węzły), między którymi występuje ta najmniejsza wartość, a całą resztę za „łączny” takson 3. W ten sposób postępujemy aż do skonstruowania drzewa, obejmującego wszystkie taksony; podobieństwo do analizy skupisk jest niewątpliwe.

Na skonstruowanym drzewie obliczyć można odległości e_{ij} między taksonami, a na ich podstawie wartość kryterium E . Teoretycznie dla znalezienia optymalnej wartości E konieczne byłoby porównanie wszystkich możliwych topologii, co jak wiemy jest niemożliwe przy większej liczbie taksonów T . W praktyce więc bada się jedynie część możliwych topologii, drzewo obliczone powyższym algorytmem poddając optymalizacji metodami heurystycznymi z grupy technik wymiany gałęzi. Matematycznie bardziej wiarygodny (choć w praktyce obliczone drzewa są podobne) i szybki algorytm obliczania drzew techniką najmniejszych kwadratów przedstawiają Rzhetsky i Nei (1992a, 1993) oraz Nei i Kumar (2000). Dla T taksonów mamy $T(T-1)/2$ odległości między taksonami d_{ij} , którym odpowiadają na nieukorzenionym drzewie odpowiednio e_{ij} , obliczone przez sumowanie odpowiednich z $2T-3$ gałęzi o długościach v_{ij} (Ryc. 4.5b). Matematycznie przedstawić więc możemy związki między długościami gałęzi a odległościami pomiędzy taksonami macierzą A o $T(T-1)/2$ rzędach i $2T-3$ kolumnach oraz elementach $A_{(ij)k}$ określonych tak, że $A_{(ij)k} = 1$, gdy gałąź k jest częścią szlaku łączącego taksony i oraz j , a gdy nie, to równa jest zero. Wówczas:

$$e_{ij} = \sum_{k=1}^{2T-3} A_{(ij)k} v_k .$$

Dla drzewa złożonego z czterech taksonów (Ryc. 4.5b) można więc zapisać:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} = \begin{bmatrix} e_{AB} \\ e_{AC} \\ e_{AD} \\ e_{BC} \\ e_{BD} \\ e_{CD} \end{bmatrix}, \text{ czyli } \mathbf{Av} = \mathbf{e}.$$

Oczywiście dla odległości doskonale addytywnych, dla każdej pary taksonów i, j $d_{ij} = e_{ij}$, jednak dla danych eksperymentalnych nigdy tak nie jest i konieczne jest wykorzystanie powyższej macierzy dla znalezienia takich wartości v_k , które minimalizują E . Obliczenia przeprowadza się za pomocą programowania liniowego bądź kwadratowego (Barrsdale i Roberts 1973) lub iteratywnie (Felsenstein 2000) albo – gdy $\alpha = 2$ i $w_{ij} = 1$ – wykorzystując algebrę liniową (Cavalli-Sforza i Edwards 1967, Olsen 1988):

$$\mathbf{v} = (\mathbf{A}^T \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{d}), \text{ a dla kryterium ważonego: } \mathbf{v} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{W} \mathbf{d}),$$

gdzie \mathbf{W} to macierz odpowiednich wag, o $T(T-1)/2 \times T(T-1)/2$ elementach (Swofford i inni 1996), gdzie wzdłuż przekątnej znajdują się wagi dla kolejnych odległości między parami taksonów, a pozostałe elementy równe są zeru.

Symulacje komputerowe przeprowadzone przez Saitou i Nei (1986) oraz Rzhetsky'ego i Nei (1992a) wykazały, że prawdopodobieństwo znalezienia poprawnej topologii drzewa przy użyciu technik najmniejszych kwadratów i pokrewnych, ważonych bądź nie, bywa często niższe niż dla niektórych innych metod opartych na odległościach. Głównym źródłem tej niedoskonałości jest tendencja tych technik do znajdowania ujemnych wartości dla długości niektórych gałęzi. Inna rzecz, że od wady tej nie są wolne i inne metody, jak choćby minimalnej ewolucji (Swofford i inni 1996, Felsenstein 2000). Ujemne długości gałęzi nie mają oczywiście sensu – nie da się ich zinterpretować ewolucyjnie – zaproponowano więc różne sposoby obejścia tego problemu. Można zwyczajnie pomijać drzewa z jakąkolwiek ujemną długością gałęzi, jednak ryzykuje się wówczas pominięcie drzew bliskich optymalnemu. Można też uznać wszystkie wartości ujemne za równe zeru, postępowanie takie zmniejsza jednak dokładność rekonstrukcji i także może prowadzić do nieznaledzenia drzewa optymalnego. Wreszcie zmodyfikowano algorytm tak, aby długości gałęzi nie mogły być ujemne (Swofford i inni 1996, Felsenstein 1997, 2000, Nei i Kumar 2000), dzięki iteratywnej optymalizacji. Kuhner i Felsenstein (1994) wykazali, że ta modyfikacja wydatnie zwiększa prawdopodobieństwo znalezienia prawidłowej topologii drzewa.

Jak wspominaliśmy, metody najmniejszych kwadratów estymują długość gałęzi nie gorzej niż maksymalizujące wiarygodność, jeżeli błędy mają rozkład normalny. Rzhetsky i Nei (1993) wykazali, że warunek ten jest spełniony dla odpowiednio długich sekwencji. Tak jak przy innych technikach, rekonstrukcje filogenezy są tym bardziej wiarygodne, im więcej danych: jeżeli sekwencje są długie (lub cech wiele), a taksonów nie za dużo. Im więcej taksonów a mniej cech, tym większe prawdopodobieństwo błędnej

rekonstrukcji. Jest jednak teoretyczny problem. Omawiane techniki minimalizują E , dobierając odpowiednio długości gałęzi drzewa. Analiza filogenetyczna wymaga jednak również – jeżeli nie przede wszystkim – znalezienia właściwej topologii drzewa. Formalnie brak jakiegokolwiek związku pomiędzy kryterium E a topologią drzewa – właściwie milcząco zakłada się, że najlepsze dopasowanie długości gałęzi do empirycznych odległości możliwe jest wyłącznie dla drzewa o prawidłowej – optymalnej – topologii. Formalnie nikt tego jednak nie wykazał. Jeżeli użyjemy nieobciążonych estymatów zmian ewolucyjnych jako odległości, obliczonych dla (niemal) nieskończonej długiej sekwencji, to $E = 0$ jedynie dla prawidłowej topologii. W praktyce sekwencje są skończonej – zwykle niewielkiej – długości, a estymaty odległości obciążone błędami, wynik może więc być różny.

Ocenić wiarygodność techniki rekonstrukcji filogenezy jest trudno (Rozdział 4.11): jeżeli taksonów jest więcej, zawsze pozostaje niepewność, i gdy różne techniki dają odmienny wynik, to nie potrafimy powiedzieć, który z nich jest lepszy. Pozostają więc nie do końca miarodajne proste przykłady bądź symulacje. Wynik tych ostatnich zależy jednak od danych, które nie zawsze są realistyczne. To tłumaczy szereg sprzecznych z sobą opinii, porównujących moc, spójność i odporność różnych technik, także z rodziny najmniejszych kwadratów, porównywanych z techniką minimalnej ewolucji (Swofford i inni 1996, Felsenstein 2000, Nei i Kumar 2000). Kuhner i Felsenstein (1994) wykazali, że techniki najmniejszych kwadratów niezezwalające na ujemne długości gałęzi dają spójne topologie dla wysokiej liczby nukleotydów. Kidd i inni (1974) wskazali, że gdy długości gałęzi nie mogą być ujemne, to techniki te dają takie same drzewa jak metoda minimalnej ewolucji.

Technika minimalnej ewolucji

Kidd i Sgaramella-Zonta (1971) zaproponowali użycie nieważonego kryterium najmniejszych kwadratów jak w technice Cavalli-Sforzy i Edwardsa (1967) dla znajdowania długości gałęzi, lecz odmiennego kryterium optymalizacji przy porównywaniu drzew, tym lepszych, im niższa wartość tego kryterium:

$$LS = \sum_{k=1}^{2T-3} |v_k|; \text{ podobne kryterium wprowadzili Rzhetsky i Nei (1992a): } S = \sum_{k=1}^{2T-3} v_k.$$

Jak widać, obie techniki zakładają minimalizację długości gałęzi v_k na obliczonym drzewie, różniąc się jedynie tym, że wartość absolutna kryterium LS sugeruje dopuszczalność długości ujemnych, natomiast ujemne długości gałęzi są wykluczone, gdy wykorzystuje się kryterium S. W praktyce metody dają bardzo podobne wyniki, bowiem ujemne wartości nie są regułą, a gdy już występują, to są bardzo bliskie zeru (Swofford i inni 1996), choć przy występowaniu losowych błędów estymatów odległości krótkie gałęzie mogą mieć długości ujemne, w następstwie przypadku, także dla prawidłowej topologii (Nei i Kumar 2000). Rzhetsky i Nei (1992a) nazwali zaproponowaną technikę metodą **minimalnej ewolucji** (*minimum evolution*, ME). Jak już pisaliśmy, algorytmiczną wersją tej techniki jest dołączanie sąsiada, obie techniki poszukują rekonstrukcji najprostszej z możliwych w świetle posiadanych danych. Technika minimum ewolucji daje spójne rekonstrukcje filogenezy i jest często stosowana, choć

argumenty o jej wyższości nad technikami najmniejszych kwadratów, sformułowane przez Rzhetsky'ego i Nei (1992b), nie są przekonujące (Kidd i inni 1974, Kuhner i Felsenstein 1994, Swofford i inni 1996), zwłaszcza że dotyczą krótkich sekwencji (gdy różnice pomiędzy takimi krótkimi sekwencjami są duże, to odległości cechuje wysoka wariancja), których użycie daje niepewne wyniki dla każdej metody. W sumie wydaje się, że dobroć technik obu grup jest podobna.

Rzhetsky i Nei (1993) wykazali, że dla nieobciążonych estymatów ewolucyjnych odległości wartość spodziewana S staje się najmniejsza dla prawdziwej topologii, bez względu na liczbę taksonów. Topologia o najniższym S nie musi jednak być nieobciążonym estymatem prawdziwej topologii (Nei i Kumar 2000). Podobnie jak w technikach najmniejszych kwadratów, teoretycznie aby znaleźć najlepsze drzewo metodą ME, należałoby zbadać wszystkie możliwe topologie, co jednak byłoby zbyt czasochłonne dla większej liczby taksonów. Zważywszy, że dla małej liczby taksonów T drzewa obliczone techniką dołączania sąsiada (NJ) są identyczne lub niemal identyczne ze skonstruowanymi techniką ME, dla większej liczby taksonów oblicza się najpierw drzewo metodą NJ, po czym oblicza wartości S dla zestawu drzew podobnych do znalezionej techniką NJ. Drzewo o najniższej wartości S uznaje się czasowo za drzewo ME, po czym znów bada drzewa doń podobne, wybiera to o najniższej wartości S , itd., by w końcu drzewo o najniższej wartości S z wszystkich znalezionych w kolejnych cyklach poszukiwań uznać za drzewo ME, czyli optymalne. Efektywną techniką modyfikacji jest algorytm wymiany bliskiego sąsiada (*close neighbor interchange*). Szczegóły obliczeń S oraz D (różnicy między S dla dwóch porównywanych drzew) przedstawiają Nei i Kumar (2000). Wybierając drzewo, powinniśmy wprawdzie zbadać statystyczną istotność różnicy D między wartościami S dla porównywanych drzew, przy wykorzystaniu testu Z (Nei i Kumar 2000), co wymaga znajomości błędu standardowego D . Rzhetsky i Nei (1992a, 1993) przedstawiają algorytm umożliwiający obliczanie błędu standardowego D dla różnych modeli substytucji, ale jego stosowanie dla większej (mniej więcej ponad 50) liczby sekwencji zaczyna być problematyczne z uwagi na czas obliczeń.

4.5. Metoda kladystyczna (redukcjonistyczna)

Spośród technik rekonstrukcji filogenezy, bezpośrednio korzystających ze stanów cech dla poszczególnych taksonów, niewątpliwie najszerzej i najczęściej stosowana jest metoda **kladystyczna**, czyli **redukcjonistyczna** (*parsimony*, *maximum parsimony*). Angielską nazwę przetłumaczyć najlepiej by można jako „skąpstwo”, bowiem „oszczędność” to termin za słaby – technika w każdym razie opiera się, podobnie jak opisane już dołączanie sąsiada czy minimalnej ewolucji, na poszukiwaniu możliwie najprostszego przebiegu rekonstruowanej ewolucji. Ponownie podkreślić należy, że technika nie ma uzasadnienia ontologicznego – rzeczywiste procesy mogły być i pewnie były daleko bardziej skomplikowane – a jedynie epistemologiczne: wprawdzie można wskazać rozwiązanie najprostsze, lecz nawet najbardziej skomplikowane można by bardziej skomplikować. Metoda redukcjonistyczna pozwala więc na uzyskiwanie rekonstrukcji obiektywnie porównywalnych między sobą, choć niekoniecznie najbliższych rzeczywistości. Dla danych molekularnych techniki *maximum parsimony* zachowują się nienajgorzej, choć ustępują metodom masyalizacji wiarygodności, nato-

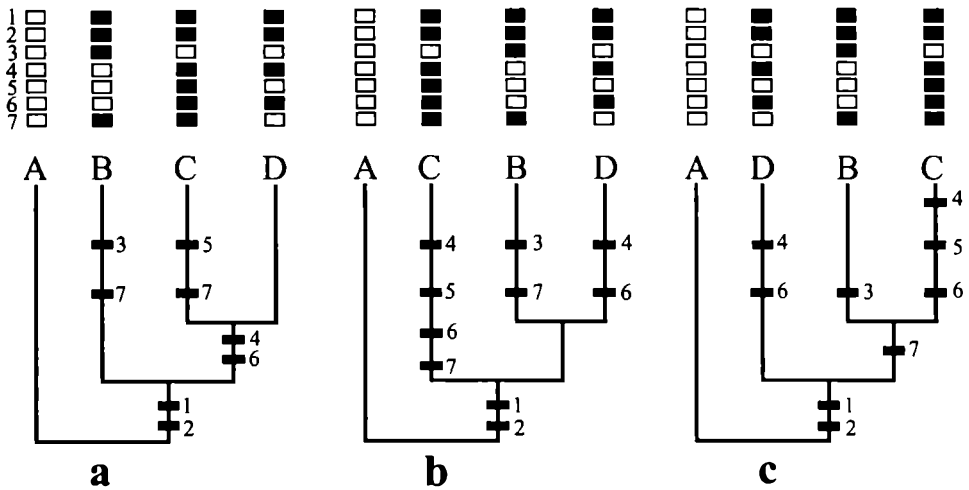
miast dla danych morfologicznych są zdecydowanie najczęściej stosowane i – zapewne – najlepsze, zwłaszcza że umożliwiają również śledzenie ewolucji stanów cech. Techniki te stanowią też istotę rekonstrukcji kladystycznych: są bliskie metodologii zaproponowanej przez Henniga (1966) i przez większość kladystów uznawane za jedyne w pełni filogenetyczne (np. Kitching i inni 1998). Musimy więc omówić je nieco szerzej.

Jeżeli u hipotetycznego przodka występował, powiedzmy, stan a pewnej cechy, a u potomka stan b , to oczywiście musiała mieć miejsce zmiana: $a \rightarrow b$, była to jedna zmiana. Licząc zmiany zachodzące na gałęziach drzewa, wynikające ze stanów cech w kolejnych węzłach, obliczyć można długość drzewa. Gdyby nie było jakichkolwiek homoplazji, łatwo znaleźlibyśmy to jedyne, właściwe drzewo. W praktyce jednak tak nigdy nie jest i musimy, z konieczności, tworzyć *ad hoc* hipotezy o homoplazjach, z których każda zwiększa liczbę zmian zrekonstruowanych na drzewie, czyli długość tego drzewa. Zgodnie z techniką kladystyczną, czyli redukcjonistyczną (można by też nazwać ją minimalistyczną), poszukujemy drzewa, dla którego liczba takich hipotez *ad hoc* jest najniższa, czyli **drzewa najkrótszego**. Formalnie można to wyrazić jako poszukiwanie – spośród wszystkich możliwych drzew – zestawu wszystkich drzew τ , których długość $L(\tau)$ ma wartość najmniejszą (Swofford i inni 1996):

$$L(\tau) = \sum_{k=1}^G \sum_{j=1}^N w_j \text{diff}(x_{k'j}, x_{k''j}),$$

gdzie G to liczba gałęzi drzewa τ , N to liczba cech, k' i k'' to dwa węzły wyznaczające każdą z gałęzi k , $x_{k'j}$ i $x_{k''j}$ to stany cech w macierzy danych dla węzłów terminalnych bądź optymalne rekonstrukcje stanów cech dla węzłów wewnętrznych drzewa, $\text{diff}(y, z)$ to funkcja określająca koszt transformacji cechy ze stanu y do stanu z wzdłuż jakiegokolwiek gałęzi drzewa. Co ważne, $\text{diff}(y, z)$ nie musi być równe $\text{diff}(z, y)$, a funkcja ta określona jest różnie dla różnych sposobów optymalizacji stosowanych w metodzie kladystycznej; w_j to waga przypisana danej cesze j . Gdy dla każdej z cech w_j ma tę samą wartość równą 1, współczynnik ten możemy we wzorze, rzecz jasna, pominąć i mówimy wówczas o **nieważonej technice kladystycznej** (*unweighted parsimony*), zaś gdy przybiera różne wartości dla różnych cech, to jest to **ważona technika kladystyczna** (*weighted parsimony*).

Przedstawianie techniki kladystycznej rozpoczniemy od prostego, Hennigiańskiego przykładu optymalizacji czterotaksonowego drzewa, na podstawie znajomości stanów siedmiu cech (Ryc. 4.6). Technika jest nieważona, każda z cech jest binarna i przejście ze stanu pierwotnego do zaawansowanego dla każdej z cech kosztuje 1, zwykle mówi się wówczas o jednym **kroku** (*step*). Zarazem wiadomo dla każdej z cech, który stan jest pierwotny (plezjomorficzny, zgodnie z konwencją znaczący jasnym prostokątem), a który zaawansowany (apomorficzny, zgodnie z konwencją znaczący czarnym prostokątem). Nad symbolami taksonów umieszczono tzw. **kostki danych** (*data boxes*), graficznie przedstawiające stany cech u każdego z taksonów. Dla każdej z trzech możliwych topologii na kolejnych gałęziach drzewa zaznaczono cechy, których stany zmieniły się z plezjomorficznych na apomorficzne: jak wiemy, zgodnie z Hennigiem (1966) takie przejście – powstanie apomorfii – jest warunkiem niezbędnym uznania kladu za naturalny. Warto zaznaczyć, że takie zmiany stanów cech zaznaczamy dla całej gałęzi, ani nie będąc w stanie, ani nie usiłując bliżej określać, w którym momencie anagenezy zachodzącej wzdłuż tej gałęzi zmiana stanu cechy miała miejsce.



Ryc. 4.6. Wszystkie trzy możliwe topologie (a, b, c) nieukorzonego czterotaksonowego drzewa. Nad każdym z taksonów terminalnych (A, B, C, D) umieszczono kostki danych, czyli graficzną prezentację stanów cech: tu siedmiu cech binarnych; zgodnie z konwencją stan plezjomorficzny jest zaznaczony prostokątem niewypełnionym (jasnym), a apomorficzny – prostokątem wypełnionym (ciemnym). Podobnie ciemnymi prostokątami zaznaczymy na samych drzewach uzyskanie stanu apomorficznego dla kolejnych cech, które musiało mieć miejsce dla danej topologii na określonej gałęzi, aby wytłumaczyć rozkład stanów cech obserwowany u taksonów terminalnych: każde takie uzyskanie to krok. Dla różnych topologii liczba kroków jest różna

Jak wiemy, w odtwarzaniu filogenezy użyteczne są jedynie synapomorfie, a więc **filogenetycznie informacyjne** (*phylogenetically informative*) mogą być jedynie te z cech, których stany zaawansowane występują co najmniej u dwóch taksonów terminalnych; pozostałe są **pozbawione informacji filogenetycznej** (*phylogenetically uninformative*): w omawianym przykładzie to cechy 3 i 5. Rozróżnienie to nie do końca jednak jest słuszne: zarówno autapomorfie, jak i utrzymujące się symplezjomorfie niosą informację filogenetyczną, wykorzystywaną choćby w technikach opartych na odległościach bądź maksymalizacji wiarygodności, a bezużyteczne są jedynie w technice kladystycznej, toteż właściwie powinniśmy je określać mianem **kladystycznie pozbawionych informacji** (*parsimony uninformative*). Odtworzone zmiany stanów cech możemy policzyć na każdym z drzew (Ryc. 4.6a–c): dla drzewa (a) 8 kroków, dla drzewa (b) 10 kroków, zaś dla (c) – 9 kroków. Drzewo (a) jest więc najlepsze, jako najkrótsze, jest to więc **najlepsza rekonstrukcja** (*most parsimonious reconstruction* – MPR). Najlepsza rekonstrukcja (MPR) to najkrótsze drzewo, ale też rekonstrukcja stanów cech w węzłach wewnętrznych przeprowadzona tak, aby drzewo było najkrótsze, choć jego topologia nie musi być najkrótszą z możliwych: mówimy więc o topologii MPR, ale też o rekonstrukcji MPR stanów cech dla danej topologii. Zauważmy, że dla rozważanych danych – jak niemal zawsze w praktyce – nie istnieje rekonstrukcja, w której każda z cech zmieniałaby stan jedynie raz: na drzewie (a) stan cechy 7 ulega zmianie dwukrotnie, na drzewie (b) stan cechy 4 dwukrotnie i stan cechy 6 trzykrotnie, natomiast na drzewie (c) – 4 dwukrotnie i 6 dwukrotnie. To są właśnie wspomniane

wcześniej hipotezy *ad hoc*, których liczbę należy minimalizować, choć zupełnie ich wykluczyć nie da się.

Sposoby optymalizacji drzewa w metodzie kladystycznej

Omówiony powyżej przykład zakładał skrajnie prostą transformację stanów cech. W rzeczywistości, jak była o tym mowa w Rozdziale 2.3, transformacje możliwe dla danej cechy mogą być różne, różne więc mogą być sposoby optymalizacji, czyli obliczania kosztów przejścia jednego stanu w drugi. Camin i Sokal (1965) zaproponowali optymalizację zakładającą nieodwracalność zmian stanów cech, zwaną **optymalizacją Camina i Sokala** (*Camin-Sokal parsimony*). Przypominamy, że raz nabyty stan cechy nigdy nie może zostać utracony, choć może go zastąpić następny stan ($a \rightarrow b \rightarrow c$, $c \neq a$). Jeżeli więc dwa węzły drzewa mają ten sam stan cechy, to węzeł, z którego pochodzą, ma ten sam stan i długość drzewa się nie zmienia. Gdy stany są różne, to stan pierwotniejszy przypisujemy węzłowi wyjściowemu, a długość drzewa rośnie o liczbę kroków pomiędzy stanami w tych dwóch potomnych węzłach. Optymalizacja Camina i Sokala możliwa jest oczywiście jedynie dla drzewa ukorzonego, którego korzeń charakteryzuje zestaw plezjomorfii – gdy tak nie jest, to ukorzenie należy zmienić. U podstaw techniki leży założenie, że stan zaawansowany cechy może zostać uzyskany więcej niż raz, lecz nigdy nie utracony – nasza wiedza o procesach ewolucji każe uznać takie założenia za mało realistyczne, toteż optymalizacja Camina i Sokala stosowana jest bardzo rzadko, a jako jedyna – dla wszystkich serii transformacyjnych – praktycznie nigdy.

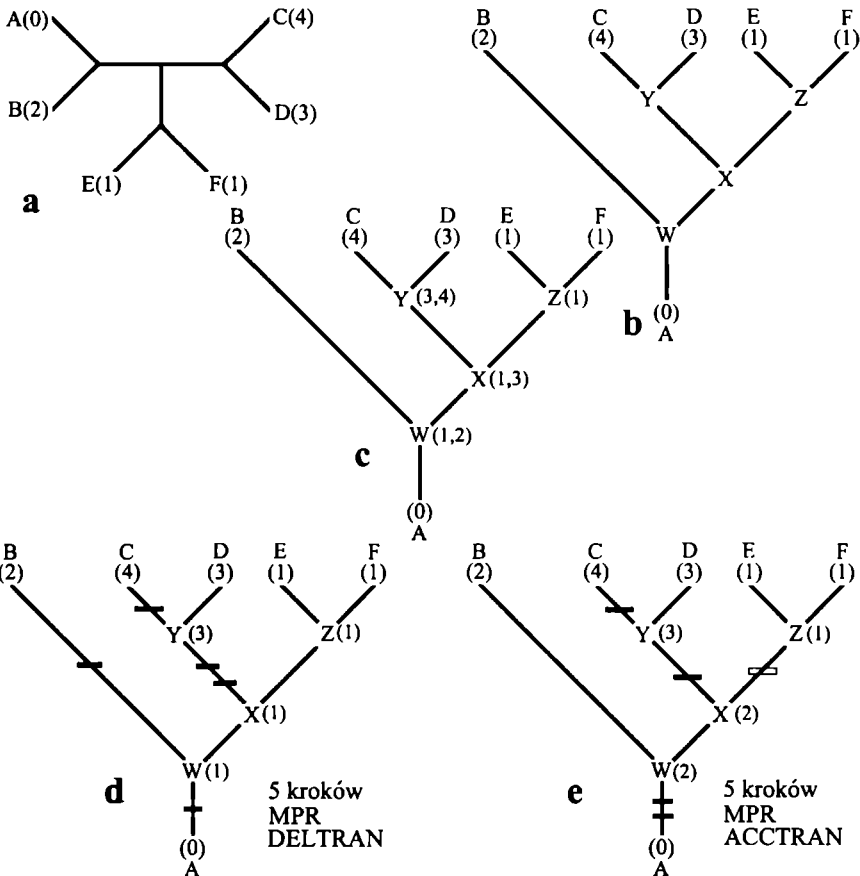
Optymalizacja zakładająca, że cechy są odwracalne i uporządkowane, nosi nazwę **optymalizacji Wagnera** (*Wagner parsimony*). Optymalizacja Wagnera nazywa się tak, bowiem koncepcja opiera się na pracy Wagnera (1961), choć sformalizowana została dopiero przez Kluge i Farris (1969) oraz Farris (1970). Dla cech odwracalnych nieuporządkowanych mówimy o **optymalizacji Fitcha** (*Fitch parsimony*; Fitch 1971). W obu przypadkach miejsce ukorzenia drzewa nie ma wpływu na wartość kryterium optymalizacji, czyli długość tego drzewa. Algorytmy obliczania długości drzewa dla tych sposobów optymalizacji przedstawiają Swofford i Maddison (1987), Maddison i Maddison (1992) oraz Swofford i inni (1996). Są one możliwe do użycia, gdy dysponuje się jedynie kartką papieru i ołówkiem, więc dla lepszego zrozumienia techniki przedstawimy je tu w uproszczonej formie, choć obecnie wszelkie obliczenia długości drzew wykonuje się za pomocą komputera i odpowiedniego programu. W obu przypadkach do obliczenia długości drzewa wystarczy jedno przejście od taksonów terminalnych w dół drzewa. Dla optymalizacji Wagnera algorytm jest następujący (Swofford i inni 1996):

- (1) Po ukorzeniu drzewa na jednym z taksonów terminalnych, wybranym dowolnie (jak wiemy, miejsce ukorzenia nie wpływa na długość drzewa przy optymalizacji Wagnera lub Fitcha) każdemu z węzłów terminalnych i , włącznie ze stanowiącym korzeń drzewa, przypisujemy stan cechy odpowiadający jej stanowi w macierzy danych i określony jako S_i ; długość drzewa przyjmujemy wstępnie za równą zeru;

(2) Gdy stany cech dla pary węzłów i i j (terminalnych lub wewnętrznych) dychotomicznego drzewa, rozchodzących się z wewnętrznego węzła k , są znane jako S_i i S_j , to stan cechy S_k dla węzła k określamy:

(2a) gdy $S_i \cap S_j \neq \emptyset$, czyli iloczyn zbiorów stanów cechy u taksonów i i j nie jest zbiorem pustym, to $S_k = S_i \cap S_j$, a więc stan cechy w węźle k odpowiada iloczynowi, czyli wspólnej części zbiorów stanów w węzłach i i j ; możemy ten stan określić obustronnie zamkniętym przedziałem $\langle a_k, b_k \rangle$;

(2b) gdy $S_i \cap S_j = \emptyset$, czyli iloczyn zbiorów stanów cechy u taksonów i i j jest zbiorem pustym, to stan cechy w węźle k odpowiada najmniejszemu obustronnie zamkniętemu przedziałowi $\langle a_k, b_k \rangle$, zawierającemu elementy z każdego z dwóch zbiorów S_i i S_j ; wówczas powiększamy długość drzewa o wartość $b_k - a_k$;



Ryc. 4.7. Optymalizacja Wagnera (a, b, c) i rekonstrukcja stanów cech, w tym rodzaju optymalizacji (d, e). Bliższe objaśnienia w tekście

- (3) Jeżeli węzeł k jest bezpośrednim potomkiem terminalnego węzła (taksonu) stanowiącego korzeń, to przejście w dół jest zakończone, należy wówczas przejść do kroku (4); jeżeli nie – należy wrócić do kroku (2);
- (4) Jeżeli stan cechy w węźle terminalnym stanowiącym korzeń drzewa (x_r) nie należy do zbioru stanów określonych dla węzła będącego jego bezpośrednim potomkiem (S_k), to należy powiększyć długość drzewa o $a_k - x_r$, gdy $x_r < a_k$, bądź $x_r - a_k$, gdy $x_r > b_k$.

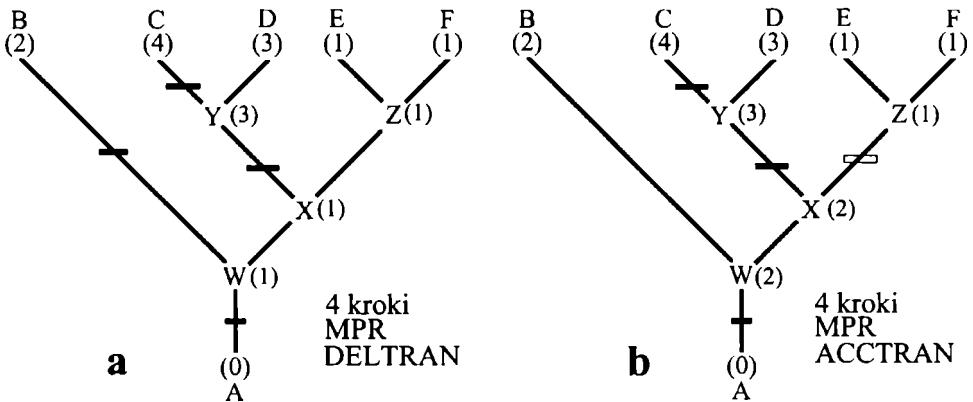
Nieukorzenione, sześciotaksonowe drzewo (Ryc. 4.7a) ukorzeniamy więc na taksonie A (Ryc. 4.7b) i w opisany wyżej sposób wyznaczamy przedziały możliwych stanów cechy dla wewnętrznych węzłów W, X, Y, Z (Ryc. 4.7c). Stany cech dla taksonów C i D (3, 4), przypisane węzłowi Y, wykluczają się, więc do zerowej na razie długości drzewa dodajemy: $4 - 3 = 1$. W węźle Z stan cechy to 1, tak jak u taksonów E i F, długość drzewa pozostaje bez zmian. W węźle X rekonstruujemy przedział wartości cechy jako $\langle 1, 3 \rangle$, jako że zbiór wspólnych wartości dla 1 i $\langle 3, 4 \rangle$ jest zbiorem pustym, a więc długość drzewa wzrasta o $3 - 1 = 2$. Wartości cechy dla węzła A (2) i X $\langle 1, 3 \rangle$ mają wspólny przedział $\langle 1, 2 \rangle$, który przyjmujemy dla węzła W i długość drzewa się nie zmienia. Wreszcie (krok 4) $2 - 0 = 2$. W sumie więc $1 + 2 + 2 = 5$ i tyle właśnie kroków liczy to drzewo dla tej cechy (Ryc. 4.7d, e). Powyższa procedura pozwala na obliczenie najmniejszej długości danego drzewa dla danej cechy, ale nie wystarcza do znalezienia najlepszej rekonstrukcji (MPR) stanów cech w węzłach wewnętrznych. Aby ją znaleźć, konieczne jest powtórne przejście drzewa: tym razem w górę, od korzenia po pozostałe taksony terminalne:

- (5) Przechodzimy do wewnętrznego węzła k , dla którego optymalna rekonstrukcja stanu cechy x_k nie została jeszcze znaleziona, ale jest znana (x_m) dla bezpośredniego przodka tego węzła, oznaczonego m ; gdy krok ten wykonywany jest po raz pierwszy, dotyczy pierwszego od dołu węzła wewnętrznego, a $m = r$, czyli bezpośrednim przodkiem jest takson terminalny, na którym ukorzeniono drzewo;
- (6) Z wyznaczonego wcześniej przedziału wartości S_k (równego $\langle a_k, b_k \rangle$) stanu cechy dla węzła k wyznaczamy wartość najbliższą x_m : gdy x_m zawiera się w S_k , to przyjmujemy $x_k = x_m$; gdy nie, to $x_k = a_k$ dla $x_m < a_k$ lub $x_k = b_k$ dla $x_m > b_k$;
- (7) Powtarzamy kroki (5) i (6), aż zrekonstruujemy stany cech dla wszystkich węzłów wewnętrznych.

W ten sposób zrekonstruowano stany cech drzewa na Ryc. 4.7d. Algorytm pozwala wyznaczyć MPR, lecz nie musi to być – i często nie jest – jedyny MPR dla danej cechy i danego drzewa: tak samo 5 kroków liczy inny możliwy MPR (Ryc. 4.7e). Swofford i Maddison (1987) oraz Maddison i Maddison (1992) opisują algorytm umożliwiający znalezienie wszystkich MPR. Porównanie rekonstrukcji wskazuje, że tak samo optymalne jest późniejsze przejście stanu 1 w stan 2 (Ryc. 4.7d: *delay transition* – DELTRAN), jak wcześniejsze przejście między tymi stanami (Ryc. 4.7e: *accelerate transition* – ACCTAN). Oczywiście rekonstrukcji może być więcej i wtedy ACCTAN i DELTRAN to przypadki skrajne zestawu możliwych rekonstrukcji. Choć formalnie tak samo dobre, odzwierciedlają różną rzeczywistość ewolucyjną: ACCTAN maksymalizuje liczbę odwróceń (rewersali), a DELTRAN liczbę paralelizmów, jako postulowanych *ad hoc* homoplazji. Musimy więc mieć jakieś podstawy, wybierając jedną z tych opcji: łatwość nabywania określonego stanu cechy przemawia

za DELTRAN-em, łatwość utraty za ACCTTRAN-em. Optymalizacja Fitcha – odpowiednia dla nieuporządkowanych cech – przebiega podobnie, wymagając jedynie modyfikacji kroków (2a), (2b), (4) i (6):

- (2a') gdy $S_i \cap S_j \neq \emptyset$, a więc iloczyn zbiorów stanów cechy u taksonów i i j nie jest zbiorem pustym, to $S_k = S_i \cap S_j$, czyli stan cechy w węźle k odpowiada iloczynowi, to znaczy wspólnej części zbiorów stanów w węzłach i i j ;
- (2b') gdy $S_i \cap S_j = \emptyset$, czyli iloczyn zbiorów stanów cechy u taksonów i i j jest zbiorem pustym, to stan cechy w węźle k odpowiada sumie stanów przypisanych węzłom i i j ($S_i \cup S_j$); wtedy powiększamy długość drzewa o 1;
- (4') Jeżeli stan cechy w węźle terminalnym stanowiącym korzeń drzewa (x_i) nie należy do zbioru stanów określonych dla węzła będącego jego bezpośrednim potomkiem (S_k), to należy powiększyć długość drzewa o 1;
- (6') Gdy x_m należy do zestawu stanów cech S_k znalezionej dla węzła k , to stan ten przyjmujemy również dla węzła k ; jeżeli nie, to przyjmujemy dla tego węzła arbitralnie jakkolwiek stan z zestawu S_k .

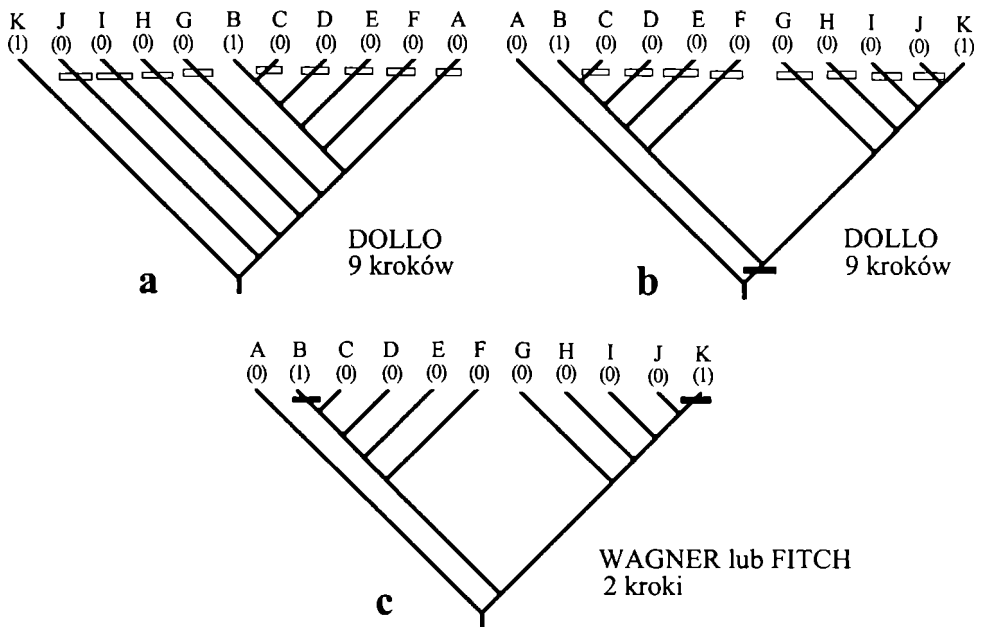


Ryc. 4.8. Rekonstrukcja stanów cech w optymalizacji Fitcha (a, b). Bliższe objaśnienia w tekście

I w tym przypadku może być więcej niż jeden MPR dla danej topologii drzewa (Ryc. 4.8a, b), algorytm obliczający wszystkie możliwe MPR przedstawił Fitch (1971). Dla cech typu Dollo mówimy o **optymalizacji Dollo** (*Dollo parsimony*; Farris 1977). Jak pamiętamy, w tym przypadku stan zaawansowany cechy nabyty może być tylko raz, choć tracony może być wielokrotnie; całość homoplazji jest więc następstwem kolejnych utrat stanu zaawansowanego. Z pozoru to założenie mało realistyczne, jednak nietrudno dostrzec jego słusność dla cech unikatowych, bardzo złożonych: łatwo uznać, że dwukrotne uzyskanie struktury morfologicznej o dużym stopniu złożoności, zbudowanej identycznie – przykładem oko kręgowca czy głowonoga – albo też niektóre złożone stany cech molekularnych – jak identyczne miejsca restrykcyjne DNA – jest niemożliwe, a przynajmniej niezmiernie mało prawdopodobne. Zarazem optymalizację Dollo stosuje się zwykle w złagodzonej postaci, zakładając bardzo wysoki koszt powtórnego uzyskania stanu zaawansowanego, niewykluczonego jednak zupełnie. Optymalizacja Dollo wymaga oczywiście znajomości polaryzacji stanów cechy, jest to

„polaryzacja uniwersalna”, lecz stany cech u przodka znane być nie muszą, a długość drzewa nie zależy od miejsca jego ukorzenia (Ryc. 4.9a, b).

Obliczanie długości drzewa za pomocą optymalizacji Dollo najlepiej rozpocząć od ukorzenia drzewa na taksonie, u którego występuje najbardziej zaawansowany stan cechy. Gdy stan cechy u dwóch taksonów pochodzących bezpośrednio od wewnętrznego węzła jest identyczny, to ten sam stan przypisujemy temu wewnętrznemu węzłowi i długość drzewa nie zmienia się. Gdy stany są różne, to wyższą (bardziej zaawansowaną) wartość przypisujemy wewnętrznemu węzłowi, a długość drzewa wzrasta o różnicę pomiędzy stanami cechy. Gdy wewnętrzny węzeł najbliższy taksonu terminalnego, na którym ukorzeniono drzewo, ma wartość cechy identyczną jak ten takson, długość drzewa nie wzrasta; gdy stany się różnią, długość drzewa wzrasta o tę różnicę. Zastosowanie optymalizacji Dollo wymaga rozwagi, bowiem użyta bezzasadnie może drastycznie podwyższyć wartość kryterium optymalizacji, czyli długość drzewa. W zilustrowanym przypadku (Ryc. 4.9b, c) optymalizacja Dollo wymaga dziewięciu kroków, podczas gdy optymalizacja Wagnera bądź Fitcha – jedynie dwóch.



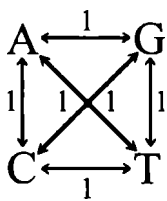
Ryc. 4.9. Długość drzewa w optymalizacji Dollo nie zależy od miejsca ukorzenia drzewa (a, b). Użycie tego rodzaju optymalizacji wymaga ostrożności, bowiem często wybór optymalizacji Dollo drastycznie zwiększa długość drzewa: na tym przykładzie optymalizacja Dollo wymaga 9 kroków (a, b), gdy Wagnera lub Fitcha – jedynie dwóch (c)

W przypadku DNA optymalizacja Fitcha (Ryc. 4.10a, b) niekoniecznie jest najodpowiedniejsza, bowiem szereg obserwacji potwierdza większą częstość tranzycji niż transwersji (Brown i inni 1982). Stąd postulowanie **optymalizacji transwersji** (*transversion parsimony*; Swofford i inni 1996). Teoretycznie adeninę i guaninę kodować można wspólnie jako puryny, a cytozynę i tyminę jako pirymidyny, co jest proste, jednak całkowite pominięcie tranzycji silnie ogranicza informację o zróżnicowaniu mię-

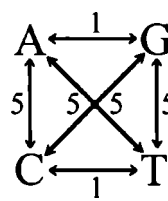
dzy bliskimi sobie taksonami. Właściwsze więc jest przypisanie większych wag transwersjom niż tranzycjom (Ryc. 4.10c, d), co można łatwo przeprowadzić w ramach tzw. **uogólnionej optymalizacji** (*generalized parsimony*). Im częściej zachodzi zmiana, tym większe jej prawdopodobieństwo, a zatem niższy koszt; im rzadziej – tym prawdopodobieństwo niższe, a koszt wyższy. Możemy więc tworzyć **macierz kosztów transformacji cech**, a nawet – jak o tym pisaliśmy w Rozdziale 2.3 – drzewo stanów cech, wskazujące możliwe przejścia. Obliczenia stają się skomplikowane i niemożliwe bez komputera, ale istniejące programy (choćby PHYLIP, PAUP, MACCLADE) w pełni je umożliwiają. Oczywiście znów warto przypomnieć, że komplikowanie modelu zwiększa wariancję rekonstrukcji, musi więc być rzeczywiście uzasadnione. Macierze kosztów dla różnych sposobów optymalizacji są następujące:

Camin-Sokal				Wagner				Fitch				Dollo				uogólniona				
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
1	-	1	2	3	-	1	2	3	-	1	1	1	-	<i>M</i>	<i>2M</i>	<i>3M</i>	-	1	3	7
2	∞	-	1	2	1	-	1	2	1	-	1	1	1	-	<i>M</i>	<i>2M</i>	5	-	2	2
3	∞	∞	-	1	2	1	-	1	1	1	-	1	2	1	-	<i>M</i>	7	4	-	1
4	∞	∞	∞	-	3	2	1	-	1	1	1	-	3	2	1	-	9	7	5	-

M to wielka liczba dodatnia. Macierz dla uogólnionej optymalizacji może być oczywiście bardzo różna. Teoretycznie wzdłuż przekątnej też możliwe są wartości różne od zera: można zaproponować model, w którym niepodleganie zmianom też coś kosztuje. Macierz dla optymalizacji transwersji jest symetryczna, bowiem zarówno tranzycje, jak i transwersje w obu kierunkach kosztują tyle samo, lecz koszt transwersji wynosi, powiedzmy, pięciokrotnie więcej niż tranzycji (Ryc. 4.10c, d). Kosztów przejść między określonymi stanami cechy nie należy mylić z wagą tej cechy, niekoniecznie równą 1, do czego wrócimy. Uogólniona optymalizacja jest wygodna i uniwersalna, o ile potrafimy sensownie określić koszt poszczególnych transformacji. Łatwo tu o zarzut arbitralności, jednak zakładanie wszystkich kosztów takich samych to także mocne i arbitralne założenie, nie jest więc bynajmniej „bardziej obiektywne”. Ogólną przesłanką jest przypisywanie kosztów tym większych, im rzadziej zmiana zachodzi, zaś tym mniejszych, im jest częstsza. Można obliczać koszt zmiany jako ujemny logarytm naturalny prawdopodobieństwa tej zmiany. Wrócimy do tego jeszcze.

**a**

	A	C	G	T
A	-	1	1	1
C	1	-	1	1
G	1	1	-	1
T	1	1	1	-

b**c**

	A	C	G	T
A	-	5	1	5
C	5	-	5	1
G	1	5	-	5
T	5	1	5	-

d

Ryc. 4.10. Optymalizacja Fitcha dla DNA, zakładająca taki sam koszt zamiany którejkolwiek z zasad na jakąkolwiek inną (a, b), niezbyt odpowiada rzeczywistości. Zwykle tranzycje zdarzają się znacznie częściej niż transwersje, choć możliwe tranzycje są zaledwie cztery, gdy transwersji osiem (c). Optymalizacja transwersji zakłada więc wyższy koszt transwersji (d), choć oczywiście wartość „pięć” dobrano tu arbitralnie

Dla sekwencji białek Eck i Dayhoff (1966) zaproponowali *maximum parsimony* uznającą każdy z 20 aminokwasów za stan cechy i zakładającą taki sam koszt przejścia między którymkolwiek dwoma z 20 stanów. Obliczane w ten sposób drzewa mają sensowne topologie (Russo i inni 1996), a technika jest obliczeniowo prosta. Takie podejście jest teoretycznie znacznym przybliżeniem, bowiem dla niektórych zmian aminokwasów wystarcza pojedyncza substytucja, gdy dla innych dwie bądź nawet trzy; ponadto niektóre pary aminokwasów są biochemicznie podobne, inne nie, a zmiany w obrębie podobnych zachodzą częściej. Dlatego też szereg badaczy (jak Moore i inni 1973, Fitch i Farris 1974, Sankoff i Rousseau 1975, Felsenstein 1988b) zaproponowało uwzględnianie minimalnej liczby podstawień w DNA dla danego przejścia przy określaniu jego kosztu. Złożoność algorytmów i mnogość założeń nie oznacza jednak, by te techniki dawały lepsze rekonstrukcje niż przybliżona i prosta metoda Ecka i Dayhoffa (1966). Zarazem złożoność pozornie prostych procesów ewolucyjnych sekwencji DNA i szybkie osiągnięcie saturacji powodują, że rekonstrukcje filogenezy oparte na sekwencjach białek często są bliższe rzeczywistym filogenezom niż te oparte na DNA.

Filogenetyczna informacja zawarta w genomach wyższych organizmów nie ogranicza się do tej, którą uzyskujemy z porównania, pozycja po pozycji, kolejnych nukleotydów w sekwencjach kwasów nukleinowych organizmów, których pokrewieństwa rekonstruujemy. Genomy zawierają szereg unikatowych, występujących u więcej niż jednego taksonu, zaawansowanych stanów cech, które wydają się nieodwracalne. Są to powtarzające się fragmenty, uzyskane w całości w którymś momencie ewolucji. Najlepiej poznane to **krótkie powtarzalne elementy** (*short interspersed repetitive elements* – SINEs) i **długie powtarzalne elementy** (*long interspersed repetitive elements* – LINEs; Singer 1982, Jurka i Smith 1988, Britten i inni 1988). Krótkie liczą 80–400 nukleotydów, długie – od kilkuset po kilka tysięcy, oba rodzaje to retropseudogeny, zdolne do replikacji; po replikacji zostają włączone w różne fragmenty genomu i pozostają tam na zawsze, jeżeli nie zostaną utracone na drodze rzadko zachodzących delecji wielkich fragmentów DNA (Nei i Kumar 2000). Oczywiście w obrębie tych elementów również zachodzą punktowe mutacje oraz drobne insercje/delecje, lecz gdy rozpatrujemy taksony dość bliskie sobie, które oddzieliły się nie wcześniej jak 50 mln lat temu, elementy obu rodzajów traktowane być mogą jako synapomorfie (Verneau i inni 1997, Nei i Kumar 2000). Homoplazji wykluczyć w tym przypadku nie można, choć jest bardzo mało prawdopodobna; potencjalnym problemem jest ewentualny ancestralny polimorfizm: występowanie tych elementów nie we wszystkich liniach gatunku wyjściowego (Nei i Kumar 2000). Podobnie użyteczne są inne markery genetyczne, jak wielkoskalowe insercje czy delecje, albo obecność lub brak intronów w obrębie genów kodujących białka – dla tych ostatnich odpowiednia wydaje się optymalizacja Dollo, bowiem dwukrotne pojawienie się intronu w tym samym miejscu jest niezwykle mało prawdopodobne, gdy utrata zawsze może się zdarzyć.

Miary dobroci drzewa i rekonstrukcji ewolucji cechy

Jak wiemy, gdyby wśród stanów cech występujących u taksonów terminalnych nie było jakichkolwiek homoplazji, rekonstrukcje zarówno drzewa, jak i ewolucji cech byłyby proste i pewne. W praktyce jednak zawsze występują homoplazje, filogenetyczny sygnał bywa słaby i jest możliwa więcej niż jedna topologia oraz rekonstrukcja ewolucji stanów cech. Ponadto im więcej homoplazji, tym mniej wiarygodne są rekon-

strukcje techniką kładystyczną. Warto więc jakoś oceniać wiarygodność zarówno topologii, jak i rekonstrukcji ewolucji cechy. Jak wiemy, w metodzie kładystycznej minimalizowanym kryterium jest długość drzewa, wyrażona liczbą kroków. Dla cechy uporządkowanej przejście 1→3 to oczywiście dwa kroki, ale – zależnie od rekonstrukcji ewolucji cech – jedna lub dwie **zmiany** (*changes*), zależnie od tego, czy na drzewie przejście nastąpiło jednokrotnie, czy też poprzez stan 2. Możemy więc mówić o **maksymalnej liczbie zmian** i **minimalnej liczbie zmian** dla danego drzewa (i paru więcej parametrach opisujących liczbę zmian – patrz Maddison i Maddison 1992). Dla cechy *i* minimalną możliwą liczbę kroków na drzewie o jakiegokolwiek możliwej topologii oznaczmy m_i , maksymalną możliwą liczbę kroków na drzewie o jakiegokolwiek możliwej topologii oznaczmy M_i , a minimalną możliwą liczbę kroków na rozważanym drzewie oznaczmy s_i . Wówczas dla rekonstrukcji cechy określamy **współczynnik spójności rekonstrukcji cechy** (*character consistency index*: Kluge i Farris 1969, Farris 1989):

$$CI_{cechy} = \frac{m_i}{s_i}, \text{ a współczynnik retencji (character retention index): } RI_{cechy} = \frac{M_i - m_i}{M_i - s_i}$$

(Archie 1989, Farris 1989). Definiuje się też **reskalowany współczynnik spójności** (*character rescaled consistency index*): $RC_{cechy} = CI_{cechy} * RI_{cechy}$ (Farris 1989). Jeżeli cecha nie wykazuje zmienności w obrębie drzewa, mianownik CI jest równy 0, więc wówczas współczynnik nie istnieje, bądź jego wartość arbitralnie przyjmowana jest za równą 0.

Analogiczne ogólne współczynniki: spójności (CI), retencji (RI) i reskalowany spójności (RC) zdefiniować możemy też dla całego drzewa (*ensemble* albo *overall consistency, retention and rescaled consistency indices*), uwzględniając wszystkie *n* cech, dla każdej z nich określając wartość wagi w_i :

$$CI_{drzewa} = \frac{\sum_{i=1}^n w_i m_i}{\sum_{i=1}^n w_i s_i}, \quad RI_{drzewa} = \frac{\sum_{i=1}^n w_i M_i - \sum_{i=1}^n w_i s_i}{\sum_{i=1}^n w_i M_i - \sum_{i=1}^n w_i m_i}, \quad RC_{drzewa} = CI_{drzewa} * RI_{drzewa}$$

Gdy brak jakichkolwiek homoplazji, $CI = 1$; im więcej homoplazji, tym niższa wartość CI, choć nie może osiągnąć zera, nawet gdy sygnału filogenetycznego brak. Współczynnik spójności CI ma i inne wady. Dla cech kładystycznie pozbawionych informacji również przybiera wartość 1, toteż gdy cech tych nie wyłączymy w obliczaniu współczynnika, to jego wartość zostanie zawyżona i może być całkiem wysoka, mimo że udział homoplazji jest duży. Ponadto ze wzrostem liczby taksonów wartość CI maleje, najczęściej również wtedy, gdy udział homoplazji nie wzrasta (Kitching i inni 1998). Dla cech kładystycznie pozbawionych informacji współczynniki RI i RC mają w mianowniku 0, a więc są nieokreślone: cechy takie trzeba więc pominąć w obliczeniach. Współczynnik retencji RI osiąga wartości w przedziale $<0, 1>$ i odzwierciedla udział synapomorfii w danych. Pozwala uniknąć wad CI, podobne własności ma reskalowany współczynnik spójności RC. Niekiedy definiuje się też **współczynnik homoplazji** (*homoplasy index*): $HI = 1 - CI$.

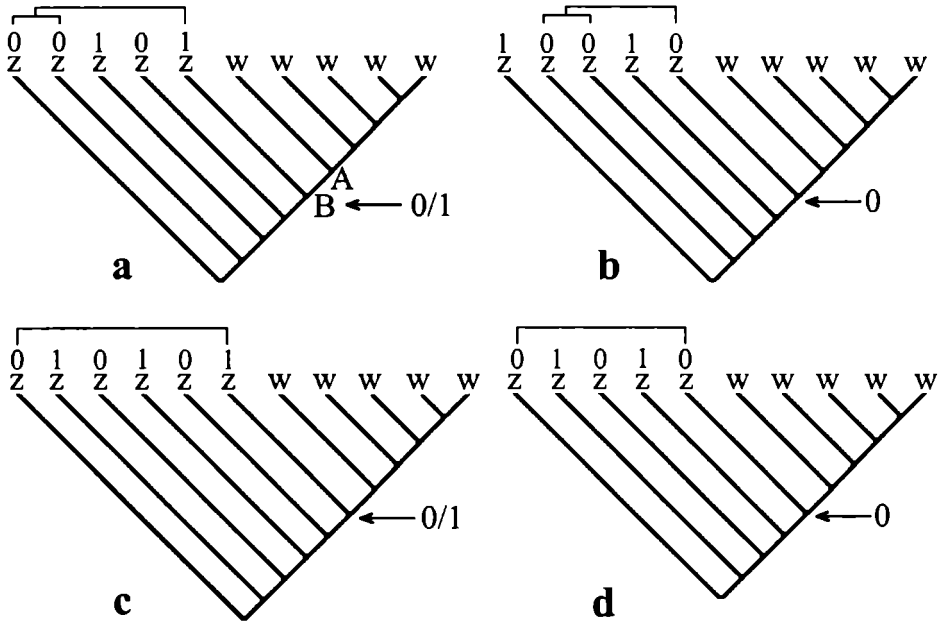
Polaryzacja cech i ukorzenianie drzewa w metodzie kladystycznej

Jak pamiętamy, Hennigiańska metoda rekonstrukcji filogenezy wymaga znajomości *a priori* polaryzacji cech, natomiast technika kladystyczna pozwala na rekonstrukcję bez takiej wiedzy, jednak w końcu drzewo należy ukorzenić i wskazać polaryzację cech. Jak wspominaliśmy, znajomość rozwoju ontogenetycznego pozwala na określenie polaryzacji cech, jako że stany zaawansowane pojawiają się w rozwoju później, lecz nie jest tak, gdy ma miejsce pedomorfoza, co bynajmniej nie zdarza się rzadko. Zagadnienie omawiają choćby Wiley (1981) czy Kitching i inni (1998). Tu zajmujemy się krótko określaniem polaryzacji za pomocą grupy zewnętrznej. Najprościej ujmując, jeżeli cecha ma dwa lub więcej stanów w obrębie grupy zewnętrznej, to stan obecny u grupy zewnętrznej jest stanem plezjomorficznym dla grupy wewnętrznej. Watrous i Wheeler (1981) zaproponowali zestaw reguł, użytecznych w tym przypadku. Wprowadzili oni metodę **funkcjonalnej grupy wewnętrznej/funkcjonalnej grupy zewnętrznej** (*functional ingroup/functional outgroup* – FIG/FOG). Do grupy taksonów o nieokreślonych pokrewieństwach (drzewo w postaci jednej politomii) stanowiących grupę wewnętrzną dołączamy takson (taksony) grupy zewnętrznej. Wybieramy jedną z cech i na jej podstawie częściowo rekonstruujemy pokrewieństwa w obrębie grupy wewnętrznej: z politomii wyodrębnia się co najmniej jedno dychotomiczne rozgałęzienie. Takson najbliższy grupie zewnętrznej uznajemy teraz za funkcjonalną grupę zewnętrzną (FOG), resztę grupy wewnętrznej za funkcjonalną grupę wewnętrzną (FIG). Wybieramy następną cechę, znów przekształcając część politomii w dychotomie, takson na drzewie najbliższy FOG uznajemy teraz za nowy FOG, a resztę grupy wewnętrznej za nowy FIG, i tak aż do pełnej rekonstrukcji filogenezy w obrębie grupy wewnętrznej.

Procedura FIG/FOG jest odpowiednia, gdy stany cech nie wykazują zmienności w obrębie grupy zewnętrznej. Jeżeli są tam zmienne, a zwłaszcza gdy wśród ich stanów są występujące również w grupie wewnętrznej, technika FIG/FOG zawodzi (Maddison i inni 1984). Wskazanie stanu w węźle stanowiącym wspólnego przodka badanej (wewnętrznej) grupy (węzeł A na Ryc. 4.11a) może prowadzić do jedynie „lokalnie” poprawnej rekonstrukcji; lepiej za pierwotny uznać stan w węźle obok, należącym już do grupy zewnętrznej (węzeł B na Ryc. 4.11a). Ten stan możemy zrekonstruować w sposób jednoznaczny bądź nie, zależnie od rozkładu stanu cechy w obrębie grupy zewnętrznej (Ryc. 4.11a–d). Ogólnie rekonstrukcja przebiega podobnie jak opisana dla optymalizacji Wagnera.

Sformułowano też dwie proste reguły określania stanu plezjomorficznego (w węźle B): **pierwszego dubletu** (*first doublet*) i **przemiennej grupy zewnętrznej** (*alternating outgroup*). Obie wymagają znajomości filogenezy w obrębie grupy zewnętrznej, co bywa problematyczne. Reguła pierwszego dubletu głosi, że jeżeli określony stan cechy występuje u jakichś dwóch najbliższych spokrewnionych taksonów (leżących obok siebie na drzewie: to właśnie dublet) grupy zewnętrznej i u taksonu z grupy zewnętrznej najbliższego grupie wewnętrznej, to stan cechy w węźle B możemy określić jednoznacznie i jest to właśnie ten stan co u dubletu (Ryc. 4.11b). Jeżeli natomiast stan w obrębie dubletu jest inny niż u taksonu najbliższego grupy wewnętrznej, to stan cechy w węźle B nie może być jednoznacznie określony (Ryc. 4.11a). Reguła przemiennej grupy zewnętrznej zestawia stany cech w obrębie grupy zewnętrznej, pomiędzy taksonem najbliższym a najdalszym grupy zewnętrznej, gdy dubletu wyróżnić się nie da: gdy są identyczne, taki sam stan jednoznacznie można przypisać węźlowi B (Ryc. 4.11d); gdy

są różne, stan cechy w węźle B nie może być jednoznacznie określony (Ryc. 4.11c). Oczywiście dołączenie kolejnego, najodleglejszego taksonu do grupy zewnętrznej może zmienić wynik rekonstrukcji polaryzacji za pomocą drugiej z reguł, choć w praktyce cechę, której stany występują przemiennie w obrębie drzewa, najczęściej eliminuje się z analizy – jako niosącą niewiele informacji filogenetycznej.



Ryc. 4.11. Znajdowanie polaryzacji cech procedurą FIG/FOG. Stan cechy binarnej w węźle A (a), czyli wspólnego przodka badanej grupy wewnętrznej (W), może nie być globalnie ancestralny, bezpieczniej przyjąć za stan ancestralny stan cechy w węźle B (a), czyli u przodka taksonu grupy zewnętrznej (Z) najbliższego wewnętrznej. Konieczna jest znajomość wzajemnych pokrewieństw taksonów grupy zewnętrznej. Reguła dubletu (a, b) pozwala jednoznacznie określić stan cechy w punkcie B wówczas, gdy ten sam stan cechy występuje u dubletu i u taksonu grupy zewnętrznej najbliższego wewnętrznej (b), inaczej stan cechy jest nieokreślony (a). Reguła przemiennie grupy zewnętrznej (c, d), stosowana gdy nie można wyróżnić dubletu, pozwala określić stan cechy w punkcie B, gdy stany cech są identyczne u taksonów grupy zewnętrznej najbliższego i najdalszego grupie wewnętrznej (d)

W literaturze spotykamy szereg opinii o kluczowym znaczeniu właściwego doboru grupy zewnętrznej, najlepiej będącej kładem siostrzanym grupy badanej. Wydają się one przesadzone: wykorzystanie taksonów nieco odleglejszych – lub całej gamy taksonów bliższych i dalszych – powinno wystarczyć. Ponadto czas nie stał w miejscu także dla kładu siostrzanego, toteż często przedstawia on zestaw zaawansowanych stanów cech (autapomorfii), a więc może być słabo użyteczny dla ukorzeniania drzewa czy określania polaryzacji cech.

Ważenie cech i kosztów transformacji

Jak już mówiliśmy, przyjęcie dla wszystkich cech takich samych wag, choć pozornie obiektywne – jako pozbawione potencjalnie subiektywnych i nieuzasadnionych założeń – nie jest właściwe, wprowadzając nierealistyczne założenie o takim samym ewolucyjnym znaczeniu wszystkich cech. Wag cech nie należy mylić z kosztami transformacji, które mogą być zróżnicowane pomiędzy kolejnymi przejściami między stanami tej samej cechy. Zarówno ważenia cech, jak i określenia kosztów transformacji dokonuje się zarówno *a priori* – przed rozpoczęciem rekonstrukcji i niezależnie od jej wyniku, jak i *a posteriori* – czyli na podstawie przynajmniej wstępnej rekonstrukcji filogenezy. Techniki *a priori* określane bywają też jako **niezależne od hipotezy** (*hypothesis independent*) bądź **niezależne od obliczonego drzewa** (*tree independent*), a metody *a posteriori* jako **zależne od hipotezy/drzewa** (*hypothesis/tree dependent*).

Techniki *a priori* opierają się na wiedzy o cechach, którą mamy jeszcze przed rozpoczęciem rekonstrukcji filogenezy. Wiadomo, że cechy łatwo ulegające adaptatywnym modyfikacjom powinny ważyć mniej. Obok cech odzwierciedlających wyłącznie sygnał filogenetyczny, a więc najbardziej użytecznych dla rekonstrukcji, spotykamy cechy niosące niewiele informacji, ale także i cechy wprowadzające w błąd – potrzeba różnicowania wag jest więc oczywista. W postaci skrajnej niektóre cechy zwyczajnie eliminujemy z analizy. Także eliminacja bądź przypisanie niskich wag cechom, których zachowanie nie spełnia warunków użycia techniki, prowadzi do redukcji błędu systematycznego rekonstrukcji. Technika kladystyczna działa najlepiej, gdy zmian na drzewie jest niewiele, a więc hipotez *ad hoc* o homoplazjach musimy tworzyć najmniej – stąd cechom bardzo zmiennym należy przypisać wagi najniższe. Techniki *a priori* wykorzystują **analizę cech** (*character analysis*) i **analizę kompatybilności cech** (*character compatibility analysis*).

Przed rozpoczęciem rekonstrukcji filogenezy warto raz jeszcze sprawdzić stany poszczególnych cech, czy jakieś dane nie są błędne. Następnie należy starannie sprawdzić raz jeszcze homologie (Neff 1986). Gdy któraś z cech wykazuje inne zróżnicowanie stanów niż pozostałe cechy, warto jej szczególnie uważnie się przyjrzeć. Dla danych morfologicznych sporo rozumiemy i taka analiza zwykle przynosi sensowne rezultaty; gorzej z danymi molekularnymi. Hillis i inni (1993) zestawili różne schematy ważenia *a priori* dla sekwencji DNA: identyczne wagi dla wszystkich pozycji, wagi zróżnicowane między pozycjami (zróżnicowanie wag dla pierwszej, drugiej i trzeciej pozycji kodonu, zróżnicowanie zależne od pozycji w drugorzędowej strukturze RNA) oraz wagi zróżnicowane w obrębie pozycji (czyli różne koszty zmian stanów cechy: transwersje ważone inaczej niż tranzycje, 12 możliwych substytucji ważonych zależnie od ich spodziewanych bądź obserwowanych częstości, w kodujących białka fragmentach sekwencji zmiany kodonów na synonimiczne ważone inaczej niż na niesynonimiczne: 70% substytucji na trzecich pozycjach kodonów jest synonimiczna, gdy wszystkie substytucje na drugiej pozycji i 96% na pierwszej nie są synonimiczne). Zważywszy, że w obrębie 61 kodonów kodujących białka różnych substytucji może być w sumie 549, obliczenia częstości różnych substytucji są długie, a jakkolwiek analiza bez komputera zupełnie niemożliwa. Możliwych tranzycji jest cztery, gdy transwersji osiem, lecz – jak wiemy – w rzeczywistych sekwencjach tranzycje zwykle przeważają nad transwersjami, szybko osiągając saturację.

W przypadku gdy pewien zestaw cech zgodnie wskazuje na określony obraz ewolucji, a inne nie, warto skorzystać z analizy kompatybilności cech (Felsenstein 1981b), pozwalającej na wskazanie największego **zestawu cech kompatybilnych** (*clique*), czyli takich, które wskazują na taką samą rekonstrukcję filogenezy, zupełnie pozbawioną homoplazji (LeQuesne 1982, Estabrook 1983). Oparte na kompatybilności metody rekonstrukcji filogenezy szybko zostały zarzucone, jako oparte na nierealistycznym założeniu, że cecha raz odrzucona z największego możliwego zestawu cech kompatybilnych (*largest clique*) może zostać uznana za nieniosącą jakiegokolwiek użytecznej informacji. Z drugiej strony, technika pozostaje użyteczna dla ważenia cech (Penny i Hendy 1985a, 1986; Sharkey 1989 opisał podobną metodę dla cech binarnych, a Estabrook i Landrum 1975, Sneath i inni 1975 oraz Fitch 1977 zajmowali się badaniem kompatybilności cech nieuporządkowanych wielostanowych, technikę ponownie rozważał Sharkey 1993). Penny i Hendy (1985a, 1986) zaproponowali obliczanie liczby E_j niekompatybilności (par cecha z cechą), spodziewanej losowo w przypadku braku jakiegokolwiek – wynikającej ze wspólnej ewolucyjnej historii – zależności między rozkładami stanów różnych cech, jak też zliczanie przypadków niekompatybilności między cechami O_j , obserwowanych w macierzy danych. Wówczas:

$$w_j = \max \left[1 - \left(\frac{O_j}{E_j} \right), 0 \right],$$

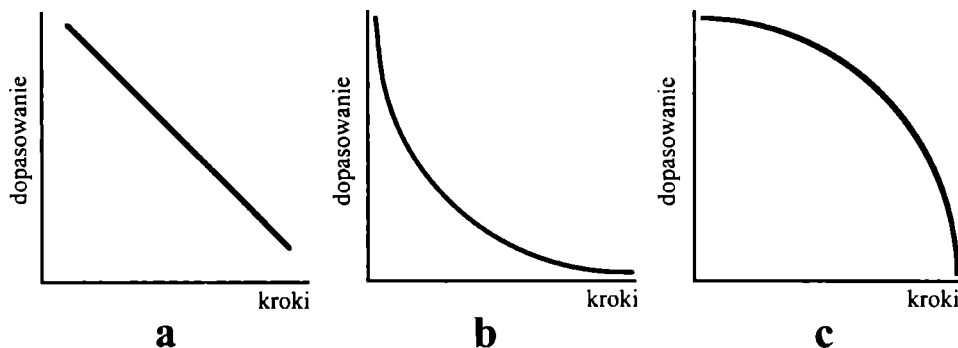
choć funkcje ważące mogą być różne. Dla powyżej zdefiniowanej cecha kompatybilna z wszystkimi pozostałymi otrzyma wagę 1, gdy cecha kompatybilna z liczbą spodziewaną losowo będzie miała wagę 0, a dla mniejszych niż E_j liczb par niekompatybilnych – wartości będą należały do przedziału (0, 1). Teoretycznie dla wartości wyższych niż E_j wagi mogą być ujemne, choć interpretacja takich wag jest niejasna i raczej przyjmuje się je wówczas za równe 0. Analiza kompatybilności wymaga badań – najlepiej techniką symulacji – bowiem jej zachowanie nie zostało dokładnie sprawdzone.

Kolej na omówienie technik ważenia *post hoc*. Farris (1969b, 1988; Kluge i Farris 1969) i Carpenter (1988) zaproponowali technikę ważenia cech *post hoc*, znaną jako **analiza wag cech metodą kolejnych przybliżeń** (*successive approximations character weighting analysis* – SACW). Istota techniki polega na wstępnym przyjęciu wag, które wydają się odpowiednie – może to być po prostu założenie takiej samej wagi, równej 1, dla wszystkich cech. Następnie znajdujemy drzewo i na jego podstawie obliczamy nowe wagi. Jako wag użyć można któregoś z omówionych wcześniej współczynników: CI, RI lub RC, odpowiednio skalowanych (wagi proporcjonalne do wielkości współczynnika). Wagi można też obliczać na podstawie liczby zmian stanów cechy, policzonych na drzewie: może to być odwrotność liczby zmian/kroków bądź też kwadrat odwrotności (Williams i Fitch 1989, 1990, Maddison i Maddison 1992). Oczywiście kwadrat odwrotności waży cechę o większej liczbie zmian znacznie niżej niż zwyczajna odwrotność. Dla nowych wag obliczamy nowe drzewo, na jego podstawie nowe wagi, wykorzystując je – następne drzewo i tak aż do czasu, gdy kolejne drzewo lub drzewa nie różnią się od obliczonych z wykorzystaniem poprzednich wag.

Procedura jest, rzecz jasna, obliczeniowo intensywna. Dodatkowym problemem jest to, że wolno ewoluujące cechy – np. wolno ewoluujące fragmenty sekwencji DNA – użyteczne są jedynie przy rekonstrukcji pokrewieństw między odległymi taksonami,

gdy szybko ewoluujące – między bliskimi, natomiast w praktyce filogenezę rekonstruuje się zwykle dla grupy taksonów różnie odległych: bliższych i dalszych. Wówczas trudno mówić o wagach odpowiednich dla całego drzewa: chcąc nie chcąc rekonstrukcja będzie obciążona w kierunku odleglejszych bądź bliższych pokrewieństw, zależnie od tego, jakie relacje filogenetyczne są częstsze pomiędzy badanymi taksonami, i będzie tak również przy wykorzystaniu technik maksymalizujących wiarygodność. Teoretyczne uzasadnienie SACW nie jest pełne, a zachowanie techniki nie do końca jasne. Intuicyjnie oczywiste wydaje się, że należy jakoś kompensować wielką liczbę zmian jednej cechy (odpowiednio niżej ważonej), gdy druga zmienia się, powiedzmy, raz w obrębie całego drzewa (powinna więc ważyć więcej).

Trudno ocenić, który z proponowanych współczynników jest najlepszy – może warto spróbować wykorzystać różne. Bywa, że kolejne rekonstrukcje nie różnią się zbyt od wcześniejszych i lepsze wyniki uzyskamy, ważąc cechy bardziej zdecydowanie (np. odwrotnością całkowitej liczby zmian podniesioną do potęgi 10: Swofford i inni 1996). Kolejnym problemem jest to, że zwykle drzew MPR mamy więcej niż jedno, więc wartości współczynników są różne i brak kryterium, których z nich (najniższe, najwyższe, średnie?) użyć. Korzystne byłoby, gdyby zastosowanie SACW prowadziło do zmniejszenia liczby kladogramów MPR. Tak bywa, lecz bywa i odwrotnie, gdy liczba MPR rośnie drastycznie. Wyniki SACW zależą też od wyjściowego drzewa, więc cecha nisko ważona na tym drzewie utrzyma stosunkowo niską wagę i w kolejnych rekonstrukcjach. Zresztą, jak to już podkreślaliśmy i jak też zauważył D.R. Maddison (1990), brak obiektywnych kryteriów porównania dwóch drzew, a więc trudno oceniać, w jakim stopniu technika poprawia rekonstrukcję filogenezy: pozostają symulacje, lecz trudno w nich stworzyć realistyczne dane, niebędące np. połączeniem „dobrych” cech z „losowym szumem” – co jest proste, lecz nie odpowiada rzeczywistości. Ważenie cech często prowadzi do zastąpienia wyjściowo bogatego zestawu danych – w następstwie przypisania większości cech wag zero lub bliskich zero – kilku zaledwie cechami o potencjalnie wątpliwej wartości. Simon i inni (1994) dokładniej omawiają różne aspekty ważenia cech.



Ryc. 4.12. Im mniej homoplazji, tym bardziej dane pasują do kladogramu. Na osi odciętych przedstawiono liczbę kroków policzonych na drzewie, a na osi rzędnych stopień dopasowania danych do tego drzewa: wykres funkcji dopasowującej może być liniowy (a), wklęsły (b) lub wypukły (c)

Technikę ważenia cech opartą na ocenie homoplazji zaproponował Goloboff (1993). Dane tym bardziej pasują do kladogramu, im mniej jest wśród nich homoplazji. Jeżeli na osi odciętych umieścimy liczbę kroków policzonych na drzewie, a na osi rzędnych stopień dopasowania danych do tego drzewa (Farris 1969b), to wykres funkcji dopasowującej może być liniowy (Ryc. 4.12a), wklęsły (Ryc. 4.12b) lub wypukły (Ryc. 4.12c). Wykres liniowy (Ryc. 4.12a) przedstawia sytuację, gdy wszystkim cechom przypisano takie same wagi – niejako przypadek wyjściowy. Krzywa wklęsła odpowiada sytuacji, gdy cechom wykazującym najmniej homoplazji przypisano wagi najwyższe. Wypukły (Ryc. 4.12c) przedstawia sytuację przeciwną – wagi cech proporcjonalne do stopnia homoplazji tych cech – a więc ważenie pozbawione ewolucyjnego sensu. Goloboff (1993) wykorzystał krzywą wklęsłą do ważenia cech w swoim programie PIWE (Goloboff 1996b). Wagi obliczane są zgodnie z formułą: $W = K/(K + Esi)$, gdzie Esi to liczba dodatkowych kroków dla danej cechy, a K to stała wklęsłości funkcji. Technika wymaga weryfikacji.

Podobnie ważyć można *post hoc* koszty kolejnych transformacji między stanami określonej cechy: Sankoff i Cedergren (1983) oraz Williams i Fitch (1990) zaproponowali **dynamiczne ważenie** (*dynamical weighting*: nazwy tej używa się też niekiedy dla SACW) kosztów transformacji, choć technika służyć też może ważeniu samych cech; wówczas podobnie działa jak SACW i budzi podobne zastrzeżenia, choć Fitch i Ye (1991) wykazali symulacjami komputerowymi, że procedura ta – przy spełnieniu szeregu założeń – zwiększa prawdopodobieństwo znalezienia prawidłowej topologii. Technika budzi wątpliwości, zwłaszcza związane z ryzykiem logiki błędnego koła, a brak teoretycznych przesłanek gwarantujących konwergencję obliczanych drzew z drzewem odzwierciedlającym rzeczywistą filogenezę. Nasza wiedza nie pozwala wciąż jednoznacznie zalecić stosowanie lub nie dynamicznego ważenia. Ponadto w praktyce metoda ta często nieznacznie zmienia początkowe rekonstrukcje lub nawet zupełnie ich nie modyfikuje. Zasada ważenia jest oczywista: im częstsza zmiana, a więc większe prawdopodobieństwo, czyli łatwość zmiany, tym niższa waga tej zmiany. Oznaczmy koszt przejścia ze stanu i do stanu j jako K_{ij} , liczbę przejść $i \rightarrow j$ jako X_{ij} , liczbę przejść ze stanu i do wszystkich stanów cechy jako X_i , a liczbę zmian na całym drzewie jako X . Wówczas koszty przejścia obliczać możemy jako: $K_{ij} = -\ln(X_{ij}/X_i)$ lub $K_{ij} = -\ln(X_{ij}/X)$ (Wheeler 1990, Maddison i Maddison 1992), albo też $K_{ij} = 1/X_{ij}$ lub $K_{ij} = 1/(X_{ij})^2$ (Williams i Fitch 1989, 1990, Maddison i Maddison 1992). Znowu trudno ocenić, która z formuł jest najwłaściwsza, znowu też – mając więcej niż jeden MPR – nie wiadomo, których wartości użyć: najniższych, średnich czy najwyższych. Jak w SACW, koszty przejścia wykorzystujemy przy konstrukcji następnego drzewa (drzew), na których podstawie obliczamy nowe koszty i tak dalej, aż kolejna rekonstrukcja nie różni się od poprzedniej.

Teoretycznie najprecyzyjniejsze estymaty prawdopodobieństw poszczególnych transformacji obliczyć można techniką maksymalizacji wiarygodności, którą zajmujemy się w następnym rozdziale. Tak obliczone wagi wykorzystać możemy do analizy techniką redukcjonistyczną. Wprawdzie technika maksymalizacji wiarygodności pozwala też obliczyć drzewo, jednak jest bez porównania bardziej intensywna obliczeniowo niż *parsimony*, więc w praktyce analizuje zaledwie część możliwych topologii i długości gałęzi, bowiem już kilka taksonów stanowi tu limit możliwości pełnych obliczeń. Tymczasem metoda redukcjonistyczna jest szybsza, można więc wagi obliczone techniką maksymalizacji wiarygodności wykorzystać do znalezienia drzewa techniką re-

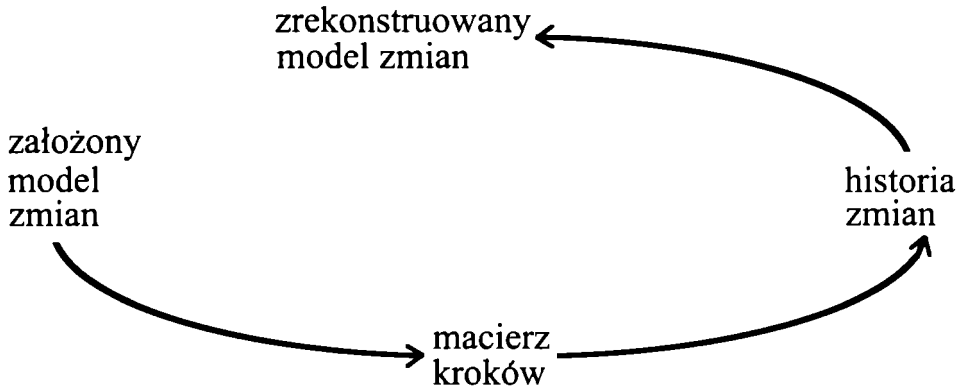
dukcyjności, po czym to drzewo optymalizować metodą maksymalizacji wiarygodności. W ten sposób stosunkowo szybko znaleźć możemy drzewa lepsze niż korzystając z nieważonej techniki redukcjonistycznej, a nawet techniki maksymalizacji wiarygodności, gdy taksonów jest więcej.

Rekonstrukcja ewolucji cech, dane wewnętrzne i zewnętrzne

Jak wynika choćby z wcześniejszego przedstawienia algorytmów kolejnych metod optymalizacji, technika kladystyczna w trakcie konstrukcji drzewa odtwarza – choć nie dla każdego węzła jednoznacznie – stany cech w kolejnych węzłach drzewa, włącznie z hipotetycznym przodkiem badanej grupy. Obok znajdowania drzewa odtwarzającego filogenezę, metoda wykorzystywana jest więc również do badania ewolucji cech. Choć pozornie proste, odtwarzanie ewolucji cech to cała dziedzina, szeroko omawiają ją choćby Maddison i Maddison (1992) oraz Maddison (1994). Głównym problemem, jak już wspominaliśmy, jest ryzyko wnioskowania zgodnie z regułą błędnego koła: rekonstrukcja ewolucji cech odbywa się na podstawie drzewa, znalezione na podstawie znajomości ewolucji tych samych cech (albo jedynie arbitralnych sądów na temat tej ewolucji). Niektórzy uważają więc (np. Coddington 1988, Brooks i McLennan 1991), że tych samych cech nie można użyć do rekonstrukcji filogenezy, a następnie badania ewolucji tych cech na podstawie tego drzewa. Teoretycznie można by obliczyć „pewne” drzewo na podstawie części cech, by następnie użyć go do śledzenia ewolucji pozostałych cech, jednak drzewo obliczone jedynie z części danych raczej nie może być zbyt „pewne”. Oczywiście można obliczyć drzewo na podstawie części cech, następnie obliczyć drzewo z wykorzystaniem wszystkich cech i jeżeli drzewa będą takie same lub niemal identyczne, spokojnie użyć drzewa do analizy ewolucji cech, jednak zwykle tak obliczone drzewa będą zdecydowanie różne.

Można określić częstość różnych przejść między stanami cechy na kladogramie, jednak uznanie ich za parametry probabilistycznego modelu transformacji budzi zastrzeżenia: próba jest niewielka (zwykle zaledwie kilka zmian), metoda redukcjonistyczna daje estymaty obciążone tendencją do minimalizacji liczby zmian, a ponadto estymat zawsze podlega **obciążeniu macierzą kroków** (*step-matrix-bias*; Mickevich 1982, Mickevich i Weller 1990). Oznacza to po prostu, że przyjęty sposób liczenia kroków – wynikający z modelu transformacji – nieuchronnie rzutuje na otrzymaną rekonstrukcję. Dlatego też Mickevich (1982) zaproponowała tzw. **analizę serii transformacyjnych** (*transformation series analysis – TSA*), która nie zaczyna od macierzy kroków ani drzewa stanów cech, a oblicza drzewa stanów cech na podstawie analizy rozkładów stanów cech na drzewie, starając się omijać jakiegokolwiek wstępne założenia rzutujące na wynik rekonstrukcji filogenezy, w tym macierz kroków (Mickevich i Weller 1990). Zasadą TSA jest konstrukcja drzew stanów cech, odzwierciedlających najlepiej hierarchię stanów cech, wynikającą z obliczonego kladogramu. To „odzwierciedlanie” (*reflectance*) nie zostało jednak matematycznie zdefiniowane ani algorytm sprawdzony pod względem znajdowania „najlepszego odzwierciedlenia”, a teoretyczne uzasadnienie wykorzystywania „odzwierciedlenia” do wyboru najlepszej hipotezy również nie zostało podane; co więcej, TSA nie sprawdzono nawet metodą symulacji (Maddison i Maddison 1992), więc użyteczność i wiarygodność TSA pozostają wątpliwe.

Obciążenie macierzą kroków niewątpliwie stanowi problem, choć raczej nie w stopniu wykluczającym rekonstrukcję ewolucji cech na podstawie filogenezy, obliczonej z wykorzystaniem macierzy kroków – wynikiem tego obciążenia są właśnie obciążone, czyli niedokładne estymaty ewolucji cech, konserwatyzm tych estymatów, czyli odchylenie w kierunku wyjściowej macierzy; stosowana ostrożnie, technika powinna działać w miarę zadowalająco (Ryc. 4.13), pozwalając potwierdzać (ale i odrzucać, choć z mniejszą czułością) i zwiększać dokładność modeli zmian stanów cech. Wydaje się, że na ogół wstępna macierz kroków nie uniemożliwia wnioskowania o ewolucji cech na podstawie rekonstrukcji, choć problem istnieje i brak jakiegokolwiek pełnego rozwiązania, pomimo wielu propozycji w literaturze (patrz Maddison i Maddison 1992). Oczywiście rekonstrukcja stanów ancestralnych bywa niepewna. Istnieją techniki estymacji parametrów procesu ewolucyjnego niebazujące na historycznych rekonstrukcjach, lecz na rozmieszczeniu stanów cechy, formie drzewa i stochastycznym modelu ewolucji (modele takie dotyczą niemal wyłącznie cech molekularnych, bowiem dla morfologicznych modele formułować trudno), zapewne ich wyniki są bardziej wiarygodne, lecz wykraczają one poza technikę redukcjonistyczną.



Ryc. 4.13. Choć ryzyko logiki błędnego koła przy rekonstrukcji ewolucji cech na podstawie filogenezy jest zawsze realne, to jednak można go uniknąć

Argumentować można, że lepiej poznać historię ewolucji cechy w przybliżeniu niż nie poznać jej wcale, zwłaszcza że technika kladystyczna może być użyta w każdym przypadku, gdy inne metody wymagają znajomości modelu i wiedzy, której często brak. Symulacje wykazały, że rekonstrukcje stanów ancestralnych metodą redukcjonistyczną są dość wiarygodne (Maddison W.P. 1990, Martins i Garland 1991), podobnie jak obliczone w ten sposób estymaty przepływu genów (Slatkin i Maddison 1989). Oczywiście często mamy więcej niż jedno drzewo MPR i więcej niż jedną rekonstrukcję ewolucji cechy MPR; niekiedy mamy podstawy do odrzucenia którychś z nich, lecz regułą jest uwzględnianie wszystkich. Przy całej ostrożności logika błędnego koła może się zdarzyć, jak zdarza się w całej systematyce (Hull 1967). Na ogół jednak procedura iteratywnej rekonstrukcji filogenezy i ewolucji cech daje niezłe wyniki (Falniowski i Szarowska 1995).

Z drugiej strony, choć także i rekonstrukcja ewolucji cech może odpowiadać modelowi GIGO, wspomnianemu przy omawianiu technik fenetycznych – arbitralne i słabo bądź w ogóle nieuzasadnione założenia modelu ewolucji cechy przynoszą niewiarygodne rekonstrukcje filogenezy, a z nich wynika jeszcze bardziej oderwany od rzeczywistości obraz ewolucji cech – to bynajmniej tak być nie musi. Często podstawy przyjęcia określonej serii transformacyjnej są solidne: opierają się na danych embriologicznych, paleontologicznych, morfologii porównawczej, znajomości procesów ewolucyjnych, genetyki, biochemii, itd. Sama analiza filogenetyczna dostarcza drzew, na których odtwarzać możemy ewolucję cech: są to tzw. **dane wewnętrzne**; wszystko, co wiemy o badanej grupie niezależnie od przeprowadzonej rekonstrukcji, to tzw. **dane zewnętrzne**, które należy wykorzystać w największym możliwym stopniu.

Zwykle jednak obserwujemy konflikt między danymi wewnętrznymi i zewnętrznymi. Wówczas musimy rozstrzygnąć, czy nasza rekonstrukcja jest błędna, czy też pozwoli ona na lepsze zrozumienie procesów ewolucyjnych w badanej grupie. Kluczowa jest jakość danych zewnętrznych. W sytuacji, gdy skrócenie drzewa o, powiedzmy, dwa kroki wymaga odrzucenia wyników powtarzanych badań, powiedzmy, embriologicznych, odrzucenie naszej rekonstrukcji jest oczywiste. Z drugiej strony, gdy nasza seria transformacyjna opiera się np. na założeniu, że kolec krótki to stadium pośrednie między długim a szczątkowym (co jest logiczne, lecz bynajmniej biologicznie nieuzasadnione), to tego rodzaju „dane zewnętrzne” z powodzeniem możemy odrzucić. Podobnie może okazać się, że dwa stany cech to różne, ekologicznie warunkowane ekspresje tego samego allela, co wytłumaczy niezgodność rekonstrukcji z danymi zewnętrznymi. Choć istotą rozwoju nauki jest znajdowanie nowych rozwiązań starych problemów, warto zachować pewien stopień konserwatyizmu: gdy nasze rekonstrukcje przy pewnych założeniach potwierdzają dotychczasowy obraz filogenezy, a przy innych, podobnie uzasadnionych założeniach jej przeczą, należy jednak przyjąć te pierwsze założenia.

Obok rekonstrukcji ewolucji poszczególnych cech interesujące jest badanie koevolucji różnych cech: czy określony stan jednej cechy wykazuje związek z określonym stanem innej. W.P. Maddison (1990b, Maddison i Maddison 1992) zaproponował **test skoncentrowanych zmian** (*concentrated changes test*), badający, czy zrekonstruowane zmiany stanu cechy binarnej w obrębie określonej części drzewa, dla której stan innej cechy ma określoną wartość, są liczniejsze niż zachodzące losowo, bez związku z określonym stanem tej drugiej cechy. Dla mniejszych drzew prawdopodobieństwa obliczane są techniką wyliczenia wszystkich możliwych kombinacji zmian, dla większych konieczne są symulacje. W ten sposób możemy sprawdzać, czy można odrzucić hipotezę o braku korelacji ewolucji między cechami.

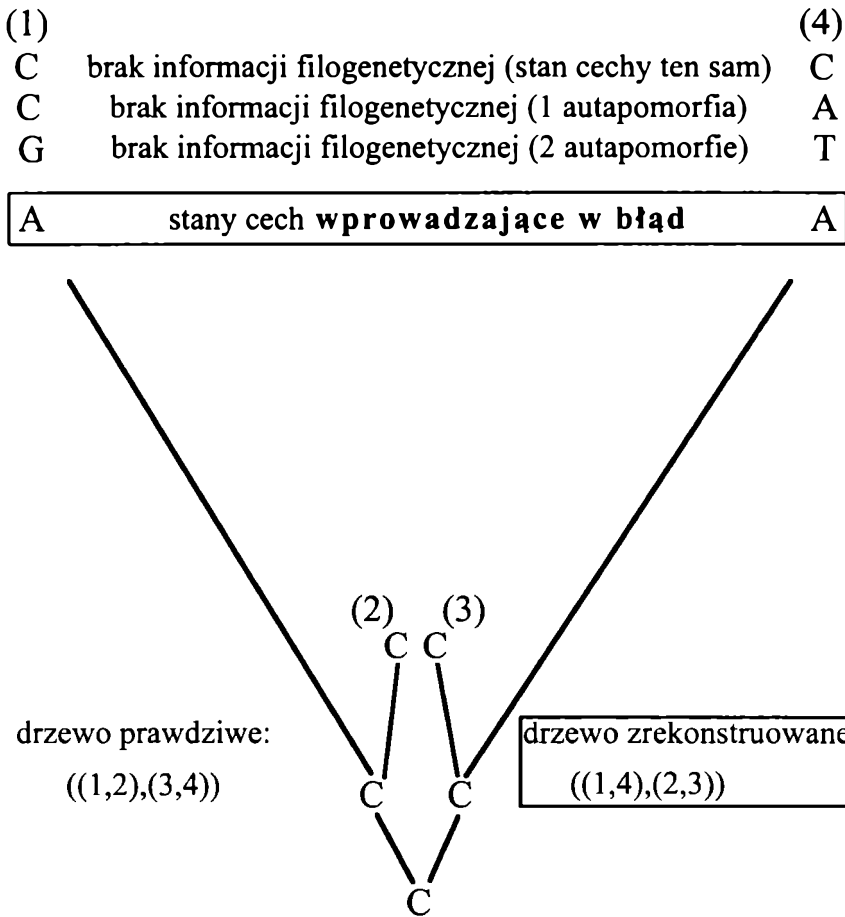
Zalety, wady i ograniczenia użyteczności metody kladystycznej

Jak zauważył Sober (1988), im mniej musimy wiedzieć o procesie ewolucyjnym, aby rekonstruować jego przebieg, tym bardziej polegać możemy na wynikach rekonstrukcji. Choć – jak już pisaliśmy – metoda kladystyczna bynajmniej nie jest, jak się często twierdzi, w pełni wolna od założeń, to jednak założeń jest tu niewiele. To raczej zaleta: nawet dla sekwencji DNA, których ewolucję wszystkie dotąd dostępne modele opisują w najlepszym razie z bardzo grubym przybliżeniem (Nei i Kumar 2000), minimalizacja liczby założeń dać może wyniki lepsze niż dla innych technik, zwłaszcza

gdy podstawień wstecznych lub równoległych (czyli homoplazji) brak lub są bardzo nieliczne, a porównywane sekwencje długie; metoda powinna dać dobre wyniki dla grupy bliskich sobie taksonów, choć w praktyce sekwencje są niezbyt długie, a mutacje wsteczne i równoległe częste, i wówczas technika zawodzi. Dla cech morfologicznych metoda pozwala na wykorzystanie całej wiedzy o seriach transformacyjnych, a także odtwarzanie ewolucji cech w trakcie filogenezy, jak przedstawiliśmy to w poprzednim podrozdziale. Warto jednak nie zapominać, że metoda kladystyczna ze swej istoty – którą jest redukcja liczby hipotez *ad hoc* o homoplazjach – dąży do opisania wszystkich podobieństw jako następstwa wspólnego pochodzenia, a więc potencjalnie zawyża – a przynajmniej maksymalizuje – udział homologii w tłumaczeniu podobieństw między taksonami; ponadto przyjmuje zawsze najprostszy możliwy przebieg ewolucji.

Zaletą techniki jest niewątpliwie stosunkowo duża wiedza o różnych jej aspektach i sprawdzone zachowanie w wielu sytuacjach, wynikające z bogatej literatury, przedstawiającej zarówno teoretyczne studia nad metodą, jak i wyniki rekonstrukcji filogenezy, prowadzonych z jej wykorzystaniem. Metoda jest też łatwa do zrozumienia, a więc zachodzi mniejsze niebezpieczeństwo jej niewłaściwego użycia bądź interpretacji wyników. Istnieje ponadto cały szereg dobrych, dopracowanych programów komputerowych, rekonstruujących filogenezę tą techniką. Pamiętajmy jednak, że choć matematycznie dopracowana, metoda redukcjonistyczna opornie i w ograniczonym stopniu poddaje się statystycznemu testowaniu, bowiem brak jakiegś jednoznacznej techniki obliczania średnich czy wariancji długości drzewa albo innych parametrów. W swej istocie metoda poszukuje *tego jednego, najlepszego* drzewa, choć to rzeczywiście znalezione nie musi być najlepsze, a poziomu ufności dla obliczonego drzewa określić się nie da: technika jest osadzona w przyczynowo-skutkowym obrazie historycznych zaszłości, bynajmniej nie w probabilistycznym podejściu do opisu procesów ewolucyjnych.

Metoda redukcjonistyczna w niektórych przypadkach daje niestety wyniki niespójne, bowiem nawet duży i nadal powiększany zestaw danych może konsekwentnie dawać obraz pokrewieństw niezgodny z rzeczywistością. Felsenstein (1978) przedstawił hipotetyczny przypadek dla czterech taksonów (sekwencji), kończących linie ewolucyjne, różniące się zdecydowanie tempem substytucji (Ryc. 4.14), choć oczywiście podobnie może się dziać i dla innych cech, np. morfologicznych. W sytuacji gdy u taksonów kończących długie linie znajdujemy na danej pozycji ten sam nukleotyd co u pozostałych dwóch taksonów, kończących wolno ewoluujące linie, informacji filogenetycznej brak. Oczywiście ten sam nukleotyd nie musi oznaczać niezmienności tej pozycji sekwencji – zmian mogło być wiele, lecz pozostają one niewykrywalne. W przypadku gdy u jednego z taksonów nukleotyd się zmienił, mamy do czynienia z autapomorfią, więc również brak sygnału filogenetycznego; podobnie gdy są dwa różne podstawienia. Jeżeli jednak na końcu obu szybko ewoluujących linii (taksony 1 i 4) – reprezentowanych więc przez długie gałęzie – znajdujemy te same nukleotydy, to technika redukcjonistyczna nieuchronnie wskaże na pokrewieństwo tych taksonów – wskaże błędnie. Przypadki takie określa się mianem **strefy Felsensteina** (*Felsenstein's zone*). Henny i Penny (1989) nazwali tę właściwość techniki redukcjonistycznej **przyciąganiem się długich gałęzi** (*long branch attraction*), a Nei (1996) – **przyciąganiem się krótkich gałęzi** (*short branch attraction*).



Ryc. 4.14. Zasada zjawiska przyciągania długich gałęzi, przedstawiona na czterotaksonowym nieukorzenionym drzewie, dla jednej pozycji sekwencji DNA. Gdy dwie gałęzie zewnętrzne (zakończone taksonami 2 i 3) i gałąź wewnętrzna są krótkie – ewolucja biegła tam wolno – a pozostałe dwie gałęzie (zakończone taksonami 1 i 4) są długie – ewolucja biegła tam szybko – to na tych długich gałęziach mógł zajść wiele zmian, których nie jesteśmy w stanie określić. Znamy jedynie stany u taksonów terminalnych, na końcach krótkich gałęzi występuje C, można się spodziewać, że przy powolnej ewolucji stan cechy nie uległ zmianie. Na końcach długich gałęzi jakiegokolwiek stany cech nie będą wówczas filogenetycznie informatywne albo będą wprowadzały w błąd, wskazując na pokrewieństwo tych taksonów

Zjawisko przyciągania długich gałęzi można zilustrować przykładem, przedstawionym przez Swofforda i innych (1996). Korzysta on z prostego modelu zmian cechy binarnej, wprowadzonego przez Cavendera i Felsensteina (1987). Zakłada, że oba stany cechy (0, 1) występują z taką samą częstością, a prawdopodobieństwo zmiany $0 \rightarrow 1$ jest takie samo jak zmiany $1 \rightarrow 0$, model jest więc w pełni odwracalny. Prawdopodobieństwo występowania stanów i, j, k i l (0 lub 1) u taksonów 1, 2, 3 i 4 oznaczmy P_{ijkl} . Znowu skorzystamy z prostego przykładu czterotaksonowego drzewa. Prawdopodobieństwo zmiany stanu cechy równe jest ϵ dla długich gałęzi, a ϕ dla gałęzi krótkich (Ryc.

4.15a). Odwracalność przyjętego modelu pozwala na ukorzenie drzewa w dowolnym punkcie (Ryc. 4.15b). Dla czterotaksonowego drzewa i danego zestawu stanów cechy binarnej u terminalnych taksonów możliwe są cztery rekonstrukcje stanów cechy w węzłach wewnętrznych (Ryc. 4.15c–f). Oczywiście jeżeli prawdopodobieństwo zmiany stanu cechy wzdłuż gałęzi wynosi ε , to prawdopodobieństwo braku zmiany dla tej samej gałęzi równe jest $(1 - \varepsilon)$ (a dla ϕ : $1 - \phi$). Zmiany (brak zmian) dla każdej gałęzi są niezależne od mających miejsce na innych gałęziach, więc prawdopodobieństwo określonej rekonstrukcji obliczymy jako iloczyn prawdopodobieństw dla wszystkich gałęzi oraz dla stanu początkowego (w miejscu arbitralnego ukorzenia drzewa) – to ostatnie, jak wynika z modelu, równe jest $1/2$. Prawdopodobieństwo wystąpienia określonego zestawu cech i, j, k, l u taksonów terminalnych równe będzie sumie prawdopodobieństw wszystkich czterech rekonstrukcji, tak więc:

$$\begin{aligned} P_{0011} &= (1/2)\varepsilon\phi(1 - \varepsilon)(1 - \phi)^2 + (1/2)\phi(1 - \varepsilon)^2(1 - \phi)^2 + (1/2)\varepsilon^2\phi^3 + (1/2)\varepsilon\phi(1 - \varepsilon)(1 - \phi)^2 \\ &= (1/2)[\phi(1 - \varepsilon)(1 - \phi)^2(\varepsilon + 1) + \varepsilon^2\phi^3]. \end{aligned}$$

Z odwracalności modelu wynika, że prawdopodobieństwo wystąpienia odwrotnych stanów cechy u taksonów terminalnych jest takie samo, a więc:

$$P_{0011} + P_{1100} = 2 P_{0011} = \phi(1 - \varepsilon)(1 - \phi)^2(\varepsilon + 1) + \varepsilon^2\phi^3.$$

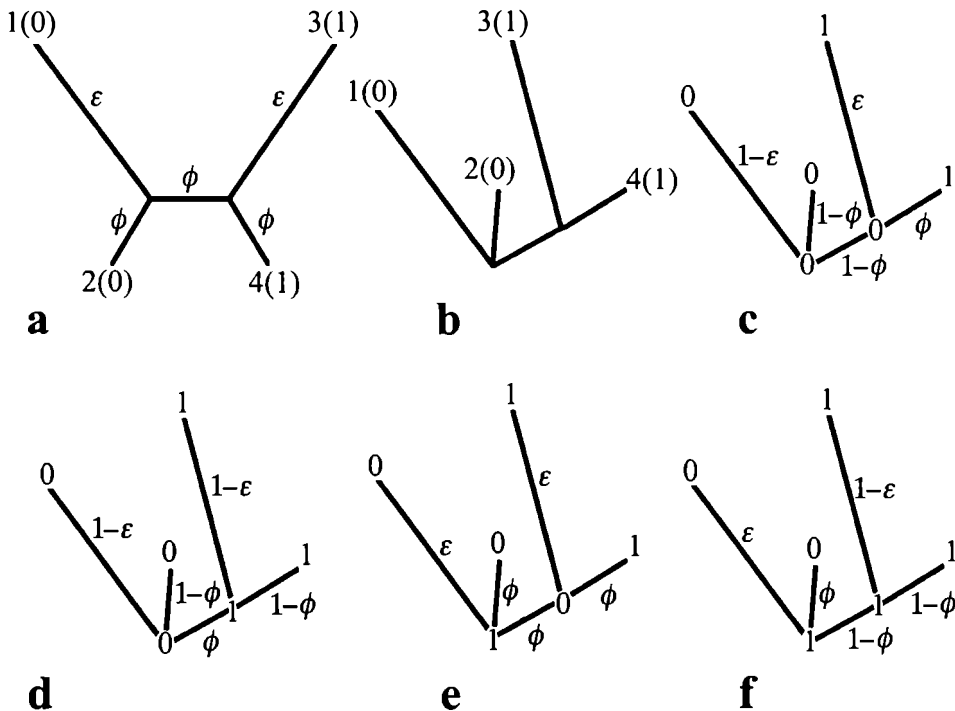
W identyczny sposób obliczyć możemy prawdopodobieństwo dwóch pozostałych możliwych rozkładów stanów cechy u terminalnych taksonów:

$$P_{0101} + P_{1010} = 2 P_{0101} = \phi^2(1 - \varepsilon)(1 - \phi)(\varepsilon + 1) + \varepsilon^2(1 - \phi)^3.$$

Filogenetycznie poprawne będzie oczywiście drzewo ((1,2),(3,4)), a nie drzewo ((1,3),(2,4)); aby nieważona technika kladystyczna je znalazła, konieczny jest u taksonów terminalnych rozkład cech, występujący z prawdopodobieństwem $P_{0011} + P_{1100}$, inaczej rekonstrukcja musi być błędna. W przedstawionym przykładzie $\varepsilon > \phi$. Dla wielu wartości spełniających tę nierówność $P_{0101} + P_{1010}$ będzie większe niż $P_{0011} + P_{1100}$. Przykładowo, dla $\varepsilon = 0,5$ i $\phi = 0,1$: $P_{0101} + P_{1010} = 0,189$, gdy $P_{0011} + P_{1100} = 0,061$, a więc prawdopodobieństwo wystąpienia rozkładu stanów cech wspierającego błędną rekonstrukcję jest trzykrotnie wyższe niż dla rozkładu umożliwiającego rekonstrukcję prawidłową. Dla $\varepsilon = 0,4$ i $\phi = 0,1$: $P_{0101} + P_{1010} = 0,1242$, gdy $P_{0011} + P_{1100} = 0,0682$, więc proporcja tych prawdopodobieństw wynosi 1,8.

Model Felsensteina bywał krytykowany jako nierealistyczny, jednak dla bardziej złożonych i całkowicie realistycznych modeli wykazać można to samo, tyle że wymaga to złożonych procedur matematycznych, jak choćby przedstawione dalej sprzężenie Hadamarda (Swofford i inni 1996). Co więcej, w pewnych warunkach zjawisko to wystąpi również wówczas, gdy tempo substytucji jest takie samo we wszystkich liniach ewolucyjnych (Hendy i Penny 1989, Zharkikh i Li 1993, Takezaki i Nei 1994, Kim 1996). Na szczęście, paradoksalnie, ryzyko wystąpienia przyciągania się długich gałęzi maleje, gdy rekonstruowane drzewa obejmują więcej taksonów. Już intuicyjnie łatwo zauważyć, że bliskie sobie długie gałęzie pochodzą od stosunkowo niedawnego wspólnego przodka, a kolejne zmiany mogły zachodzić równolegle w sposób identyczny, natomiast dla odległych długich gałęzi jest to mało prawdopodobne. W strefie Felsen-

steina ogólna reguła, że im więcej danych, tym bardziej spójna rekonstrukcja, niestety nie obowiązuje i zwiększanie zestawu danych jedynie upewnia nas, że błędny wynik jest prawdziwy. Można więc próbować prowadzić rekonstrukcję na mniejszych zestawach danych, licząc na to, że któreś z rekonstrukcji wskażą prawidłowe rozwiązanie. Sensowne jest spróbowanie rekonstrukcji na podstawie kilku najbardziej konserwatywnych cech (np. fragmentów sekwencji), a włączenie pozostałych do analizy dopiero po ustaleniu położenia taksonów leżących na końcach długich gałęzi. Najpewniej jest jednak unikać takich gałęzi: drzewo o niektórych gałęziach długich powinno budzić nieufność, a dodanie do analizy więcej taksonów, także zewnętrznych dla badanej grupy, najczęściej pozwala zlikwidować przynajmniej większość długich gałęzi. Inna sprawa, że grupa zewnętrzna wręcz z reguły powoduje pojawienie się na drzewie długich gałęzi, a jej dołączanie służy przecież wyłącznie ukorzenieniu drzewa: czasem więc lepiej pozostać przy drzewie nieukorzenionym albo też pokrewieństwa w obrębie grupy wewnętrznej zrekonstruować na drzewie nieobejmującym grupy zewnętrznej, a dołączyć ją dopiero później – dla wskazania miejsca ukorzenienia grupy wewnętrznej, dla której zachowujemy topologię sprzed ukorzenienia. Ryzyko niespójności jest też mniejsze dla techniki kladystycznej ważonej, lecz nawet metody oparte na maksymalizacji wiarygodności nie są od niego wolne.



Ryc. 4.15. Nieukorzenione czterotaksonowe drzewo o dwóch gałęziach znacznie dłuższych od pozostałych (a) ukorzeniamy w dowolnym punkcie (b) i obliczamy prawdopodobieństwa wszystkich możliwych rekonstrukcji stanu cechy binarnej w obu wewnętrznych węzłach (c, d, e, f), dla danego rozkładu stanów cechy u taksonów terminalnych (a), gdy prawdopodobieństwo zmiany stanu cechy na krótkiej gałęzi równe jest ϕ , a na długiej ϵ

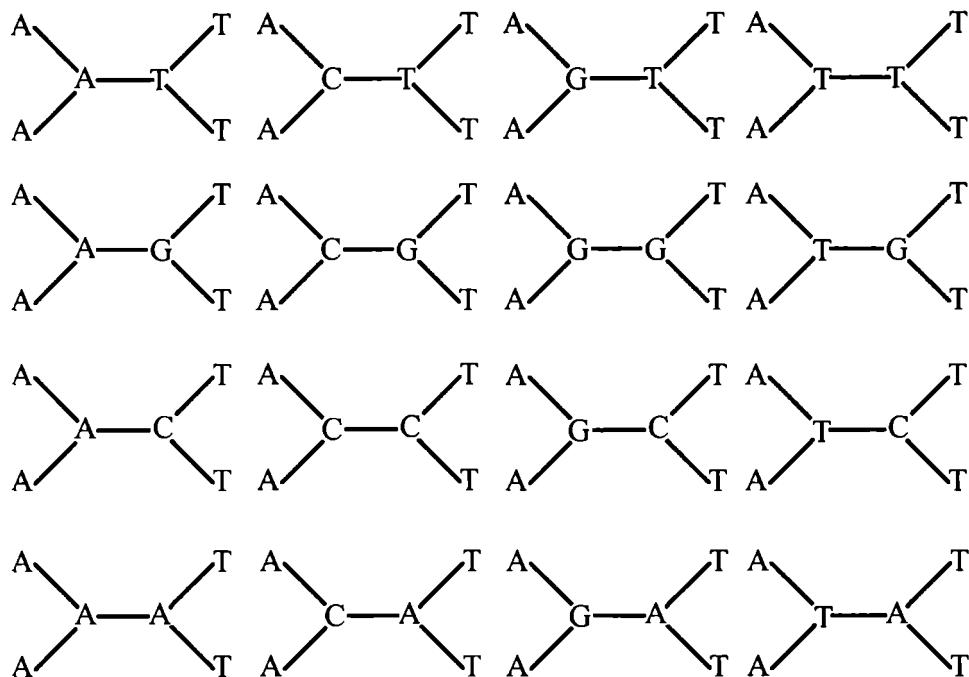
4.6. Metody oparte na maksymalizacji wiarygodności

Metody oparte na maksymalizacji wiarygodności (*maximum likelihood* – ML) zaproponowali Cavalli-Sforza i Edwards (1967) dla częstości alleli, Felsenstein (1981a) dla sekwencji kwasów nukleinowych, a Kishino i inni (1990) dla sekwencji białek. Goldman (1990) przedstawia dobre wprowadzenie do ML. Pojęcie wiarygodności (*likelihood*) przedstawiliśmy już w Rozdziale 2.12, omawiającym odległości dla sekwencji kwasów nukleinowych i białek, przypomnijmy jednak i uzupełnijmy to wprowadzenie. Formalnie, jeżeli mamy jakieś dane D i hipotezę H , to wiarygodność L jest daną wzorem:

$$L_D = \Pr(D|H),$$

czyli jest to prawdopodobieństwo zaistnienia danych D przy spełnieniu hipotezy H . Jest to więc prawdopodobieństwo warunkowe. Przypomnijmy: technika estymuje wiarygodność danych przy przyjęciu hipotezy H (szacuje zgodność danych z tą hipotezą), natomiast nie mówi nic o wiarygodności samej hipotezy. Zilustrować to można anegdotą. Gdy w latach sześćdziesiątych dwudziestego wieku szach Iranu odwiedził Polskę, wizyta ta była oczywista dla pewnego młodego chłopca: „przyjechał, gdyż obiecał to Koziołkowi Matołkowi”. Zważywszy, że tak dostojny monarcha zapewne dotrzymywał obietnic, wiarygodność przedstawionego wytłumaczenia była wysoka, choć z drugiej strony wiarygodność zarówno samej obietnicy, jak i w ogóle istnienia bohatera książki Kornela Makuszyńskiego pozostają bliskie zeru. Wiarygodności nie należy mylić z prawdopodobieństwem: prawdopodobieństwa sumują się do 1, wiarygodności – nie: w naszym przypadku sumujemy wiarygodności dla danego drzewa i przyjętego modelu ewolucji, i suma ta będzie zawsze (znacznie) mniejsza niż 1. Dopiero zsumowanie wiarygodności dla wszystkich możliwych drzew dałoby wartość 1, lecz takie sumowanie nie ma uzasadnienia. Oczywiście wybieramy takie drzewo, dla którego wiarygodność osiąga największą wartość.

Technika maksymalizacji wiarygodności jest szeroko stosowana w statystycznym testowaniu hipotez. W naszym przypadku – rekonstrukcji filogenezy – danymi są stany cech i optymalizowane drzewo, przy czym – odmiennie niż w technice kladystycznej – optymalizuje się nie tylko topologię, lecz i długość gałęzi. To ostatnie pozwala na ominięcie niektórych mankamentów metody kladystycznej – wiadomo, że prawdopodobieństwo zmiany na długiej gałęzi jest większe niż na krótkiej – jednak technika, choć mniej podatna, również nie jest wolna od niebezpieczeństwa przyciągania długich gałęzi, do czego jeszcze wrócimy. Musimy dysponować modelem ewolucyjnych zmian – co jest siłą, lecz i słabością metody, bowiem korzystne jest dokładne sprecyzowanie założeń procesu, lecz z drugiej strony dla wielu danych modelu brak. Ponadto model musi być wspólny dla całości danych: nie dziwi więc, że technikę stosuje się dla cech molekularnych, natomiast zupełnie sporadycznie – dla morfologicznych. Plusem metody jest możliwość statystycznego testowania różnych hipotez o przebiegu ewolucji. Modele zmian przedstawimy pokrótce dla najczęściej tą techniką analizowanych sekwencji kwasów nukleinowych i białek.



Ryc. 4.16. Dla czterotaksonowego drzewa, gdzie u dwóch taksonów terminalnych występuje A i u pozostałych dwóch T, możliwe jest 16 różnych rekonstrukcji stanów w dwóch węzłach wewnętrznych. Choć wiele z nich jest niezmiernie mało prawdopodobnych, to wszystkie muszą być uwzględnione, co dobrze ilustruje złożoność obliczeń w technice maksymalizacji wiarygodności

Modele ewolucji kwasów nukleinowych i białek

Dla czterotaksonowego drzewa, w którym u dwóch z taksonów terminalnych występuje na rozpatrywanej pozycji sekwencji DNA adenina, a u dwóch pozostałych tymina, możliwych kombinacji stanów ancestralnych w dwóch wewnętrznych węzłach może być 16 (Ryc. 4.16). I choć oczywiście prawdopodobieństwa wystąpienia kolejnych z nich są różne, niektóre relatywnie niskie, to jednak każdą z możliwości technika ML musi uwzględnić, obliczając wiarygodność, i ilustruje to złożoność i intensywność obliczeń. Oczywiście liczba możliwych kombinacji rośnie gwałtownie z liczbą terminalnych taksonów T (analizowanych sekwencji), będąc równa dla kwasów nukleinowych $4^{(T-2)}$ (a dla białek $20^{(T-2)}$). Rzecz jasna, prawdopodobieństwa wystąpienia w tych węzłach określonych nukleotydów – i prawdopodobieństwa określonych długości poszczególnych gałęzi, które również podlegają ewaluacji – zależą od przyjętego modelu ewolucji DNA. Podkreślić należy, że ani nie istnieje jakiś uniwersalny model, ani też którykolwiek z istniejących modeli nie odzwierciedla dokładnie rzeczywistości w jakimkolwiek przypadku. Szereg modeli omówiliśmy już w Rozdziale 2.12 i do rozdziału tego odsyłamy. Tutaj przypomnimy jedynie i uzupełnimy niektóre, prostsze oraz

częściej stosowane, za Swoffordem i innymi (1996). Jeśli omawiając różnice pomijaliśmy pozycje z identycznymi nukleotydami, tutaj musimy je uwzględnić, bowiem prawdopodobieństwo/koszt braku zmiany także ma określoną wartość, którą też trzeba estymować. W najogólniejszej formie macierz częstości zmian nukleotydów (liczby substytucji na pozycję na jednostkę czasu: każdy z elementów Q_{ij} przedstawia częstość zamiany nukleotydu i przez nukleotyd j w czasie dt) jest następująca:

$$Q = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_C + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{bmatrix},$$

gdzie kolejne rzędy i kolumny odpowiadają kolejno A, C, G i T; μ to średnia częstość substytucji, g to proporcjonalny współczynnik częstości dla danej substytucji nukleotydu i nukleotydem j ($a, b, c, d, e, f, g, h, i, j, l$), odpowiednio modyfikujący wartość μ , a π_A, π_C, π_G i π_T to częstości nukleotydów A, C, G i T. Zakładamy, że częstości te nie zmieniają się w czasie. Elementy przekątnej macierzy obliczamy tak, aby elementy każdego rzędu macierzy sumowały się do zera. Przypominamy, że rozpatrujemy **model Markova** (użyteczność tego modelu rozpatrują m.in. Felsenstein 1981a, Barry i Hartigan 1987, Kishino i Hasegawa 1989, Rodriguez i inni 1990), w którym zmiany nukleotydów zachodzą niezależnie od historii danej pozycji, czyli wcześniejszych podstawień, a także **stacjonarny**, czyli parametry procesu zmian nie różnią się w czasie, a więc pozostają takie same na całym zrekonstruowanym drzewie. Założenie niezależności zmian jest kluczowe, bowiem umożliwia obliczanie łącznej wiarygodności jako iloczynu wiarygodności cząstkowych. Macierz Q często wygodnie jest poddać dekompozycji na macierze R i Π :

$$R = \begin{bmatrix} - & \mu a & \mu b & \mu c \\ \mu g & - & \mu d & \mu e \\ \mu h & \mu j & - & \mu f \\ \mu i & \mu k & \mu l & - \end{bmatrix}, \quad R_{\text{odwracalny}} = \begin{bmatrix} - & \mu a & \mu b & \mu c \\ \mu a & - & \mu d & \mu e \\ \mu b & \mu d & - & \mu f \\ \mu c & \mu e & \mu f & - \end{bmatrix}, \quad \text{a } \Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{bmatrix}.$$

Najczęściej rozpatruje się modele **odwracalne w czasie** jako prostsze, a ponadto wygodne, bowiem ewaluowane drzewo można wówczas ukorzenieć w dowolnym miejscu, zaś ukorzeniecie wydatnie upraszcza obliczenia, zmniejszając liczbę gałęzi, których długość trzeba ewaluować (patrz niżej). Odwracalność ($R_{\text{odwracalny}}$) oznacza, że częstość przejścia nukleotydu i w nukleotyd j jest taka sama jak nukleotydu j w nukleotyd i . Najogólniejszy, odwracalny w czasie model *GTR* (Barry i Hartigan 1987, Rodriguez i inni 1990, Yang i inni 1994) można przedstawić następująco:

$$Q = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{bmatrix}.$$

Większość z pozostałych używanych częściej modeli można uzyskać, ograniczając parametry modelu *GTR*. Najprostszy, zaproponowany przez Jukes'a i Cantora (1969), zapisać więc można w postaci:

$$Q = \begin{bmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{bmatrix}, \text{ a podstawiając } \alpha = \frac{\mu}{4} : Q = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}.$$

Jeżeli dopuścimy zróżnicowanie częstości nukleotydów, powyższy model staje się modelem *F81* (Felsenstein 1981a), znanym też jako *equal input model* (Tajima i Nei 1982):

$$Q = \begin{bmatrix} -\mu(\pi_Y + \pi_G) & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_R + \pi_T) & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & -\mu(\pi_Y + \pi_A) & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & -\mu(\pi_R + \pi_C) \end{bmatrix},$$

a dwuparametrowy model *K2P* (Kimura 1980), dla którego $a = c = d = f = 1$, a $b = e = \kappa$ (a więc częstość tranzycji $\alpha = \mu\kappa/4$, częstość transwersji $\beta = \mu/4$, czyli $\kappa = \alpha/\beta$):

$$Q = \begin{bmatrix} -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa \\ \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) \end{bmatrix} = \begin{bmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{bmatrix},$$

Z kolei model *HKY85* (Hasegawa i inni 1985, Rzhetsky i Nei 1995) jest modyfikacją modelu *K2P*, dopuszczającą zróżnicowanie częstości nukleotydów:

$$Q = \begin{bmatrix} -\mu(\kappa\pi_G + \pi_Y) & \mu\pi_C & \mu\kappa\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\kappa\pi_T + \pi_R) & \mu\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\kappa\pi_A + \pi_Y) & \mu\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\kappa\pi_C + \pi_R) \end{bmatrix},$$

gdzie $\alpha = \mu$, $\beta = \mu\kappa$, $\pi_R = \pi_A + \pi_G$, a $\pi_Y = \pi_C + \pi_T$, czyli jest to modyfikacja modelu *GTR*, gdy $a = c = d = f = 1$ i $b = e = \kappa$. Model *F84* (Felsenstein 1984, Kishino i Hasegawa 1989, Tateno i inni 1994) dzieli częstość substytucji na dwie składowe:

ogólną częstość substytucji i wewnątrzgrupową częstość substytucji, będących wyłączenie tranzycjami. I ten model można uzyskać z modelu *GTR*, przyjmując $a = c = d = f = 1$, $b = (1 + K/\pi_R)$, $e = (1 + K/\pi_Y)$:

$$Q = \begin{bmatrix} - & \mu\pi_C & \mu\pi_G(1 + K/\pi_R) & \mu\pi_T \\ \mu\pi_A & - & \mu\pi_G & \mu\pi_T(1 + K/\pi_Y) \\ \mu\pi_A(1 + K/\pi_R) & \mu\pi_C & - & \mu\pi_T \\ \mu\pi_A & \mu\pi_C(1 + K/\pi_Y) & \mu\pi_G & - \end{bmatrix},$$

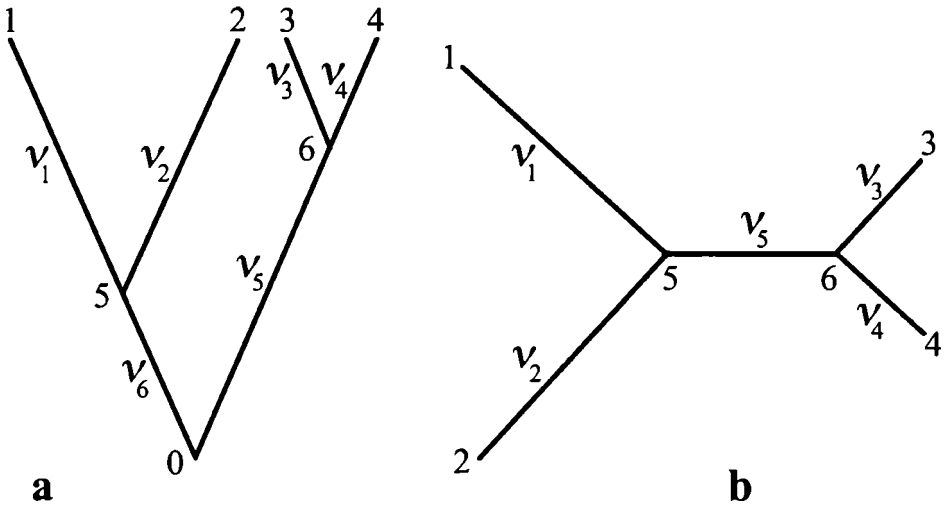
gdzie K jest parametrem określającym proporcję tranzycji do transwersji (dla $K = 0$ model staje się modelem *F81*, ze wzrostem wartości K tranzycji jest coraz to więcej niż transwersji), $\pi_R = \pi_A + \pi_G$, a $\pi_Y = \pi_C + \pi_T$. Przedstawiliśmy jedynie modele najprostsze; jak już pisaliśmy w Rozdziale 2.12, szereg modeli istnieje w modyfikacji Γ , dopuszczającej zróżnicowanie częstości substytucji zależnie od położenia nukleotydu w obrębie sekwencji; poważnie komplikuje to te modele, lecz bywa niezbędne. Modele uwzględniające rozkład Γ omawiają m.in. Kocher i Wilson 1991, Yang 1993, 1994a, Sullivan i inni 1995, Yang i Kumar 1996. Gu i inni (1995) oraz Waddell i Penny (1996) zaproponowali rozkład Γ z parametrem α estymowanym jedynie dla części pozycji, gdy dla innych nie, a Felsenstein i Churchill (1996) – ukryty model Markowa dla uwzględniania zmienności częstości substytucji w obrębie sekwencji; powinien on dawać podobne wyniki, jak zaproponowany przez Yanga (1994b) model dyskretnych (czyli nieciągłych) rozkładów Γ . Oczywiście dla zmian aminokwasów w sekwencjach białek również można formułować modele – jak już o tym pisaliśmy w Rozdziale 2.12 – choć są one mniej dopracowane i bardziej złożone. Ponadto – zamiast macierzy 4×4 jak dla kwasów nukleinowych – konieczne jest uwzględnianie macierzy 20×20 . Modele dla białek przedstawiają m.in. Dayhoff i inni (1978), Jones i inni (1992), Hasegawa i Fujiwara (1993).

Obliczanie wiarygodności drzewa dla kwasów nukleinowych

Wracając do zdefiniowanej wcześniej wiarygodności, dla sekwencji kwasów nukleinowych najogólniej przedstawić ją możemy wzorem:

$$L = f(\mathbf{x}; \theta),$$

gdzie f to odpowiednia funkcja, \mathbf{x} to zestaw analizowanych sekwencji nukleotydów, a θ to zestaw parametrów, takich jak długości gałęzi drzewa, frekwencje nukleotydów i parametry procesów substytucji użyte w przyjętym modelu. Warto już teraz zwrócić uwagę, że w równaniu brak jakiegokolwiek elementu, odzwierciedlającego topologię drzewa. Zestaw parametrów może być bardzo bogaty, tu poprzestaniemy oczywiście na najprostszych modelach.



Ryc. 4.17. Zasada obliczania wiarygodności drzewa na przykładzie prostego czterotaksonowego drzewa. Odwracalność zmian zakładana przez większość modeli ewolucji kwasów nukleinowych pozwala zamiast ukorzenionego (a) rozpatrywać nieukorzenione drzewo (b), dzięki czemu można pominąć węzeł 0 i gałąź 6; v_i to spodziewana liczba substytucji na gałęzi i

Aby przedstawić zasadę obliczeń, znów najlepiej skorzystać z przykładu czterotaksonowego drzewa (Ryc. 4.17). Załóżmy, że mamy cztery sekwencje o długości n , współosiowane bez insercji czy delecji, porównywane w pozycji k , w której mamy u kolejnych taksonów nukleotydy x_1, x_2, x_3 i x_4 , a w pozostałych węzłach x_0, x_5 i x_6 , choć nie wiemy, które z nukleotydów występowały w tych trzech węzłach (x_i to A, C, G lub T). Oznaczmy $P_{ij}(t)$ prawdopodobieństwo, że w czasie t nukleotyd i zastąpiony zostanie nukleotydem j . ML dopuszcza różną częstość r substytucji na różnych gałęziach, toteż spodziewaną liczbę substytucji w czasie t wygodnie oznaczać jako $v = rt$, a więc dla gałęzi i $v_i = r_i t_i$; v_i w technice ML są estymowane jako parametry określające długość gałęzi, tak aby osiągnąć najwyższą wartość wiarygodności. Dla drzewa z Ryc. 4.17a funkcja określająca wiarygodność dla pozycji k dana jest wzorem:

$$l_k = \pi_{x_0} P_{x_0 x_5}(v_5) P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2) P_{x_0 x_6}(v_6) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4),$$

gdzie π_{x_0} jest wyjściowym prawdopodobieństwem, że w węźle 0 występował nukleotyd x_0 . Parametr ten można przyjąć za równy średniej częstości danego nukleotydu w porównywanych sekwencjach, lecz można go również estymować techniką ML. Aby ustalić wartość $P_{ij}(v)$, musimy przyjąć któryś z modeli ewolucji kwasów nukleinowych. Dla modelu F81 (Felsenstein 1981), oznaczając π_i częstość nukleotydu i :

$$P_{ii}(v) = \pi_i + (1 - \pi_i) e^{-v}, \text{ dla } i \neq j; P_{ij}(v) = \pi_j (1 - e^{-v}), \text{ a warunek: } \pi_i P_{ij}(v) = \pi_j P_{ji}(v)$$

oznacza odwracalność modelu zmian, spełnianą przez *F81* i większość innych modeli. Wówczas wartość wiarygodności pozostaje niezmienna i również dobrze rozpatrywać możemy drzewo nieukorzenione (Ryc. 4.17b), niemające węzła 0; oznacza to też, że bez względu na miejsce ukorzenia wartość $(v_5 + v_6)$ pozostaje niezmienna, toteż $v_5 + v_6$ z drzewa (a) zastąpić można v_5 z drzewa (b). Wówczas:

$$L_k = \pi_{x_5} P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2) P_{x_5 x_6}(v_5) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4).$$

Ponieważ w praktyce nie wiemy, jakie nukleotydy odpowiadały x_5 i x_6 , więc wiarygodność dla pozycji k musimy obliczyć jako sumę wiarygodności dla wszystkich możliwych nukleotydów w węzłach 5 i 6:

$$L_k = \sum_{x_5} \sum_{x_6} \pi_{x_5} P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2) P_{x_5 x_6}(v_5) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4) = \sum_{x_5} [\pi_{x_5} P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2)] \sum_{x_6} [P_{x_5 x_6}(v_5) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4)].$$

Oczywiście wiarygodność drzewa obliczana jest dla całych sekwencji, nie dla jednej z pozycji. Całkowita wiarygodność dla zestawianych sekwencji równa jest iloczynowi wiarygodności dla wszystkich pozycji, bowiem w przyjmowanych modelach kolejne pozycje ewoluują niezależnie (co, jak wiemy, nie całkiem jest prawdziwe). Wiarygodności są jednak wartościami bardzo małymi, zwykle poza zakresem kalkulatorów, toteż w praktyce wygodniej posługiwać się ich logarytmami, zatem logarytm wiarygodności dla wszystkich sekwencji jest sumą logarytmów wiarygodności dla wszystkich pozycji:

$$\ln L = \sum_{k=1}^n \ln L_k.$$

Zmieniając wartości parametrów v_i , możemy teraz maksymalizować wartość $\ln L$ dla danej topologii, zapisać drzewo o najwyższej wartości $\ln L$, po czym tak samo optymalizować długość gałęzi dla kolejnych topologii, a finalnie wybrać tę topologię (z odpowiednio zoptymalizowanymi długościami gałęzi), dla której wartość $\ln L$ była najwyższa. Optymalizowanych parametrów – długości gałęzi – jest $2T - 3$; jeżeli możemy założyć działanie zegara molekularnego, wówczas długości gałęzi stają się miarami upływu czasu (Felsenstein 2000), a drzewo jest ultrametryczne – liczba estymowanych parametrów maleje niemal dwukrotnie, do $T - 1$ (gdyż odpowiednie gałęzie są sobie równe: $v_1 = v_2$, $v_3 = v_4$, a $v_1 + v_5 = v_4 + v_6$ na drzewie przedstawionym na Ryc. 4.17). Oczywiście optymalizacja teoretycznie obejmować powinna wszystkie możliwe topologie (patrz Rozdział 4.3), a tych już dla $T = 10$ jest 2 027 025. Intensywność obliczeń jest więc duża i już dla niewielu taksonów wyklucza pełne poszukiwanie globalnego optimum. Oczywiście wydajność algorytmu ma tu znaczenie kluczowe. Skorzystanie z drugiej postaci formuły na obliczanie L_k znacznie skróci czas obliczeń. Dla danego drzewa użyć można ogólnej techniki wielowymiarowej optymalizacji metodą Newtona (Edwards 1972) lub inną techniką numeryczną; Adachi i Hasegawa (1996) przedstawiają szybki algorytm, iteratywne algorytmy sformułowali m.in. Olsen i inni (1992), Tillier (1994) oraz Lewis i inni (1996). Kishino i inni (1990) przedstawili algorytm dla sekwencji białek. Algorytmy ML estymują najbardziej zgodne z modelem

długości gałęzi, lecz mogą także estymować parametry samego modelu (Felsenstein 2000: DNAML w pakiecie PHYLIP, Yang 1999: PAML, Swofford 1998: PAUP). Przykłady obliczeń ML zamieszczają Swofford i inni (1996) oraz Nei i Kumar (2000).

Poszukiwanie najlepszego drzewa ML

Jak wskazaliśmy, długość czasu obliczeń w technice ML jest czynnikiem ograniczającym możliwości jej użycia w stopniu daleko większym niż dla innych metod i już dla mniej niż 10 taksonów terminalnych wyklucza znajdowanie globalnego optimum w sposób pewny. Pozostają więc techniki przybliżone, których użyteczność nie jest taka sama jak np. w metodzie redukcjonistycznej (Nei i Kumar 2000). Przykładowo rozkład gwiazdy (SD: Saitou 1988, Adachi i Hasegawa 1969) nie daje drzew tak dobrych jak rozcinanie i powtórne łączenie drzewa (TBR) (Nei i Kumar 2000). Nei i inni (1998) sugerowali, że użycie wymiany najbliższego sąsiada z niewielu cyklami iteracji może znaleźć nie gorsze drzewo niż pełne poszukiwanie, bez porównania dłuższe. Symulacje komputerowe przeprowadzone przez Takahashi i Nei (2000) wskazują, że stopniowe dodawanie, po którym następuje szereg cykli wymiany najbliższego sąsiada (NNI), jest zwykle nie mniej efektywne niż daleko bardziej czasochłonna technika TBR. Jak i inne metody, ML wykazuje skłonność do znajdowania rekonstrukcji niezgodnych z rzeczywistym przebiegiem filogenezy zwłaszcza wówczas, gdy taksonów (sekwencji) jest wiele, a sekwencje krótkie (Nei i inni 1998, Nei i Kumar 2000), więc długotrwałe obliczenia i tak nie gwarantują wówczas bardziej spójnego wyniku.

Porównywanie wiarygodności dla różnych modeli i różnych drzew

Mając różne modele ewolucyjnych zmian, a także różne drzewa i obliczone wiarygodności dla każdego z nich, wybieramy najlepsze, czyli mające najwyższe wartości wiarygodności. Techniki ML pozwalają na statystyczne porównywanie istotności różnic między wartościami L metodą testu **proporcji wiarygodności** (*likelihood ratio test*). Jak wiemy, wartości L są bardzo małe i wygodniej używać ich logarytmów, toteż proporcja przyjmuje postać różnicy:

$$\Delta = \ln L_1 - \ln L_0,$$

gdzie L_1 to wiarygodność hipotezy H_1 , a L_0 to wiarygodność hipotezy H_0 . Jeżeli hipoteza H_0 jest specjalnym przypadkiem hipotezy H_1 , różniącym się brakiem jednego lub więcej parametrów uwzględnianych w hipotezie H_1 , to rozkład wartości 2Δ zbliża się do rozkładu χ^2 , o liczbie stopni swobody równej różnicy liczb parametrów, uwzględnianych przez modele. Jeżeli jedna z hipotez nie jest szczególnym przypadkiem drugiej, rozkład 2Δ nie odpowiada χ^2 i pozostaje próbkowanie numeryczne. Możemy przykładowo testować hipotezę o zgodności danych z występowaniem zegara molekularnego:

$$2\Delta = \ln L_{\text{brak zegara}} - \ln L_{\text{zegar obecny}},$$

gdyż rozkład 2Δ jest zbliżony do rozkładu χ^2 o $T - 2$ stopniach swobody. Zamiast testu χ^2 użyć możemy też testu G (Sokal i Rohlf 1995). Whelan i Goldman (1999) badali

symulacjami zgodność rozkładu 2Δ z χ^2 , przy liczbie stopni swobody równej różnicy liczb dobieranych parametrów w porównywanych modelach i stwierdzili, że zgodność jest dostateczna dla modeli nieuwzględniających rozkładu Γ , lecz nie dla modeli włączających ten rozkład. Dwa drzewa porównać też możemy testem, zaproponowanym przez Kishino i Hasegawę (1989):

$$\Delta = \sum_{i=1}^k (\ln L_{k, drzewo1} - \ln L_{k, drzewo2}) = \ln_{drzewo1} - \ln_{drzewo2}.$$

Test zakłada, że rozkład Δ zbliża się do normalnego, toteż gdy przedział błędu standardowego Δ zawiera 0, drzewa nie są statystycznie różne. Akaike (1974) sformułował dla technik ML **kryterium informacyjne Akaike'a** (*Akaike's information criterion – AIC*):

$$AIC = -2\ln L + 2p,$$

gdzie $\ln L$ to logarytm wiarygodności dla danego modelu, a p to liczba parametrów estymowanych.

Zastosowanie powyższych testów, choć rutynowe, budzi szereg wątpliwości. Przede wszystkim są to typowe testy dużych prób, więc dla stosunkowo krótkich sekwencji mogą działać źle. Jeden z porównywanych modeli musi być szczególnym przypadkiem drugiego, co nie zawsze ma miejsce. Dla estymacji parametrów modelu zmian musimy mieć prawidłowo zrekonstruowane drzewo, a tego nigdy nie możemy być pewni. Założenia modelu muszą też odpowiadać rzeczywistości, co w praktyce nigdy w pełni nie zachodzi, a często rozbieżności są znaczne: test wówczas daje zawyżone, innym razem zaniżone wartości. Ponadto dla krótkich sekwencji błąd standardowy estymowanych częstości nukleotydów jest znaczny.

Kryterium *AIC* możemy też użyć dla porównania dwóch modeli, z których jeden nie jest szczególnym przypadkiem drugiego, jeśli topologia drzewa pozostaje taka sama (Nei i Kumar 2000). Uważa się, że model jest lepszy, gdy wartość *AIC* jest niższa: ponieważ kryterium może osiągać wysokie wartości nawet dla bardzo małego $\ln L$, kiedy parametrów jest wiele, kryterium to w sposób oczywisty faworyzuje modele prostsze. Ta oczywistość niekoniecznie się sprawdza w analizie filogenetycznej prowadzonej techniką ML – dla rzeczywistych danych *AIC* jest niemal zawsze niższe dla modeli bardziej złożonych (Nei i Kumar 2000), co jednak nie jest dowodem ich wyższości. Russo i inni (1996) wykazali, że właściwie brak korelacji pomiędzy *AIC* a prawdopodobieństwem uzyskania prawidłowej topologii. Gaut i Lewis (1995) stwierdzili, że złożony model niekoniecznie daje lepsze wyniki niż prostszy, podobnie Yang (1997). Takahashi i Nei (2000) wygenerowali „sztuczne” dane: wykorzystując model *HKY* + Γ , uzyskali 48 sekwencji, każda o długości 1000 nukleotydów. Drzewa ML obliczono wykorzystując ten sam model *HKY* + Γ , a także nieodpowiedni dla tych danych, najprostszemu modelowi *JC*. Co znamienne, choć wiarygodności drzew obliczonych przy wykorzystaniu prawidłowego, złożonego modelu były znacznie wyższe od obliczonych dla modelu prostego, to jednak właśnie topologie drzew znalezionych, przyjmując prostszy model, były bliższe topologiom poprawnym. Tak więc przynajmniej w przypadkach wielu sekwencji, które są krótkie, model prostszy daje zwykle bardziej spójne rekon-

strukcje. Warto zauważyć, że stawia to pod znakiem zapytania użyteczność zarówno kryteriów wyboru modelu, jak i całej techniki ML.

Tworząc i wykorzystując coraz to bardziej złożone modele ewolucji sekwencji, dąży się do jak najwierniejszego odzwierciedlenia rzeczywistego procesu. Rzeczywistość jednak pozostaje i tak daleko bardziej złożona, a estymować się musi coraz więcej parametrów, każdorazowo z pewnym błędem, tak więc im bardziej złożony model, tym większa wariancja wyników. Szereg prób wykorzystania bardzo skomplikowanych modeli, w nadziei uzyskania lepszych rekonstrukcji (m.in. McArthur i Koop 1999, Silberman i inni 1999, Zardoya 1999), przyniosło wyniki nieraz zupełnie niezadowolające. Navidi i inni (1991) omawiają wybór modelu, Posada i Crandall (1998) napisali program, estymujący parametry modelu substytucji na podstawie analizowanych sekwencji. Szczególnych trudności nastroczają modele $+G$. Yang (1994b) stworzył technikę ML estymacji parametru α obliczanego wcześniej techniką redukcjonistyczną, jednak później (Yang 1996) wykazał, podobnie jak Gu i Zhang (1997), że estymaty zmieniają się w szerokim zakresie w zależności od przyjętego modelu substytucji. W sumie wydaje się, że najlepiej wykorzystywać modele jak najprostsze, zwiększając liczbę uwzględnianych parametrów z wielką ostrożnością, jak też traktując z rezerwą formalne kryteria dobroci rekonstrukcji parametrów i drzewa.

Zalety, wady i ograniczenia użyteczności ML

Jak już pisaliśmy, techniki ML opierają się na dokładnie sformułowanym modelu, więc różne parametry modelu możemy odpowiednio dobrać i założenia te są jasno określone, co sprzyja spójności rekonstrukcji, gdy w pozostałych metodach dopiero właściwości techniki pozwalają wnioskować, jakich założeń wymaga – niejednokrotnie więc badacz zapomina o tych założeniach, ponadto założenia można jedynie w pewnym stopniu modyfikować, dostosowując do konkretnego przypadku, dla którego rekonstruujemy filogenezę. To są plusy. Z drugiej strony, nie zawsze dysponujemy odpowiednim modelem – to choćby niemal zupełnie wyklucza stosowanie ML dla danych morfologicznych. Oczywiście parametry modelu na ogół nie są dokładnie znane, a estymując je na podstawie drzewa, które skonstruowaliśmy, wykorzystując te parametry, łatwo wpadamy w logikę błędnego koła, podobnie jak w omawianej już technice redukcjonistycznej. Iteratywne estymowanie drzewa i parametrów jest zarazem niezwykle intensywne obliczeniowo.

Jak i w innych technikach rekonstrukcji filogenezy, jakość rekonstrukcji ML w znacznym stopniu zależy od tego, jak silny jest filogenetyczny sygnał w analizowanych danych. Ocenic to można techniką **analizy względnej uwidaczniającej się synapomorfii** (*relative apparent synapomorphy analysis* – RASA), niezależnej od topologii rekonstruowanego drzewa, zaproponowanej przez Lyonsa-Weilera i innych (1996). Aby sprawdzić, czy cały zestaw danych wykazuje filogenetyczny sygnał (t_{RASA}), obserwowane tempo wzrostu podobieństwa kladystycznego między parami taksonów na jednostkę fenetycznego podobieństwa pomiędzy taksonami ($\beta_{\text{obserwowana}}$) porównywana jest z prostą regresji β_0 , obliczoną dla sytuacji, w której kladystyczny sygnał i fenetyczne podobieństwa rozmieszczone są losowo między parami taksonów (Lyons-Weiler i inni 1996).

Zdaniem szeregu autorów ML przewyższa inne techniki, nawet dla krótkich sekwencji (Hasegawa i Fujiwara 1993, Kuhner i Felsenstein 1994, Huelsenbeck 1995, Wilke i inni 2001), odznaczając się mniejszą wariancją wyników. Znaczy to, że często

wynik mniej zależy od losowych błędów w danych czy też generowanej losowo niepełnej reprezentatywności dla taksonów, których filogenezę chcemy zrekonstruować, tej właśnie części danych, którą mamy możliwość analizować. Podobnie bywa niezbyt wrażliwa na niespełnianie niektórych założeń modelu (Swofford i inni 1996), zwłaszcza że przynajmniej część założeń jest na ogół spełniana przynajmniej przez większość danych. Warto wspomnieć, że Cavalli-Sforza i Edwards zaproponowali metodę redukcyjną jako przybliżenie ML, obliczeniowo znacznie prostsze, więc nie może dziwić znaczne podobieństwo tych dwóch technik i dawanych przez nie wyników.

Należy przypomnieć, że przedstawione wcześniej formuły obliczania wiarygodności nie uwzględniają topologii drzewa, skądinąd kluczowej dla rekonstrukcji filogenezy. W oryginalnym ujęciu Cavalli-Sforzy i Edwardsa (1967) topologia traktowana była jako zmienna losowa, co słabo oddaje proces specjacji – kladogenezę, czyli tworzenie gałęzi estymowanego drzewa. Rannala i Yang (1996) oraz Yang i Rannala (1997) próbowali stworzyć bardziej realistyczny model specjacji, lecz wynik wciąż niedostatecznie odzwierciedla rzeczywiste procesy ewolucyjne. Współcześnie nadal brak techniki ML estymującej topologię drzewa (Yang i inni 1995, Nei 1987, 1996). Podobnie jak w omówionej wcześniej metodzie najmniejszych kwadratów, milcząco zakłada się, że najwyższe wartości wiarygodności będzie mieć drzewo o zoptymalizowanej długości gałęzi, mające topologię najbardziej zgodną z danymi i modelem; tak jednak bynajmniej być nie musi, można też znaleźć przykłady danych, dla których inne techniki dadzą lepsze topologie. Fukami i Tateno (1989) wykazali, że ML daje globalnie optymalne rekonstrukcje, jednak Steel (1994b) zakwestionował ich dowód, wykazując w nim błąd i przedstawiając przykład przeczący temu stwierdzeniu. Podobnie Rogers i Swofford (1999), Olsen i inni (1994), Saitou (1988) oraz Takahashi i Nei (2000) wykazali istnienie więcej niż jednego optimum, choć wymienieni badacze różnili się w ocenie, jak bardzo poważnym problemem jest to w przypadku prawdziwych, nie symulacyjnych danych.

Szczególnych problemów spodziewać się trzeba w przypadkach, gdy częstość substytucji zmienia się, zależnie od pozycji w sekwencji i/lub gałęzi drzewa. Gaut i Lewis (1995) wykazali niespójność ML, gdy model zakłada homogeniczność, a proces przebiega różnie w różnych pozycjach. Gdy analizujemy kodujące białka fragmenty sekwencji, dla odległych taksonów sekwencje protein dają lepsze rekonstrukcje niż sekwencje kwasów nukleinowych (Reeves 1992, Russo i inni 1996). Alternatywnie analizować można wówczas sekwencje DNA, przyjmując model ewolucji o 61 stanach (Muse i Gaut 1994, Goldman i Yang 1994). Gdy częstość substytucji jest zdecydowanie różna dla różnych gałęzi, to błędne topologie ML mogą być częstsze niż prawidłowe, nawet gdy sekwencje są długie (wysokie wartości n) (Huelsenbeck 1995). Pamiętać też warto, że warunki ciągłości i różniczkowalności funkcji wiarygodności, zapewniające asymptotyczne własności estymatorów ML, nie są spełniane w rekonstrukcji filogenezy (Yang i inni 1995).

Kolejnym problemem są same algorytmy estymujące wiarygodność. Intensywność obliczeniowa sprawia, że poszukuje się coraz to szybszych, bardziej wydajnych algorytmów. Tymczasem różne algorytmy użyte dla tych samych danych, nawet gdy przyjęty model jest ten sam, dają nieco inne wartości wiarygodności. Wprawdzie względne wartości pozostają podobne – zależność między wartościami dla tych samych drzew zoptymalizowanych różnymi algorytmami ma zwykle charakter funkcji monotonicznej – jednak niezgodności te mogą być problemem, zwłaszcza gdy taksonów T jest wiele. W miarę

wzrostu T różnice wartości L dla różnych rekonstrukcji stają się coraz mniejsze, więc potencjalnie wynik analizy filogenetycznej zależeć może od użytego algorytmu.

Powyższe uwagi krytyczne nie są podstawą do odrzucenia ML jako techniki rekonstrukcji filogenezy, zwłaszcza że wiele z zastrzeżeń dotyczy i innych metod. Są jednakże wystarczające, aby zakwestionować – jakże częste – rutynowe i niejednokrotnie bezkrytyczne stosowanie tej metody, traktowanej jako najlepsza, dająca najbardziej spójne i pewne rekonstrukcje. Niewątpliwie ML nadaje się do optymalizacji długości gałęzi drzewa, porównywania alternatywnych hipotez ewolucyjnych i estymacji parametrów procesu ewolucyjnych zmian stanów cech. Do samej rekonstrukcji filogenezy należy jej jednak używać z dużą ostrożnością i raczej zawsze równolegle do innych technik, jak redukcjonistyczna czy minimalnej ewolucji.

4.7. Analiza spektralna

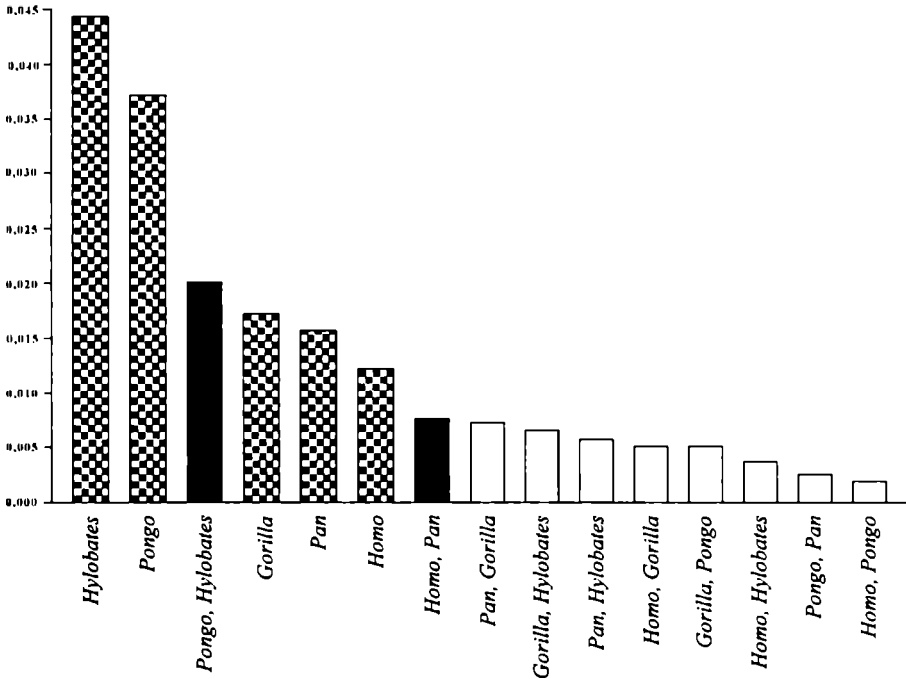
Analiza spektralna (*spectral analysis* – SA) pozwala na rekonstrukcję drzewa, ale też i na wskazanie konfliktu w danych, jak również ocenę, w jakim stopniu znaleziona najlepsza rekonstrukcja jest rzeczywiście lepsza od pozostałych. Opiera się na analizie **cząstkowych rekonstrukcji** (*splits*), które zdefiniowaliśmy w Rozdziale 4.2, omawiając technikę rozkładu cząstkowych rekonstrukcji. Metodę wprowadzili Hendy (1991) oraz Hendy i Penny (1993); Lento i inni (1995) przedstawiają dobre wprowadzenie do SA. Analiza spektralna jest niezbyt często stosowana, a aparat matematyczny dość skomplikowany, przedstawimy ją więc jedynie w zarysie.

Dla T taksonów istnieje 2^{T-1} rekonstrukcji cząstkowych. Konwencjonalnie (Hendy i Penny 1993) numerujemy je od 0 do $2^{T-1} - 1$. Dla konkretnego nieukorzonego drzewa liczba ta jest jednak niższa, równa $2T - 3$. Warto zauważyć, że z tej liczby jedna rekonstrukcja to wszystkie taksony razem, a T rekonstrukcji to pojedyncze taksony terminalne, przeciwstawione całej reszcie – takie rekonstrukcje dostarczają oczywiście informacji o odrębności poszczególnych taksonów, jednak niczego nie wnoszą do obrazu pokrewieństw. Analiza spektralna jest ograniczona do cech binarnych; istnieją rozszerzenia SA na cechy o większej liczbie stanów (omawiają je Swofford i inni 1996), lecz wymaga to przekodowania stanów cech na zestaw stanów binarnych, co zdecydowanie komplikuje obliczenia. Dla każdej rekonstrukcji cząstkowej każdy z taksonów albo należy (1), albo nie należy (0) do danej rekonstrukcji. Dla pięciu taksonów mamy więc następujące rekonstrukcje cząstkowe:

takson 1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
takson 2	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
takson 3	0	0	0	0	1	1	1	1	0	0	0	1	1	1	1	1
takson 4	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
takson 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
numer																
rekonstrukcji	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

W najprostszej formie SA przedstawia, w jakim stopniu dane uzasadniają wyodrębnianie każdej z możliwych rekonstrukcji cząstkowych, choć oczywiście nie każda z nich istnieje na danym drzewie. Mając macierz stanów cechy binarnej, liczymy, ile razy występuje każda z rekonstrukcji cząstkowych, przeliczając wyniki na częstości.

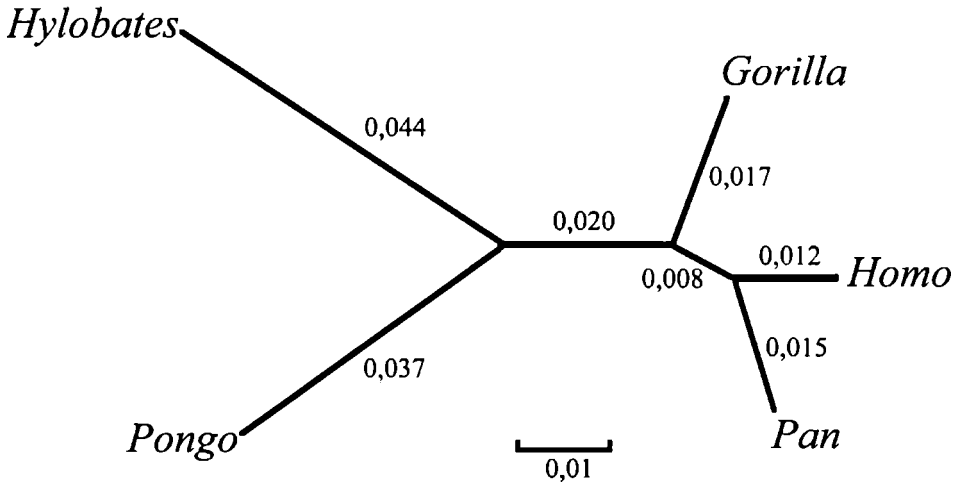
Oczywiście liczymy też cechy, których stany są identyczne dla wszystkich taksonów (rekonstrukcja 0) – jeżeli analiza dotyczy cech molekularnych (jak np. sekwencji DNA), to większość cech (pozycji) przemawiać będzie rzecz jasna za tą rekonstrukcją, więc częstości występowania pozostałych rekonstrukcji cząstkowych będą zwykle bardzo małe. Obliczone częstości przedstawiamy w postaci **spektrum**, czyli diagramu słupkowego, szeregując je od największych po najmniejsze. Wygodnie jest posłużyć się tu przykładem SA, przeprowadzonej dla sekwencji mtDNA współczesnych przedstawicieli *Hominidae* (Ryc. 4.18), zamieszczonym przez Page'a i Holmesa (1998).



Ryc. 4.18. Spektrum częstości cząstkowych rekonstrukcji na podstawie mtDNA współczesnych *Hominidae*. Kratkowane słupki przedstawiają częstości dla taksonów terminalnych przeciwstawionych reszcie – mogą być one miarą odrębności taksonu, lecz nie niosą informacji pozwalającej na rekonstrukcję pokrewieństw. Na czarno zaznaczono częstości dla najlepiej uzasadnionych węzłów, na biało – dla węzłów, których nie da się pogodzić z topologią zawierającą przedstawione czarnymi słupkami węzły (*Homo* – człowiek, *Hylobates* – gibbon, *Gorilla* – goryl, *Pan* – szympan, *Pongo* – orangutan). Wg Page'a i Holmesa (1998)

Jak widać na spektrum, najbardziej uzasadnione są rekonstrukcje wyodrębniające gibbona i orangutana, dobrze uzasadnione są też te, które oddzielają goryla, szympana i człowieka od pozostałych taksonów. Większość ewolucyjnych zmian sekwencji miała więc miejsce już po oddzieleniu tych taksonów. Z pozostałych 10 rekonstrukcji wszystkie są w jakimś stopniu uzasadnione – dla każdej częstość występowania w danych jest większa od zera. Dla pięciu taksonów $2T - 3 = 7$, od tego odjąć musimy

pięć rekonstrukcji filogenetycznie pozbawionych informacji (oddzielających kolejne taksony terminalne od całej reszty drzewa), tak więc dla nieukorzonego drzewa istnieją jedynie dwie filogenetycznie informatywne rekonstrukcje, uzasadniające utworzenie wewnętrznego węzła na tym drzewie. W idealnym przypadku dane powinny w takim samym stopniu uzasadniać istnienie dwóch węzłów, mogących się znaleźć na tym samym drzewie; tak jednak bywa rzadko, jeśli kiedykolwiek. W naszym przypadku najlepiej uzasadniony jest węzeł orangutan-gibbon, musi się więc znaleźć na zrekonstruowanym drzewie. Węzeł człowiek-szympan jest jedynie nieznacznie lepiej poparty danymi niż węzeł szympan-goryl, więc ta część rekonstrukcji jest mniej pewna; dość wysokie poparcie węzłów szympan-goryl, goryl-gibbon, itd. nie daje się natomiast pogodzić z topologią, określoną przyjętymi węzłami o największym uzasadnieniu danymi. Na tej podstawie rekonstruujemy więc drzewo (Ryc. 4.19), zwane **najbliższym drzewem** (*closest tree*: rozumiane jako drzewo najbliższe analizowanym danym). Częstości cech popierających określone rekonstrukcje cząstkowe są tu proporcjonalnymi długościami gałęzi.

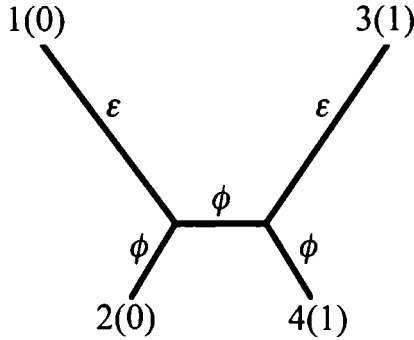


Ryc. 4.19. Najbliższe drzewo dla danych z Ryc. 4.18. Wartości przy gałęziach, proporcjonalne do ich długości, to częstości z Ryc. 4.18 (*Homo* – człowiek, *Hylobates* – gibbon, *Gorilla* – goryl, *Pan* – szympan, *Pongo* – orangutan). Wg Page'a i Holmesa (1998)

Przy większej liczbie taksonów analiza spektralna nie dałaby się przeprowadzić w przedstawiony sposób, bowiem liczba możliwych rekonstrukcji cząstkowych rośnie gwałtownie, osiągając ponad pół miliona dla 20 sekwencji. To stanowi ograniczenie stosowania SA, zaś już dla kilku taksonów proste liczenie byłoby niemożliwe. Istnieje jednak aparat matematyczny zwany **sprzężeniem Hadamarda** (*Hadamard conjugation*), znany też jako **transformacja Hadamarda** (*Hadamard transform*), wprowadzony przez Hendy'ego i Penny'ego (1993), a w uproszczeniu omówiony przez Swofforda i innych (1996). Tu przedstawimy jedynie najogólniejszy zarys metody, w ujęciu Swofforda i innych (1996), zwłaszcza że jej zrozumienie wymaga znajomości rachunku macierzowego. **Macierz Hadamarda H** to kwadratowa macierz o elementach 1 i -1 , a każdy wiersz

jest ortogonalny do pozostałych wierszy i każda kolumna ortogonalna do pozostałych kolumn:

$$\mathbf{H}^{(1)} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \mathbf{H}^{(n+1)} = \begin{pmatrix} \mathbf{H}^{(n)} & \mathbf{H}^{(n)} \\ \mathbf{H}^{(n)} & -\mathbf{H}^{(n)} \end{pmatrix}.$$



Ryc. 4.20. Czterotaksonowe drzewo identyczne z drzewem z Ryc. 4.15a; ε i ϕ to prawdopodobieństwa zmiany stanu cechy odpowiednio na długiej i krótkiej gałęzi

Dla prostego czterotaksonowego drzewa (Ryc. 4.20; powtórzone tu dla wygody Czytelnika, jest drzewem z Ryc. 4.15a) macierz Hadamarda liczy $2^{T-1} = 8$ wierszy i kolumn, a 2^{T-1} -elementowy wektor \mathbf{p} przedstawia prawdopodobieństwa ε i ϕ zmiany stanu cechy na określonej gałęzi, a więc określonych rekonstrukcji cząstkowych. Przyjmijmy: $\varepsilon = 0,4$, $\phi = 0,1$; p_0 uznaje się za równe 0, a rekonstrukcje cząstkowe 6 i 7 nie istnieją na tym drzewie ($2T - 3 = 5$, a więc wektor zawiera 5 niezerowych wartości):

$$\text{macierz } \mathbf{H} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}, \text{ wektor } \mathbf{p} = \begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ \dots \\ p_{2^{T-1}-1} \end{pmatrix} = \begin{pmatrix} 0 \\ \varepsilon \\ \phi \\ \phi \\ \varepsilon \\ 0 \\ 0 \\ \phi \end{pmatrix} = \begin{pmatrix} 0 \\ 0,4 \\ 0,1 \\ 0,1 \\ 0,4 \\ 0 \\ 0 \\ 0,1 \end{pmatrix}.$$

Przypomnijmy, że w naszym przykładzie korzystamy z prostego modelu Cavendera i Felsensteina (1987), dającego się użyć dla kwasów nukleinowych – gdy przyjmiemy np. puryny (A i G) za 0, a pirymidyny (C i T) za 1 – zmiany odpowiadać więc będą wyłącznie transwersjom. Model ten można traktować wówczas jako uproszczoną, dwustanową wersję modelu Jukesa i Cantora (1969). Wówczas wektor obserwowanych różnic \mathbf{p} można przekształcić w wektor spodziewanych łącznych zmian stanu cechy

przypadających na pozycję sekwencji, zgodnie ze wzorem, będącym specjalnym przypadkiem (gdy $B = 1/2$) wzoru na odległość między sekwencjami dla modelu F81:

$$q_i = -1/2 \ln(1 - 2p_i),$$

gdzie q_i to spodziewana liczba zmian przypadająca na pozycję w obrębie gałęzi i . Obliczone wartości q_i pozwalają zdefiniować wektor $\gamma(T)$, czyli **spektrum długości gałęzi** (*branch-length spectrum*):

$$\gamma(T) = \begin{pmatrix} -\sum_{i=1}^{m-1} q_i \\ q_1 \\ q_2 \\ \dots \\ q_{m-1} \end{pmatrix} = \begin{pmatrix} -1,94415 \\ 0,80472 \\ 0,11157 \\ 0,11157 \\ 0,80472 \\ 0 \\ 0 \\ 0,11157 \end{pmatrix}.$$

Następnie definiujemy wektor $s(T)$, czyli wektor **spodziewanego spektrum sekwencji** (*expected sequence spectrum*), którego elementami s_k są przewidywane częstości cech o stanach przemawiających za wyodrębnieniem każdej możliwej **bipartycji**, czyli podziału taksonów na dwie wykluczające się grupy; przypominając drzewo przedstawione na Ryc. 4.17, s_3 odpowiada tam $P_{0011} + P_{1100}$, a s_5 odpowiada $P_{0101} + P_{1010}$. Dla obliczeń posłużymy się **sprzężeniem Hadamarda**:

$$s(T) = \mathbf{H}^{-1} \exp[\mathbf{H}\gamma(T)],$$

gdzie funkcja wykładnicza stosowana jest oddzielnie dla każdego elementu $\mathbf{H}\gamma$, a \mathbf{H}^{-1} to odwrotność macierzy \mathbf{H} , czyli macierz skonstruowana tak, że iloczyn macierzy $\mathbf{H}\mathbf{H}^{-1} = \mathbf{I}$, gdzie \mathbf{I} to macierz identyczności, złożona z jedynek na przekątnej i zer w pozostałych pozycjach. Dla danych z naszego przykładu zaczniemy od obliczenia wektora uogólnionych odległości ρ , a następnie uogólnionych odległości \mathbf{r} . Odległości są *uogólnione* w tym sensie, że każda z nich odzwierciedla nie tylko odległości między parami taksonów, lecz również między rozłącznymi grupami, z których każda zawiera parzystą liczbę taksonów. Każdy z elementów wektora ρ odpowiada $-2\delta_i'$, gdzie δ_i' to **skorygowana uogólniona odległość**, a każdy z elementów wektora \mathbf{r} to **obserwowana uogólniona odległość**: $r_i = 1 - 2d_i'$ (Swofford i inni 1996). Dla naszych danych:

$$\boldsymbol{\rho} = \mathbf{H}\boldsymbol{\gamma} = \begin{pmatrix} 0 \\ -2,05572 \\ -0,66942 \\ -1,83258 \\ -1,83258 \\ -3,44202 \\ -2,05572 \\ -3,66516 \end{pmatrix}, \quad \mathbf{r} = \exp(\boldsymbol{\rho}) = \begin{pmatrix} e^{\rho_0} \\ e^{\rho_1} \\ e^{\rho_2} \\ e^{\rho_3} \\ e^{\rho_4} \\ e^{\rho_5} \\ e^{\rho_6} \\ e^{\rho_7} \end{pmatrix} = \begin{pmatrix} 1 \\ 0,12800 \\ 0,51201 \\ 0,16000 \\ 0,16000 \\ 0,03200 \\ 0,12800 \\ 0,02560 \end{pmatrix}.$$

Odwrotność macierzy Hadamarda (tu o 2^{T-1} rzędów i kolumn) ma prostą postać, co jest wygodne:

$$\mathbf{s}(T) = \mathbf{H}^{-1}\mathbf{r} = \left(\frac{1}{2^{T-1}} \mathbf{H} \right) \mathbf{r}; \text{ podstawiając nasze dane uzyskujemy:}$$

$$\mathbf{s}(T) = \frac{1}{8} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0,12800 \\ 0,51201 \\ 0,16000 \\ 0,16000 \\ 0,03200 \\ 0,12800 \\ 0,02560 \end{pmatrix} = \begin{pmatrix} 0,26820 \\ 0,18180 \\ 0,06180 \\ 0,06820 * \\ 0,18180 \\ 0,12420 * \\ 0,05220 \\ 0,06180 \end{pmatrix} = \begin{pmatrix} P_{0000} + P_{1111} \\ P_{1000} + P_{0111} \\ P_{0100} + P_{1011} \\ P_{1100} + P_{0011} \\ P_{0010} + P_{1101} \\ P_{1010} + P_{0101} \\ P_{1001} + P_{0110} \\ P_{1110} + P_{0001} \end{pmatrix}.$$

Gwiazdkami oznaczyliśmy wartości obliczone dla przypadku przedstawionego na Ryc. 4.17 – jak widać, tamte algebraiczne obliczenia dały te same wartości co wyliczone sprzężeniem Hadamarda. Tu jednak obliczono prawdopodobieństwa dla wszelkich możliwych bipartycji, ponadto metodę można uogólnić na obliczenia rozkładu stanów cech, przyjmując również bardziej realistyczne – a więc bardziej złożone – modele, a także na większą liczbę taksonów, choć już dla 10 taksonów wektor liczy 512 elementów, dla 15 – 16 384 elementy, dla 20 – 524 288 elementów, a dla 25 – 16 777 216 elementów, więc dość szybko natrafimy na kres możliwości obliczeniowych komputera.

Kolejną zaletą sprzężenia Hadamarda jest jego odwracalność. Jeżeli przez $\hat{\mathbf{s}}$ oznaczymy obserwowane spektrum sekwencji, przez $\hat{\boldsymbol{\gamma}}$ estymat spektrum $\boldsymbol{\gamma}$ długości gałęzi drzewa, a spodziewaną liczbę zmian na gałąź i przez \hat{p}_i , to sprzężenie Hadamarda przedstawić możemy w postaci:

$$\hat{\gamma} = \mathbf{H}^{-1} \ln(\mathbf{H} \hat{\mathbf{s}}) = \mathbf{H}^{-1} \mathbf{p} = \left(\frac{1}{2^{T-1}} \right) \mathbf{p}, \text{ a długości gałęzi: } \hat{p}_i = \frac{1 - e^{-2q_i}}{2},$$

tak więc mając spektrum sekwencji, możemy estymować długości gałęzi. W praktyce, mając większy zestaw taksonów i zwykle niezbyt jednoznaczne dane, po obliczeniu wektora $\hat{\gamma}$ wykorzystać musimy dla znalezienia drzewa jedną z paru dostępnych metod. Wspomniana już technika najbliższego drzewa (Hendy 1991), często stosowana w takim przypadku, formalnie polega na minimalizacji odległości Euklidesowej pomiędzy poszukiwanym wektorem $\mathbf{q}(\tau)$ a wektorem $\hat{\gamma}$, gdzie τ to dane drzewo o K gałęziach. Minimalizujemy wielkość daną wzorem:

$$\Gamma^2(\tau, \hat{\gamma}) = \sum_{e_i \notin e(\tau)} \hat{\gamma}_i^2 + \frac{\left(\hat{\gamma}_0 + \sum_{e_i \in e(\tau)} \hat{\gamma}_i \right)^2}{K+1},$$

gdzie $e_i \notin e(\tau)$ to gałęzie nienależące do badanego drzewa τ , a $e_i \in e(\tau)$ to gałęzie należące do tego drzewa (Hendy i Penny 1993). Najbliższe drzewo jest drzewem o najniższej wartości tak określonego parametru, znajdowane jednym z przedstawionych wcześniej (Rozdział 4.3) algorytmów poszukiwania drzewa o najlepszej wartości przyjętego kryterium optymalizacji. Zdarzyć się może, że jakieś wartości wektora $\hat{\gamma}$ będą ujemne, w następstwie błędów losowych w danych lub niepełnego spełniania założonego modelu – wszystkie drzewa zawierające takie wartości należy wówczas odrzucić.

Drzewo znaleźć można też skorygowaną (ważoną) techniką redukcjonistyczną, traktując spektrum $\hat{\gamma}$ jako transformację oryginalnej macierzy danych w nową macierz liczącą 2^{T-1} cech (gdy wyjściowe dane są binarne), z których każda odpowiada cząstkowej rekonstrukcji odpowiadającej określonemu wierszowi wektora $\hat{\gamma}$, a wartości tego wektora traktowane są jako wagi cech, przy czym występujące niekiedy wartości ujemne traktujemy jako równe 0 (Swofford i inni 1996). Metoda ta dla modelu Cavendera i Felsensteina daje zawsze spójne rekonstrukcje (Steel i inni 1993). Symulacje przeprowadzone przez Charlestona (1994) wskazują, że technika jest bardzo wydajna w niektórych sytuacjach, a ogólnie przewyższa metodę najbliższego drzewa, jak i inne techniki znajdowania drzewa w SA (Swofford i inni 1996).

Drzewo można też znaleźć metodą ważonej kompatybilności cech, poszukującej największej ważonej grupy (*clique*) cech kompatybilnych, o których mówiliśmy, omawiając ważenie cech w technice redukcjonistycznej (Rozdział 4.5). Można też wykorzystać technikę, łączącą metodę najbliższego drzewa z techniką kompatybilności cech. Przy zgodności danych z modelem i braku błędów losowych, obniżających represen-

tatywność analizowanych danych, $2T - 3$ elementów wektora $\hat{\gamma}$ będzie miało wartości dodatnie, gdy pozostałe (poza $\hat{\gamma}_0$) – czyli reprezentujące gałęzie nieobecne na rozpatrywanym drzewie – będą równe zeru. A więc suma kwadratów odchyłeń od zera, obliczonych dla elementów wektora $\hat{\gamma}$ nieobecnych na rozpatrywanym drzewie – odpowiadająca pierwszemu elementowi prawej strony równości formułującej kryterium najbliższego drzewa – jest miarą niezgodności drzewa z danymi i modelem, estymowaną techniką najmniejszych kwadratów:

$$\Delta^2(\tau\gamma) = \sum_{e, e \in \tau} \hat{\gamma}_e^2,$$

wartość tę należy więc minimalizować. Zdaniem Waddella (1995) drugi element sumy prawej strony równania kryterium najbliższego drzewa znacznie mniej wpływa na wartość kryterium wartościującego drzewo, więc pozwala to na użycie powyższej równości, której minimalizacja to właśnie wykorzystanie metody kompatybilności cech; tu znów wartości ujemne uznajemy za zerowe (Swofford i inni 1996). Wazenie wartości wektora $\hat{\gamma}$ estymowanym błędem losowym zwiększa dobroć techniki (Waddell 1995, Swofford i inni 1996).

Oprócz znajdowania najlepszego drzewa technika SA umożliwia też analizę samych danych użytych do rekonstrukcji, jak ocenę stopnia niezgodności z modelem, który może być za prosty (nie uwzględniając różnej częstości zmian dla różnych pozycji, przyjmując nieodpowiedni model substytucji), albo wykrywanie występowania rekombinacji lub braku niezależności zmian pomiędzy pozycjami. Metody uwidaczniania konfliktów w obrębie danych i oceny jakości filogenetycznego sygnału omawiają Lento i inni (1995), Waddell i inni (1994), Waddell (1995) oraz Swofford i inni (1996). Istnieją też techniki umożliwiające SA dla cech o czterech stanach, choć są one bardziej intensywne obliczeniowo (4^{T-1} zamiast 2^{T-1}), a także dopuszczające różne częstości substytucji dla różnych pozycji (Swofford i inni 1996), co jeszcze bardziej komplikuje i aparat matematyczny, i same obliczenia. Wówczas:

$$s(T) = \mathbf{H}^{-1}[(\alpha - \rho)/\alpha]^{-\alpha} [\mathbf{H}\gamma(T)],$$

gdzie α to znany nam już parametr, określający kształt rozkładu Γ . Transformację Hadamarda zastosować możemy również dla macierzy odległości między parami taksonów (Hendy i Penny 1993, Swofford i inni 1996), a nie bezpośrednio dla sekwencji, estymując wówczas spektrum długości gałęzi $\hat{\gamma}_D$ i znajdując najlepsze drzewo dla tego spektrum. Odległości wprowadzamy jako elementy wektora uogólnionych odległości ρ , choć problemem jest wówczas występowanie w tym wektorze również odległości między grupami taksonów – estymujemy je metodą Hendy’ego i Penny’ego (1993). Symulacje wykazały, że wariancje elementów wektora obliczonych tą techniką są mniejsze od wy-

liczanych bezpośrednio dla sekwencji, bowiem mniejszą wariancję mają estymowane odległości między grupami taksonów (Waddell 1995, Swofford i inni 1996). Drzewa obliczone na podstawie odległości są więc bardziej wiarygodne (Charleston 1994), choć z drugiej strony ta odmiana transformacji Hadamarda jest mniej czuła w stosunku do niezgodności z modelem, słabiej je wykazuje. Zapewne warto jej użyć jako pomocy w wyborze odpowiedniej odległości dla naszych danych.

4.8. Specjalne odmiany metody kładystycznej

Przedstawimy tu trzy techniki, które stosowane bywają nieczęsto, więc omówimy je w dużym skrócie. Pierwsza z nich zaproponowana została dla sekwencji kwasów nukleinowych i z klasyczną techniką kładystyczną łączy ją głównie nazwa, zapewne nie będzie ona w przyszłości stosowana częściej. Druga wykorzystuje „normalne” programy rekonstruujące filogenezę techniką redukcjonistyczną, odmiennie natomiast koduje dane, opierając się na nieco innym ich rozumieniu; wprowadzona niedawno i budząca kontrowersje, być może będzie szerzej wykorzystywana w przyszłości. Ostatnia to zupełna nowość, w pewnych sytuacjach okazać się może przydatna.

Metoda niezmiennych Lake’a

Metoda niezmiennych Lake’a (*Lake's method of invariants* – MI), zwana też przez jej autora **ewolucyjną techniką redukcjonistyczną** (*evolutionary parsimony*), przedstawiona została przez Lake’a (1987), omawiają ją też i rozwijają Cavender i Felsenstein (1987), Cavender (1989), Jin i Nei (1990), Sankoff (1990), Sidow i Wilson (1990, 1992) oraz Navidi i inni (1991). Technikę stosuje się do analizy sekwencji kwasów nukleinowych, zakładając niezależność substytucji na określonej pozycji sekwencji, równowagę między określonymi rodzajami transwersji (wystarczy przyjąć, że równie prawdopodobne są obie transwersje dla określonej zasady: np. A do C i do T) oraz to, że insercje i delecje możemy bezpiecznie pominąć, jako nieobecne bądź występujące rzadko. Metoda dopuszcza natomiast odmienną częstość substytucji dla różnych pozycji.

Jak pamiętamy, nieważona technika redukcjonistyczna nie kompensuje zmian, które zaszły wzdłuż gałęzi, nawet dłuższej, zakładając jedną substytucję, gdy stan cechy uległ zmianie na tej gałęzi. MI próbuje kompensować nieobserwowane transwersje (pomijając znacznie przecież częstsze, choć filogenetycznie mniej ważne tranzycje), wykorzystując dla tej kompensacji tranzycje niepotwierdzające grupowania. Gdy przykładowo u taksonów A i B występuje G, u C jest T, a u D – G, a więc zakładamy, że miały miejsce dwie równoległe transwersje, to dla nieważonej techniki redukcjonistycznej taka (znana) sytuacja nie niesie informacji o grupowaniu C z D, natomiast w MI – takiemu grupowaniu przeczy. Z założenia obie transwersje są równie prawdopodobne, zatem częstość równoległych zmian na ten sam nuleotyd jest równa częstości równoległych zmian na różne nuleotydy. Tak więc odjęcie liczby transwersji dających różne nuleotydy od liczby wszystkich transwersji powinno dać liczbę homologicznych transwersji (synapomorfii).

Wybiera się kwartet współosiowanych sekwencji (A, B, C, D) i znajduje w nich pozycje, na których w dwóch sekwencjach występują puryny i w dwóch pirymidyny. Na-

stępnie rozważa się trzy możliwe grupowania dla każdej z tych pozycji: AB/CD (A razem z B, C razem z D) określone jako X, AC/BD – jako Y i AD/BC – jako Z. Wszystkie przypadki, gdy puryny występują jednocześnie u A i B, a pirymidyny u C i D oraz pirymidyny jednocześnie u A i B, a puryny u C i D (to znaczy pozycje potwierdzające grupowanie AB/CD, czyli X), oznaczamy jako X^+ , zaś pozostałe z wybranych wcześniej pozycji jako X^- (przeczące grupowaniu). Analogicznie obliczamy Y^+ i Y^- oraz Z^+ i Z^- . Suma tych sześciu wartości będzie równa łącznej liczbie znalezionych wcześniej pozycji. Obliczamy wartości $X = X^+ - X^-$, $Y = Y^+ - Y^-$ i $Z = Z^+ - Z^-$, które określają, w jakim stopniu dane potwierdzają kolejne grupowania sekwencji. Dla dwóch z grupowań wartości powinny być bliskie zeru, dla trzeciego – znacząco różne od zera. Lake (1987) zaproponował użycie testu χ^2 o jednym stopniu swobody, zgodnie z formułą: $\chi^2_X = X^2 / (X^+ + X^-)$ (analogicznie dla Y i Z), aby określić, czy wartość różni się istotnie od zera; dla małych wartości test nie jest odpowiedni i powinien być zastąpiony dokładnym dwumianowym (Swofford i inni 1996). Wartości mogą być negatywne, choć statystycznie istotne – zdaniem Lake’a potwierdzają one zasadność grupowania, lecz interpretacja taka jest dyskusyjna, bowiem wynik może wskazywać na selekcję lub inne nielosowe procesy (Swofford i inni 1996).

Technika, jak wiemy, nie uwzględnia tranzycji. Tak więc dla bliskich taksonów może być bezużyteczna, bowiem same tranzycje nie dają tu filogenetycznego sygnału. Z drugiej strony, tranzycje zawsze występują, tymczasem metoda zakłada, że całość obserwowanych zmian to następstwa transwersji. Często więc domniemane dwie równoległe transwersje bądź jedna transwersja przed rozdzieleniem gałęzi to w rzeczywistości wynik jednej transwersji i jednej lub więcej tranzycji bądź tranzycji poprzedzonej dwu równoległymi transwersjami. Jak długo tranzycji jest niewiele, czułość techniki spada, lecz da ona wciąż wynik spójny; gdy tranzycje są liczne, filogenetyczny sygnał jest przez nie zupełnie maskowany, tak że rozkład obserwowanych (domniemanych) transwersji staje się losowy (Li i inni 1987). Technika redukcjonistyczna ważona lub optymalizowana transwersjami (*transversion parsimony*: zamiast odejmowania sumuje ona X^+ z X^- , Y^+ z Y^- i Z^+ z Z^-) w wielu przypadkach dadzą więc lepsze wyniki (Hillis i inni 1994). Dla spójnych wyników MI wymaga bardzo długich sekwencji, co w praktyce na ogół wyklucza jej bezpieczne użycie. Hillis i inni (1994) wykazali symulacjami, że dla czterotaksonowego drzewa, modelu K2P i sytuacji, gdy występuje przyciąganie długich gałęzi, wymagana jest sekwencja długości 10^8 nukleotydów jedynie po to, aby prawdopodobieństwo znalezienia prawidłowej rekonstrukcji przewyższyło 1/3, czyli prawdopodobieństwo losowego wyboru takiego drzewa. W tych samych warunkach wystarczyło 5000 nukleotydów dla osiągnięcia prawdopodobieństwa 0,95, stosując technikę maksymalizacji wiarygodności, choć potencjalnie MI może być spójna także w warunkach, gdy ML zawodzi.

Stwierdzenie o trzech taksonach

Stwierdzenie o trzech taksonach (*three-item statement* – TIS), znane też jako **analiza stwierdzeń o trzech taksonach** (*three-item statement analysis*), zaproponowane zostało przez Nelsona i Platnicka (1991) dla zwiększenia precyzji klasycznej techniki redukcjonistycznej (kladystycznej). Technika spotkała się ze zdecydowaną krytyką (Harvey 1992, Kluge 1993, 1994, Farris i inni 1995), na którą autorzy techniki raczej przekonująco odpowiedzieli (Nelson 1992, 1993, Nelson i Ladiges 1993), lecz

dostrzeżono też zalety TIS (de Pinna 1996, Kitching i inni 1998). U podstaw metody leży świadomość, że analiza filogenetyczna pozwala formułować jedynie hipotezy zarówno o następstwie przodek-takson potomny, jak i o ewolucji cech; co więcej, hipotez takich nie da się obiektywnie sprawdzać. Tym natomiast, co możemy wykazać i testować, są określane synapomorfiami kłady siostrzane (oczywiście wymaga to poprawnego rozpoznania homologii, jednak te ostatnie testujemy przeciw filogenezie). Zamiast więc – jak w tradycyjnej analizie kladystycznej (redukcjonistycznej) – zakładać niesprawdzalne hipotezy o seriach transformacyjnych, poprzestajemy na obserwowanym rozkładzie apomorficznych stanów cech u badanych taksonów i na tej podstawie taksony te grupujemy. W jakimś sensie TIS opiera się na najbardziej ortodoksyjnym rozumieniu taksonomii filogenetycznej, która przeciw zajmuje się badaniem związków między cechami, genami, taksonami czy zasięgami biogeograficznymi, odkrywaniem ich hierarchicznej struktury. Z drugiej strony TIS, jak zobaczymy, w określonych przypadkach zwiększa czułość rekonstrukcji, pozwalając na wskazanie pokrewieństw, których nie wykaże analiza kladystyczna w tradycyjnym ujęciu.

Zaczynamy od kodowania danych jako serii binarnych lub wielostanowych cech, odzwierciedlającego wstępnie rozpoznane homologie. Dla czterotaksonowego drzewa A, B, C, D, taksony C i D mają określony, apomorficzny stan cechy (1), a u taksonów A i B go brak (0). Każda z binarnych cech reprezentuje odrębną homologię, jest więc niezależna od pozostałych, natomiast kolejne stany cechy wielostanowej niezależne nie są, bowiem opisują tę samą homologię. TIS nie przedstawia danych jako zmiennych binarnych lub wielostanowych, lecz redukuje je do zestawu wyrażającego wszystkie związki pokrewieństwa, rozpisane dla kolejnych trójtaksonowych elementów analizowanego zestawu OTU. A więc dla naszego przykładu sformułować możemy dwa trójtaksonowe stwierdzenia: A(CD) i B(CD). Stwierdzają one, że taksony CD łączy związek, wyróżniający je od taksonu A i wyrażony obecnością stanu cechy, którego brak u A, a także odróżniający je od taksonu B – również pozbawionego tego stanu cechy. Połączenie tych dwóch stwierdzeń: A(CD) + B(CD) = AB(CD), czyli powstanie kladogram, na którym C z D utworzą kład, zaś A i B trychotomię z kładem (CD). Możemy więc stworzyć macierz, gdzie kolejnym taksonom przypisujemy kolejne stwierdzenia.

Pojawia się oczywiście problem, jak w naszym przykładzie potraktować stan cechy dla taksonu B w pierwszym, a A w drugim stwierdzeniu. Otóż znaczymy je znakiem zapytania (?). W ten sposób, gdy wyjściowa, „tradycyjna” matryca miała postać: A: 0, B: 0, C: 1, D: 1, to dla techniki TIS zapisujemy ją: A: 0?, B: ?0, C: 11, D: 11. Oczywiście zwiększa to rozmiary macierzy danych, która staje się mniej przejrzysta, a wielka liczba znaków zapytania, jak wiemy kodujących także polimorfizm bądź brakujące dane – jak o tym pisaliśmy, takich przypadków lepiej unikać – to zapewne główny powód wspomnianej, zdecydowanej krytyki tej techniki (Harvey 1992, Kluge 1993, 1994). Jak podkreślają jej twórcy (Nelson i Ladiges 1993), znaki zapytania interpretuje się tutaj jako „nie dotyczy”, choć programy komputerowe i tak mogą je interpretować w sposób, likwidujący na kladogramach politomie mimo braku dostatecznego oparcia w danych. W naszym prostym przykładzie tę samą wartość kryterium optymalizacji będą miały trzy kladogramy: (AB,(CD)), (A,(B,(CD))) i (B,(A,(CD))), wynik analizy przedstawić więc musimy w postaci **drzewa pełnej zgodności** (*strict consensus tree*: jak je obliczyć przedstawiemy w Rozdziale 4.10), na którym kład (CD) tworzy trychotomię z taksonami A i B.

Dla cechy binarnej liczba stwierdzeń o trzech taksonach zależy oczywiście od liczby analizowanych taksonów T oraz liczby taksonów n mających apomorficzny stan cechy, i dana jest zależnością: $(T - n)n(n - 1)/2$. Dla cechy wskazującej na pokrewieństwo ABC(DE) stwierdzeń jest trzy: A(DE), B(DE) i C(DE), a dla AB(CDE) – sześć: A(CD), A(CE), A(DE), B(CD), B(CE) i B(DE). Cecha wielostanowa w TIS odpowiada zestawowi trzystanowych stwierdzeń, z których każde występuje tylko raz (Nelson i Ladiges 1992). Dla uporządkowanej cechy wielostanowej przedstawiającej pokrewieństwo A(B(CD)) istnieją cztery stwierdzenia o trzech taksonach: A(BC), A(BD), A(CD) i B(CD), podczas gdy przekodowanie na dwie cechy binarne reprezentowane przez A(BCD) i AB(CD), wyrazić musimy pięciu stwierdzeniami: A(BC), A(BD), A(CD), A(CD) i B(CD), a więc o jedno stwierdzenie więcej, bowiem A(CD) występuje dwukrotnie, czyli waży podwójnie: tak więc przy przekodowaniu cechy trójstanowej na dwie binarne informacja filogenetyczna pozornie się zwiększa, czego wynikiem może być błędna rekonstrukcja filogenezy.

Pozostając przy czterotaksonowym drzewie ABCD, posłużmy się przykładem macierzy danych dla 10 cech binarnych, przedstawionym za Kitchingiem i innymi (1998). Kodowanie tradycyjne przypisuje kolejnym taksonom stany kolejnych cech:

	1	2	3	4	5	6	7	8	9	10
A	0	0	0	0	0	0	0	1	1	1
B	0	0	0	1	1	1	1	1	1	1
C	1	1	1	1	1	1	1	0	0	0
D	1	1	1	1	1	1	1	0	0	0

Na ich podstawie obliczyć można techniką redukcjonistyczną drzewo o długości 13 i topologii (A,(B,(CD))). Kodowanie tych samych danych techniką TIS to następujące stwierdzenia o trzech taksonach:

	1	2	3	4	5	6	7	1	2	3	4	5	6	7	4	5	6	7	8	9	10	8	9	10
A	0	0	0	0	0	0	0	?	?	?	0	0	0	0	0	0	0	0	1	1	1	1	1	1
B	?	?	?	?	?	?	?	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C	1	1	1	1	1	1	1	1	1	1	1	1	1	1	?	?	?	?	0	0	0	?	?	?
D	1	1	1	1	1	1	1	1	1	1	?	?	?	?	1	1	1	1	?	?	?	0	0	0

Jak widać, jest tu siedem stwierdzeń A(CD), trzy B(CD), cztery A(BC), cztery A(BD), trzy C(AB) i trzy D(AB), czyli razem 24. Najlepsza topologia będzie taka sama jak dla kodowania tradycyjnego i mieć długość 30. Z 24 stwierdzeń 18 znalazło się na kladogramie (topologia kladogramu je potwierdza) – oznaczamy je jako **włączone w kladogram** (*accommodated three-item statements* – ATS). Pozostałe sześć to **niewłączone w kladogram** (*nonaccommodated three-item statements* – NTS). Oczywiście – jak pamiętamy – w technice redukcjonistycznej najlepsze drzewo to drzewo o najniższej wartości kryterium optymalizacji; dla TIS takie drzewo powinno zawierać jak najwięcej ATS i jak najmniej NTS. Długość kladogramu określona jest w TIS zależnością:

$$\text{długość} = \sum \text{ATS} + 2 \sum \text{NTS},$$

a więc w naszym przykładzie $18 + 2 \times 6 = 18 + 12 = 30$. Pamiętajmy, że choć programy do obliczania drzew metodą redukcjonistyczną, jak PAUP (Swofford (1998) czy HENNIG86 (Farris 1988), z powodzeniem obliczają dane kodowane techniką TIS, bezbłędnie znajdując najlepsze drzewa, to jednak znaczenie „kroków” jest wówczas zupełnie inne, odzwierciedlając nie liczbę zmian stanów cech (ewentualnie ważonych), lecz właśnie liczby ATS i NTS.

Jak pamiętamy, technika redukcjonistyczna zakłada wzajemną niezależność cech i choć niezależność ta bywa nieraz wątpliwa, to nie powinniśmy używać razem cech, co do których wiemy, że niezależne być nie mogą. Tymczasem przy kodowaniu TIS niezależność z założenia nie zawsze ma miejsce. Bywa tak, gdy liczba taksonów mających apomorficzny stan cechy $n > 2$. Wystarczy prosty przykład związku A(BCD): opisać go należy trzema stwierdzeniami o trzech taksonach: A(BC), A(BD) i A(CD), lecz jakiegokolwiek dwa z tych trzech stwierdzeń opisują jednoznacznie ów związek, a więc jedynie dwa z trzech stwierdzeń są niezależne. Z ogólnej liczby $(T - n)n(n - 1)/2$ stwierdzeń o trzech taksonach, $(T - n)(n - 1)$ to stwierdzenia niezależne (Nelson i Ladiges 1992), czyli $2/n$ wszystkich stwierdzeń. Choć w pewnych przypadkach część stwierdzeń nie jest niezależna, brak podstaw do eliminacji któregośkolwiek z nich. Rozwiązaniem jest natomiast wprowadzenie **proporcjonalnego ważenia** (*fractional weighting*), którego nie należy mylić z omówionym w Rozdziale 4.5 ważeniem cech.

Logiczne jest przypisanie każdemu z trzech stwierdzeń opisujących A(BCD) wagi $2/3$, bowiem wszystkie trzy łącznie ważą tyle co dwa, jako że dwa wystarczają do opisu związku; dla odmiany, trzy stwierdzenia opisujące związek ABC(DE) mają wagi 1, bowiem żadnego z nich pominąć nie można, a więc są one niezależne. Optymalizacja w przypadku proporcjonalnego ważenia polega na znalezieniu kladogramu o najniższej łącznej wadze, niekoniecznie identycznego z włączającym największą liczbę ATS (Nelson 1993, Nelson i Ladiges 1994). Proporcjonalne ważenie omija niebezpieczeństwo pozornego wzrostu informacji o niektórych ze związków między taksonami, należy je więc stosować w TIS raczej z zasady. Wagi stwarzają jednak pewien problem: jak widzimy, obliczamy je jako proporcje, czyli ułamki, tymczasem programy komputerowe pracują w systemie dziesiętnym, wartości więc przybliżą i będzie to źródłem znaczącego błędu. Ponadto większość programów rekonstruujących filogenezę dopuszcza jedynie liczby całkowite jako wagi i długości drzewa, co błąd jeszcze zwiększy (np. $2^{2/3}$ będzie zastąpione wartością 3). Należy więc wszystkie wagi tak przemnożyć, aby ułamków uniknąć: zatem w naszym przykładzie po pomnożeniu przez 3 stwierdzenia dla A(BCD) ważyc będą po 2, a dla ABC(DE) – po 3.

Charakterystyczne dla TIS częste występowanie znaków zapytania w macierzy prowadzić może do znajdowania drzew, których w pełni dychotomiczne fragmenty mogą być konstruowane na podstawie jednoczesnego występowania tych właśnie znaków zapytania – czyli odzwierciedlać „pokrewieństwa” określane nie sygnałem filogenetycznym, ani nawet nie danymi, a jedynie sposobem kodowania. Dlatego Nelson (1992) wprowadził pojęcie **minimalnego kladogramu** (*minimal cladogram*), który odpowiada w TIS **kladogramowi w pełni popartemu danymi** (*strictly supported cladogram*) w technice redukcjonistycznej analizującej tradycyjnie kodowane dane. Minimalny kladogram to ten z kladogramów o najmniejszej długości, którego wszystkie węzły mają uzasadnienie w danych. W naszym przykładzie z początku rozdziału klad (CD) tworzy trychotomię z A i B i to jest właśnie minimalny kladogram: oczywiście możemy skonstruować zarówno kladogram (A,(B,(CD))), jak i (B,(A,(CD))) i oba będą

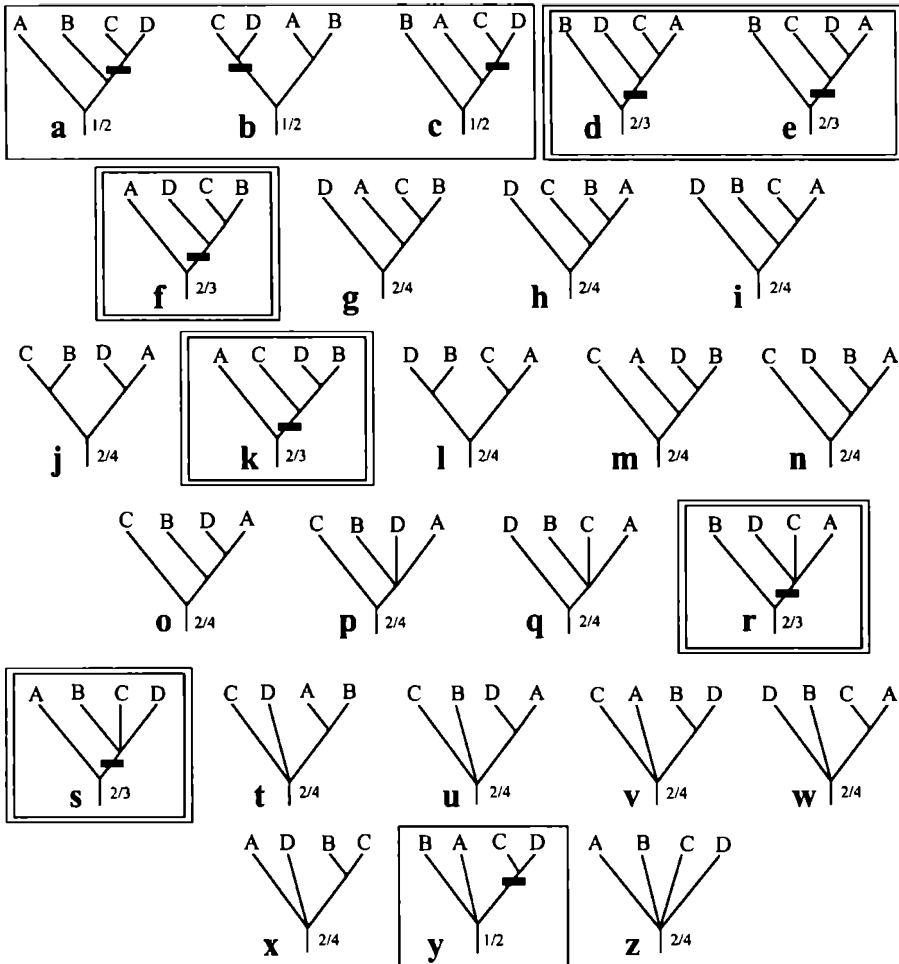
miały tę samą długość co kladogram minimalny, jednak w obu przypadkach rozkład trychotomii na dwie dychotomie nie będzie miał uzasadnienia w danych. Także minimalnych kladogramów może być więcej niż jeden. Współczynnik spójności rekonstrukcji CI będzie w technice TIS przybierał wartości 1 (stwierdzenie włączone w kladogram) lub 0,5 (niewłączone), więc nie będzie przydatny dla oceny dobroci rekonstrukcji; użyteczny natomiast będzie współczynnik retencji RI (Platnick 1993). W technice TIS $RI = ATS/(ATS + NTS)$ i odzwierciedla dobroć odwzorowania danych na drzewie.

TIS w porównaniu z tradycyjnym kodowaniem daje raz mniej, raz więcej kladogramów o najniższej wartości kryterium optymalizacji, kladogramy są raz identyczne jak przy kodowaniu tradycyjnym, innym razem nie. Wspominana już i wskazywana przez Nelsona i Platnicka (1991) większa czułość TIS występuje w niektórych przypadkach, lepiej wykorzystując względną informatywność cech. Metoda okaże się użyteczna wówczas, gdy w macierzy danych występuje wyraźny konflikt. Nelson (1996) przedstawia przykład czterech taksonów, dla których stwierdzono następujące stany trzech cech binarnych; oczywiście 0 to stan plezjomorficzny, 1 – apomorficzny:

	1	2	3
A	0	0	0
B	0	1	1
C	1	0	1
D	1	1	0

Analiza wykorzystująca tradycyjne kodowanie da sześć najkrótszych kladogramów, a obliczone na ich podstawie drzewo pełnej zgodności będzie jedną politomią, wskazującą na brak jakiegokolwiek informacji filogenetycznej w analizowanych danych. Dla odmiany kodowanie techniką TIS da trzy kladogramy, dla których drzewo pełnej zgodności będzie miało topologię A, (BCD), a więc taksony BCD utworzą grupę, odrębną od taksonu A. TIS wskaże więc na pokrewieństwo między B, C i D, choć konflikt w obrębie danych uniemożliwił uchwycenie tej zależności techniką tradycyjną. Inna rzecz, że można wskazać na mniejszą rangę takiego związku, nie mówiąc już o wyodrębnianiu grupy nieokreślonej synapomorfia, co jest niezgodne z kladyzmem, przynajmniej otodoksyjnym. Z drugiej strony, jest to użyteczne podejście dla grób politetycznych, a tak definiowane grupy przynajmniej nie są określane symplezjomorfiami.

Inny przykład, znów czterotaksonowy, gdzie taksony A i B dzielą stan plezjomorficzny, a C i D apomorficzny binarnej cechy, przytaczają Platnick i inni (1996). Istnieje 26 możliwych topologii takiego drzewa, wliczając wszystkie możliwe politomie (Ryc. 4.21). Przy każdej z topologii podano na rycinie wartość kryterium optymalizacji: dla kodowania tradycyjnego/dla kodowania TIS. Najkrótsze z nich – o wartościach 1/2 – to drzewa (a)–(c) oraz (y), otoczone pojedynczymi ramkami, identyczne dla obu technik. Kodowanie tradycyjne dla wszystkich pozostałych 22 drzew dało długość 2. Kodowanie TIS natomiast dało lepszą wartość kryterium – 3 – dla drzew (d)–(f), (k) oraz (r)–(s), otoczonych podwójnymi ramkami, niż dla 16 pozostałych (tam wartość 4). Podczas gdy suboptymalne kladogramy (f) i (g) mają przy kodowaniu tradycyjnym taką samą dobroć, to używając TIS kladogram (f) jest lepszy. Dla większej liczby cech, gdy takich przypadków jest więcej, wynikiem mogą być rekonstrukcje TIS lepiej odzwierciedlające filogenezę niż przy kodowaniu tradycyjnym.



Ryc. 4.21. 26 możliwych topologii czterotaksonowego ukorzonego drzewa, uwzględniając wszystkie możliwe politomie (a–z). Taksony A i B dzielą stan plezjomorficzny, a C i D apomorficzny cechy binamej. Dla każdej topologii podano wartość kryterium optymalizacji: kodowanie tradycyjne/kodowanie TIS. W pojedynczej ramce (a, b, c, y) kladogramy najlepsze wg obu metod (1/2), bez ramki kladogramy najlepsze (2/4) wg obu metod, w podwójnej ramce (d, e, f, k, r, s) kladogramy, dla których wartość kryterium optymalizacji w technice TIS (3) jest pośrednia między tą dla kladogramów w pojedynczej ramce a kladogramów bez ramki, gdy tradycyjne kodowanie tych kladogramów nie wyróżnia. Wg Platnicka i innych (1996)

Kodowanie TIS pozwala też uniknąć zdarzającego się nieraz nieuprawnionego wykorzystywania symplezjomorfii do grupowania taksonów. Choć wiemy, że taksony A i B „łączy” plezjomorfia, a jedynie C i D synapomorfia, to niejednokrotnie *ad hoc* traktuje się stany cechy u A i B jako np. następstwa odwróceń lub po prostu zapomina, że ten sam stan cechy nie odzwierciedla pokrewieństw, więc wynikiem rekonstrukcji bywa (A,B),(C,D), co w technice TIS nie jest możliwe. Podsumowując, wydaje się, że zdecydowana krytyka TIS nie jest uzasadniona i technika ma pewne zalety. Nie znaczy

to, że należy ją stosować zamiast kodowania tradycyjnego, a używać jej należy ostrożnie i koniecznie z wykorzystaniem proporcjonalnego ważenia, ostatecznie znajdując minimalny kladogram. Niewątpliwie jednak warto jej spróbować w sytuacjach, gdy liczne konflikty w macierzy danych powodują przy tradycyjnym kodowaniu znajdowanie wielu drzew najkrótszych.

Technika najsilniejszego świadectwa

Salisbury (1999) przedstawił technikę **najsilniejszego świadectwa** (*strongest evidence* – SE), skrytykowaną następnie przez Farrisa (2000); na tę krytykę autor odpowiedział (Salisbury 2000), przedstawiając nowe argumenty i symulację. Choć Salisbury (1999, 2000) uznaje SE za odrębną technikę, poniekąd zbliżoną do metody maksymalizacji wiarygodności, to wydaje się że SE można uznać za wariant techniki redukcjonistycznej, zwłaszcza że SE korzysta z liczenia kroków na drzewie, a nie zakłada jakichkolwiek hipotez *a priori* o przebiegu ewolucji. Technika jest kontrowersyjna i wymaga sprawdzenia, toteż przedstawimy ją tutaj jedynie w najkrótszym zarysie. Jak pisaliśmy w Rozdziale 4.5, nieważona technika redukcjonistyczna często zawodzi, a teoretycznie budzi wątpliwości, natomiast ważenie cech w praktyce zawsze nie jest wolne od arbitralności. Właśnie te wady ma eliminować bądź ograniczać zaproponowana technika SE. Także w SE cecha wymagająca mniejszej liczby kroków (a więc mniej podatna na zmiany) silniej popiera filogenezę, lecz związek – w odróżnieniu od ważonej techniki redukcjonistycznej – nie jest liniowy, zależąc od liczby stanów cechy i topologii rozpatrywanego drzewa. Konceptyjnie SE różni się od „normalnej” techniki redukcjonistycznej bardzo poważnie: zamiast zwyczajnego minimalizowania liczby hipotez *ad hoc* o homoplazjach, formułuje hipotezę zerową H_0 , z którą porównuje optymalizowane drzewo.

Hipoteza H_0 zakłada brak związku między topologią drzewa a występowaniem określonych stanów cech u określonych taksonów. Po ustaleniu liczby zmian na danej topologii dla danej cechy (czyli długości drzewa dla tej cechy), SE oblicza prawdopodobieństwo, że dla tej samej topologii liczba zmian tej cechy będzie taka sama lub niższa, gdy stany cechy zostaną losowo przypisane taksonom. W tym celu permutuje się stany cech między taksonami (a więc liczba określonych stanów danej cechy występujących na drzewie się nie zmienia, inne jest tylko ich rozmieszczenie u taksonów), oczywiście wielokrotnie (im więcej razy, tym lepiej), każdorazowo określa się długość drzewa i następnie zlicza drzewa tej samej długości lub krótsze; ich proporcja do łącznej liczby permutacji określa prawdopodobieństwo, że przy spełnieniu hipotezy H_0 drzewo będzie równie krótkie lub krótsze, a więc że dana cecha nie reprezentuje filogenetycznego sygnału. Im mniejsze to prawdopodobieństwo, tym bardziej dana cecha popiera daną topologię, tym więcej reprezentuje filogenetycznego sygnału. Przeprowadza się tę procedurę kolejno dla wszystkich filogenetycznie informatywnych cech. Wiemy, że przy niewielkich wartościach prawdopodobieństw wygodniej jest operować ich logarytmami. Dla każdej z cech definiuje się więc **uwidaczniający się sygnał filogenetyczny** (*apparent phylogenetic signal* – APS), jako: $-\log(P)$.

Zgodnie z ogólnym założeniem techniki redukcjonistycznej cechy są od siebie niezależne, a więc określone dla nich prawdopodobieństwa również: poziom homoplazji dla jednej cechy nie ma nic wspólnego z takim poziomem dla innej. Tak więc, jako dla zdarzeń niezależnych, prawdopodobieństwa mnoży się, aby uzyskać prawdopodobień-

stwo, że na danym drzewie wszystkie cechy łącznie nie reprezentują filogenetycznego sygnału. APS_d dla całego drzewa definiujemy jako sumę APS_i dla wszystkich n cech, drzewo o najwyższej wartości APS_d jest w SE drzewem najlepszym:

$$APS_d = \sum_{i=1}^n APS_i .$$

Jak widać, SE nie wymaga jakichkolwiek potencjalnie arbitralnych decyzji podejmowanych przy ważeniu cech, a z drugiej strony nie traktuje każdej z cech jako równie ważącej w rekonstrukcji. Nie zakłada też rzadkości zmian stanów cech, czyli powolnej ewolucji. Z drugiej strony, SE eliminuje wykorzystanie rzeczywistej wiedzy o ewolucji cech, pochodzącej z innych źródeł niż sama rekonstrukcja. Tak więc wydaje się, że SE najsensowniej wykorzystać można wówczas, gdy homologie i wagi cech pozostają całkowicie niejasne. SE w porównaniu z nieważoną techniką redukcjonistyczną ma tendencję do koncentrowania homoplazji w niektórych z cech, co może być zaletą – pozwalać lepiej identyfikować homoplazje – bowiem „normalna” technika redukcjonistyczna doda dodatkowy krok równie dobrze dla cechy reprezentującej niski sygnał filogenetyczny, jak i do cechy, która w rzeczywistości jest zupełnie pozbawiona homoplazji. SE znajduje też zwykle mniej najlepszych drzew niż „normalna” technika redukcjonistyczna. Wymaga jednak znacznie dłuższych obliczeń, Salisbury (1999) przedstawia iteratywne techniki skracające obliczenia SE. Symulacje wykazały, że w pewnych przypadkach SE jest gorsza od „normalnej” techniki redukcjonistycznej, w innych lepsza (Salisbury 2000), lecz zachowanie SE wymaga dokładniejszego sprawdzenia; w sumie wydaje się jednak techniką obiecującą.

4.9. Próbkowanie numeryczne i drzewa losowe

Dotąd przedstawiliśmy kolejne techniki analizy filogenetycznej, stosowanej dla rzeczywistych danych, choć większość z nich testowano metodą symulacji, jak wielokrotnie wskazywaliśmy. Można by więc zapytać: czemu ma służyć analiza danych otrzymywanych losowo bądź losowo generowanych drzew? Odpowiedź jest mniej więcej ta sama co w Rozdziale 2.8, gdy omawialiśmy liczby losowe. Otóż i tutaj po dane i drzewa losowe sięgamy wówczas, gdy proces jest zbyt złożony, aby opisać go technikami analitycznymi bądź choćby statystyką parametryczną. Rzeczywistość opisywana filogenezą jest daleko bardziej złożona niż dopuszczają to istniejące modele statystyczne, więc wykorzystanie randomizacji, a szerzej technik Monte Carlo wszelkiego rodzaju, jest tu szczególnie obiecujące. Nie ulega wątpliwości, że zastosowanie tych metod w rekonstrukcji filogenezy będzie coraz szersze, w tym rozdziale przedstawimy na skrótowym wskazaniu najczęstszych zastosowań. Nieco dokładniej zajmujemy się próbkowaniem numerycznym danych w Rozdziale 4.11, omawiającym wiarygodność zrekonstruowanych drzew. Najogólniej ujmując, losowo uzyskujemy dane bądź drzewa zgodne z hipotezą H_0 , by móc je porównać z danymi lub drzewami które badamy, by móc odrzucić lub nie tę hipotezę.

Śledząc procesy ewolucji cech i kladogenezy, szukamy odpowiedzi na dwa pytania: w jakim stopniu nasze rekonstrukcje odzwierciedlają rzeczywistość oraz jakie prawi-

długości i ewentualnie stojące za nimi mechanizmy dają się stwierdzić w zrekonstruowanej filogenezie. Dane i drzewa losowe są pomocą w odpowiedzi na oba pytania. Możemy więc, wykorzystując generator liczb pseudolosowych, dla danego zestawu taksonów i liczby cech, stworzyć zupełnie losowy zestaw „danych” i użyć ich do konstrukcji drzew. Dla tak uzyskanych „cech” obliczyć możemy takie współczynniki, jak CI, RI czy RCI, a także prześledzić rozkład długości drzew. Porównanie z rekonstrukcjami przeprowadzonymi dla naszych danych wskaże, na ile intensywny jest filogenetyczny sygnał w naszych danych. Możemy też generować losowo zachodzące zmiany stanów cech wzdłuż gałęzi drzewa i porównywać z ewolucją cech na drzewie opartym na naszych danych, co pozwoli na ocenę, na ile losowo odbywała się ewolucja na drzewie obliczonym dla naszych danych. Parametry procesu, jak prawdopodobieństwa zmian stanu cechy czy długość gałęzi, możemy przy tym odpowiednio modyfikować – pozwoli to na weryfikację niektórych hipotez o ewolucji cech.

Inne podejście to losowe przemieszanie danych w całej macierzy lub jej części, aby tak uzyskanych losowych danych użyć dla oceny intensywności sygnału filogenetycznego w naszych rzeczywistych danych. Ogólnie mówiąc, gdy dane losowe – losowo przemieszane dane oryginalne – dają drzewa zdecydowanie gorsze od obliczonych dla danych analizowanych, to wiarygodność rekonstrukcji rośnie. Rośnie też dla kladogramu lub jego części, gdy próbkowanie numeryczne generuje dane, na podstawie których oblicza się drzewa lub fragmenty drzew, najczęściej nieodbiegające od oryginalnego kladogramu. Wrócimy do tego w dalszych rozdziałach.

Losowo generować możemy nie tylko dane, aby użyć ich później do konstrukcji drzew, lecz także same drzewa (Maddison i Slatkin 1991). Jak wiemy, liczba teoretycznie możliwych topologii rośnie drastycznie wraz z rosnącą liczbą taksonów, tymczasem znalezione przez nas najlepsze drzewo bądź drzewa niejednokrotnie powinniśmy porównać z pozostałymi – dla oceny wartości takich współczynników, jak CI czy RI, a ponadto dla oceny rozkładu długości drzew (Fitch 1979, Huelsenbeck 1991) – rozkład zbliżony do normalnego wskazuje na dane niehierarchiczne, a więc niosące słaby sygnał filogenetyczny (Rozdział 4.11). Rozwiązaniem będzie więc losowe wygenerowanie, powiedzmy, paru tysięcy równieprawdopodobnych drzew i wykorzystanie ich długości zamiast niemożliwego do obliczenia pełnego zestawu rzeczywistych długości. Równieprawdopodobne drzewa to drzewa takie, że prawdopodobieństwo wylosowania każdego z nich ze zbioru wszystkich możliwych dychotomicznych i ukorzenionych drzew ma tę samą wartość.

Inna technika budowania losowych drzew to losowe łączenie kolejnych taksonów: z terminalnych OTU losowo wybiera dwa i łączy w kład będący nowym OTU, z pozostałej po tym grupy $T - 1$ OTU wybiera znów dwa i łączy w kład, znów wybiera losowo dwa z $T - 2$ OTU i tak dalej, aż do połączenia wszystkich taksonów w dychotomiczne drzewo. W ten sposób modeluje się proces podziału panmiktycznej populacji na subpopulacje potomne albo też procesy specjacji, gdy w każdej z linii ewolucyjnych specjacja jest równie prawdopodobna (Harding 1971, Simberloff i inni 1981, Maddison i Slatkin 1991). Drzewa losowe można też generować poprzez losowe dzielenie zestawu taksonów terminalnych. Najpierw dzielimy na pół, przy czym każdy z taksonów może się znaleźć w, powiedzmy, prawej połowie z prawdopodobieństwem 0,5 – w ten sposób powstają pierwsze dwa kłady. Następnie w każdej z grup dzieli się powtórnie, i tak dalej, aż każdy z taksonów utworzy odrębny kład. Powstają w ten sposób drzewa skrajnie symetryczne, skrajnie uporządkowane – takim ekstremalnym drzewom trudno przypisać sensowną

interpretację biologiczną, jest to raczej model krańcowo uporządkowanej topologii, służący porównaniom z empirycznymi drzewami. Zarazem jest to przeciwny biegun do skrajnie asymetrycznych, równieprawdopodobnych drzew – interesująca może być ocena, bliżej której skrajności leżą badane przez nas empiryczne drzewa. Losowo mogą być też traktowane politomie – rozkładane losowo na zestawy dychotomii, co umożliwi ocenę różnych rozwiązań, gdy politomia ma wiele gałęzi.

4.10. Wnioskowanie na podstawie więcej niż jednego drzewa i zestawu danych

Wynikiem wszystkich metod filogenetycznych przedstawionych w poprzednich rozdziałach (a także należącej do technik fenetycznych analizy skupisk) są drzewa. Dla wszystkich opartych na kryterium optymalizacji często – może wręcz zazwyczaj – wynikiem jest więcej niż jedno „najlepsze” drzewo. Nieraz też mamy drzewa obliczone różnymi technikami, gdy brak przesłanek do wyboru tej najlepszej albo też drzewa oparte na różnych zestawach danych; alternatywnie wyjściowe dane mają różny charakter – najbardziej typowa i częsta sytuacja to z jednej strony dane morfologiczne, z drugiej molekularne – i trudno zdecydować, czy można je wszystkie analizować łącznie. Teoretycznie, gdy drzewa obliczone dla różnych zestawów danych są takie same – lub „niemal” takie same – jest to silnym potwierdzeniem wiarygodności wyników (Penny i Henny 1986, Swofford 1991), jednak bywa tak nieczęsto. Oczywiście trudno po prostu przedstawić wszystkie „najlepsze” drzewa, zresztą stwierdzenie w rodzaju: „A to kład siostrzany B albo C lub F, zależnie od kladogramu” niewiele wyjaśnia. Konieczne jest więc stosowanie metod redukcji danych. Taka redukcja pociągać za sobą musi utratę części informacji, więc korzystać z niej trzeba rozważnie (deQueiroz 1993). Warto podkreślić, że brak jakiegokolwiek rutynowej techniki, której użyć można zawsze bądź niemal zawsze, a literatura obfituje w przeciwstawne opinie o tych samych metodach. Porównujemy zawsze ten sam zestaw taksonów, choć do analizy łącznej wprowadzić można ostatecznie – jak zobaczymy – taksony, dla których części danych brak: przykładem organizmy kopalne, dla których nie znamy części miękkich, analizowane wraz ze współczesnymi, czy też gatunki, dla których dysponujemy jedynie danymi o morfologii, analizowane wraz z innymi, dla których obok morfologicznych znamy też stany cech molekularnych.

Więcej niż jedno drzewo – porównanie filogenezy, drzewo zgodności

Teoretycznie najwygodniej jest przedstawić drzewo, podsumowujące całość informacji zawartej w porównywanych drzewach. W Rozdziale 4.8 wprowadziliśmy, chwilowo bez wyjaśnienia, pojęcie **drzewa pełnej zgodności** (*strict consensus tree*), jako drzewa przedstawiającego wynik techniki redukcyjnej dla danych kodowanych techniką TIS. Jest to jeden z wariantów **drzewa zgodności** (*consensus tree*). Zaproponowano wiele odmian takiego drzewa (Adams 1972, Nelson 1979, Barthélemy i Monjardet 1981, Margush i McMorris 1981, Sokal i Rohlf 1981, Schuh i Farris 1981, Schuh i Polhemus 1981, Stinebrickner 1984, Finden i Gorden 1985, Barthélemy i McMorris 1986, Bremer 1990, Page 1993), tu omówimy jedynie najczęściej stosowane. Warto podkreślić, że wielość metod obliczania drzewa zgodności wskazuje, iż po-

dejmowano i wciąż podejmuje się próby znalezienia możliwie najlepszej techniki, jednak nadal żadna nie jest w pełni zadowalająca. Jak wiemy, pełny opis drzewa obejmuje jego topologię i długości kolejnych gałęzi, choć nie dla każdego rodzaju drzewa te długości są estymowane. Drzewo zgodności zajmuje się jedynie topologią; liczba taksonów musi być dla wszystkich drzew taka sama. Oczywiście aby porównanie miało sens, taksony muszą być te same, choć niekiedy drzewo oblicza się dla zupełnie różnych taksonów – gdy np. badamy koewolucję różnych grup na tym samym terenie, koewolucję faun na terytoriach oddzielanych stopniowo powstającymi barierami, żywicieli i pasożytów czy drapieżników i ich ofiar albo kwiatów i zapylających je owadów (Mitter i Brooks 1983, Brooks 1988, 1990, Page 1988, Wiley 1988, Farrell i Mitter 1990, Hafner i Nadler 1990, Ronquist i Nylin 1990, Brooks i McLennan 1991, 1993, Futuyma i McCafferty 1991, Hafner i Page 1995, Hedges i inni 1996, Cooper i Penny 1997, Hoberg i inni 1997, Paterson i Gray 1997).

Porównując topologię dwóch drzew, często używa się **odległości topologicznej** (*topological distance*), zdefiniowanej przez Penny'ego i Hendy'ego (1985b). Dla dwóch nieukorzenionych dychotomicznych drzew odległość równa jest podwójnej liczbie wewnętrznych gałęzi, łączących taksony różne w porównywanych drzewach. Inaczej ujmując, jest to liczba wewnętrznych gałęzi, na których podział drzewa na dwie części da części różniące się pomiędzy porównywanymi drzewami. Jeżeli na porównywanych drzewach występują politomie, obliczanie odległości jest bardziej skomplikowane i wówczas możemy użyć ogólnego wzoru (Rzhetsky i Nei 1992a):

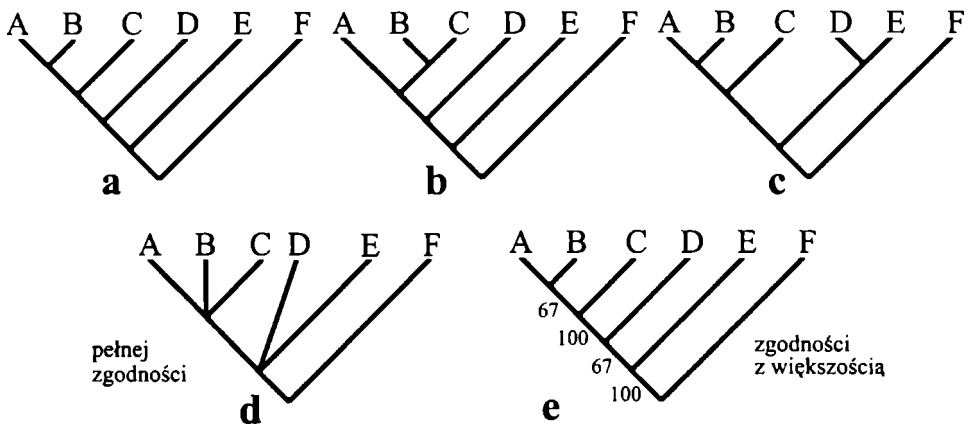
$$d_t = 2[\text{Min}(q_1, q_2) - p] + |q_1 - q_2|,$$

gdzie q_1 i q_2 to całkowite liczby możliwych podziałów, czyli wewnętrznych gałęzi dla drzew 1 i 2 ($q_1 \neq q_2$, gdy na jednym lub obu drzewach występują politomie), p to liczba podziałów (wewnętrznych gałęzi) takich samych dla obu drzew, a $\text{Min}(q_1, q_2)$ to mniejsza z wartości q_1 , i q_2 . Oczywiście dla drzew w pełni dychotomicznych ($q_1 = q_2$) powyższa formuła sprowadza się do $2q - p$. Ponieważ nieukorzenione dychotomiczne drzewo ma $T - 3$ gałęzi wewnętrznych, największa możliwa wartość d_t równa jest $2(T - 3)$. Długość odpowiednich gałęzi – a więc miarę anagenezy w ich obrębie – porównać można w ten sposób, że przyjmujemy jedną z topologii i dla drugich danych konstruujemy drzewo o takiej samej topologii (wygodnie to przeprowadzić programem MACCLADE: Maddison i Maddison 1992), po czym obliczamy współczynnik korelacji długości odpowiednich gałęzi (Omland 1994, Falniowski i inni 1996); poziom istotności bezpieczniej jest obliczyć techniką permutacji, jak to przedstawiliśmy w Rozdziale 2.8 dla testu Mantela. Oczywiście jako długości gałęzi przyjąć możemy – najbezpieczniej – średnią liczbę zrekonstruowanych zmian, choć niekiedy bezpieczniej będzie przyjąć liczbę minimalną, a można się też zastanowić nad maksymalną, jeżeli są jakieś przesłanki dotyczące odtwarzanego procesu.

Choć wielość drzew będących wynikiem analizy bywa przytłaczająca, to jednak one właśnie odzwierciedlają hierarchiczną strukturę danych, są to więc **drzewa podstawowe** (*fundamental trees*), a wszelkie obliczane na ich podstawie **pochodne** (*derivative*) drzewa zgodności są jedynie uproszczeniem, w dodatku często gorszym pod względem optymalizowanego kryterium od drzew podstawowych. Stąd szereg badaczy kwestionuje ich użyteczność, jako na ogół słabiej odzwierciedlających wyjściowe dane niż którekolwiek z drzew podstawowych (Miyamoto 1985, Carpenter 1988), jednak

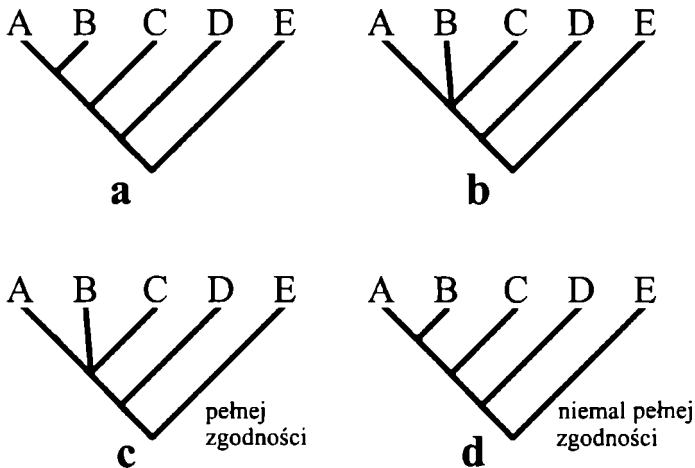
większość uważa, że drzewa zgodności bywają dobrym podsumowaniem danych. Niewątpliwie pozwalają na zorientowanie się, w jakim stopniu porównywane drzewa są podobne, które fragmenty rekonstrukcji są zawsze lub najczęściej takie same, a które klady mają najmniej stabilną pozycję. Bardziej dyskusyjne jest użycie drzew zgodności jako podstaw klasyfikacji, właśnie z powodu braku ich bezpośredniego oparcia w danych, często jednakże nie ma innego rozwiązania (np. Anderberg i Tehler 1990), jako że więcej niż jedno drzewo jako wynik analiz to wręcz reguła.

Dla ukorzonego dychotomicznego drzewa o T taksonach istnieje $T - 2$ informatywnych elementów, czyli wewnętrznych gałęzi. **Drzewo pełnej zgodności** (*strict consensus tree*), wprowadzone przez Schuha i Polhemusa (1981) oraz Sokala i Rohlf (1981), to drzewo, które przedstawia wyłącznie te gałęzie, które spotykamy na wszystkich podstawowych drzewach. W pozostałych przypadkach drzewo pełnej zgodności pokazuje politomie. Dla drzew (a), (b) i (c) z Ryc. 4.22 takie samo jest jedynie położenie taksonów A i F, a także odrębność B, C od D, E (Ryc. 4.22d). Jak widzimy, konstruując drzewo pełnej zgodności, tracimy wiele danych, lecz jest to jedyna technika w pełni przedstawiająca zgodność drzew, pokazująca wyłącznie te fragmenty, które występowały na wszystkich drzewach. Warto zauważyć, że jeżeli podstawowe drzewa różniły się gałęziami mającymi pełne poparcie w danych, to drzewo pełnej zgodności będzie miało gorszą wartość kryterium optymalizacji (np. będzie dłuższe w technice redukcjonistycznej) niż drzewa podstawowe. Jeżeli natomiast odmienne fragmenty topologii były następstwem optymalizacji przyjmującej różne, podobnie uprawnione rekonstrukcje ewolucji cech, to drzewo pełnej zgodności będzie miało tę samą długość co drzewa podstawowe, likwidując jedynie gałęzie o zerowej długości, i drzewo takie będzie właśnie wspomnianym już w Rozdziale 4.8 **kladogramem w pełni popartym danymi** (*strictly supported cladogram*), czyli o wszystkich węzłach popartych danymi, a więc w pełni uprawnioną – choć niecałkowicie dychotomiczną, „gubiącą” część informacji i rozdzielczości – podstawą klasyfikacji. Interpretacja drzew zgodności wszystkich pozostałych rodzajów wymaga ostrożności.



Ryc. 4.22. Dla trzech różnych dychotomicznych drzew podstawowych (a, b, c) obliczono drzewo pełnej zgodności (d) i zgodności z większością (e). Blizsze objaśnienia w tekście

Drzewo pełnej zgodności – jak jeszcze zobaczymy – często przedstawia wyłącznie lub niemal wyłącznie politomie, poza tym nie może wskazać części topologii, które były obecne choćby we wszystkich drzewach podstawowych z wyjątkiem jednego. Wygodnie jest więc obliczyć **drzewo zgodności z większością** (*majority-rule consensus tree*: Margush i McMorris 1981), w przypadku większej liczby podstawowych drzew często lepsze od drzewa pełnej zgodności (Swofford 1991). Przedstawia on te gałęzie, które występują u większości porównywanych drzew. Większość z definicji oznacza ponad 50% (konwencjonalnie już tę wartość technika uwzględnia), choć można przyjąć wartość wyższą, a niekiedy wygodnie jest ustalić próg poniżej 50% – oczywiście nie jest to wówczas „większość”, choć technika działa tak samo. Zwykle podaje się częstości przy gałęziach obliczonego drzewa zgodności. Na naszych drzewach (Ryc. 4.22a–c) kład ((A,B),C) występuje w 2/3, czyli 67% przypadków, podobnie topologia (((A,B,C),D),E),F); a więc drzewo zgodności z większością jest w pełni dychotomiczne, o gałęziach popartych odpowiednio 100 i 66% drzew podstawowych (Ryc. 4.22e). Obok wygodnego podsumowania informacji zawartej w wielu podstawowych drzewach, drzewo zgodności z większością jest szeroko stosowane przy ocenie wiarygodności poszczególnych kładów, podsumowując wyniki analizy techniką *bootstrap* (patrz Rozdział 4.11).

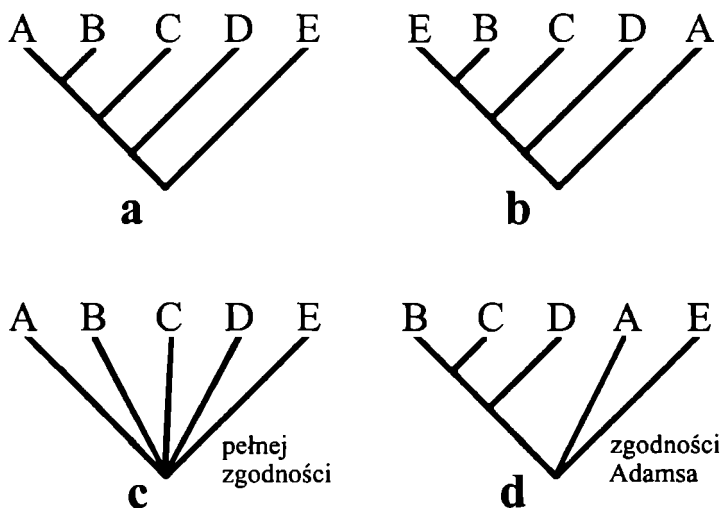


Ryc. 4.23. Gdy na jednym z podstawowych drzew (a, b) występują politomie (b), politomie te pojawią się także na drzewie pełnej zgodności (c), lecz nie na drzewie niemal pełnej zgodności (d). Przy braku politomii drzewa pełnej i niemal pełnej zgodności są identyczne

Gdy jedno z podstawowych drzew zawiera politomię, będzie ona obecna także na drzewie pełnej zgodności (Ryc. 4.23a–c). Zwykle taka politomia jest miękka, czyli wynika z niedostatku danych, a więc jakiegokolwiek jej rozłożenie na zestaw dychotomii nie może być z nią sprzeczne. Dla takich właśnie przypadków – podsumowania informacji niesprzecznej między drzewami – wprowadzono **drzewo niemal pełnej zgodności**, czyli **połączenia niesprzecznych składników** (*semi-strict consensus tree, combinable component consensus tree*: Bremer 1990). Zastępuje ono politomie występujące na którychś z drzew dychotomiami, jeżeli były one identyczne na wszystkich pozosta-

łych drzewach (Ryc. 4.23d). Oczywiście dla drzew w pełni dychotomicznych drzewo niemal pełnej zgodności będzie identyczne z drzewem pełnej zgodności.

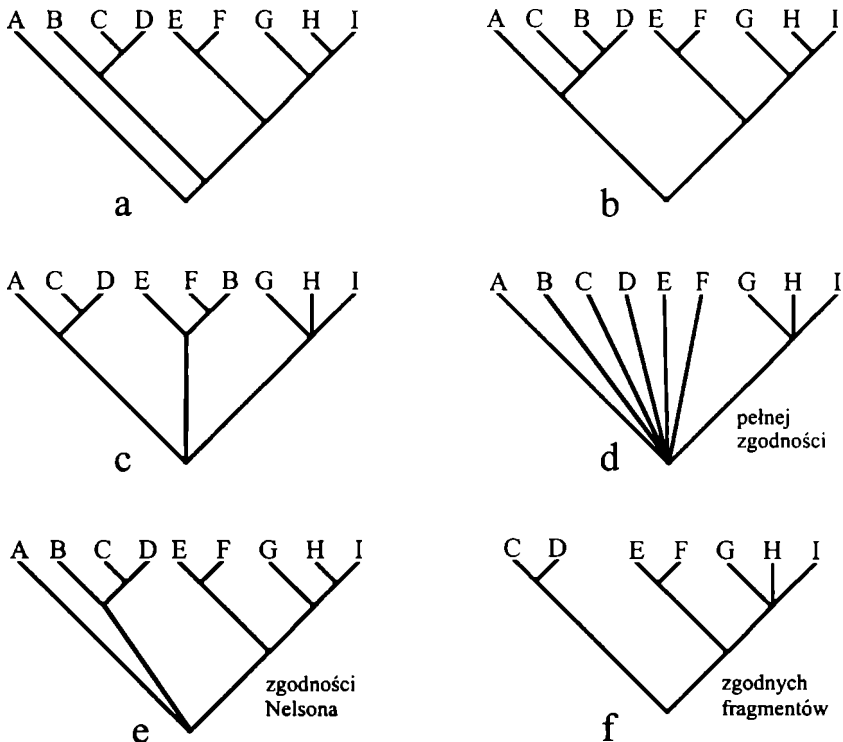
Bywa, że drzewo pełnej zgodności obliczone dla dwóch topologii (Ryc. 4.24a–c) jest jedną politomią, czyli że drzewa te są – formalnie – całkowicie różne. Bliższe przyjrzenie się drzewom z Ryc. 4.24a–b wskazuje jednak, że różnią się one jedynie przeciwnym położeniem taksonów A i E, podczas gdy wzajemne położenie pozostałych taksonów (B, C, D) jest identyczne. Dla takich przypadków wygodnie jest obliczyć **drzewo zgodności Adamsa** (*Adams consensus tree*: Adams 1972), wskazujące na stopień stałości struktury w obrębie badanych drzew. Drzewo zgodności Adamsa (Ryc. 4.24d) obliczamy, przenosząc taksony występujące na niezgodnych pozycjach w podstawowych kladogramach do najbliższego węzła wspólnego dla tych kladogramów. Jest ono bardzo pomocne w przypadkach, gdy porównywane topologie wydają się bardzo różne, a zwłaszcza gdy zawierają jeden lub więcej taksonów, których położenie drastycznie się zmienia w zależności od drzewa. Pamiętajmy jednak, że jego interpretacja bywa problematyczna: w naszym przykładzie taksony B, C i D nie tworzą przecież kladu, nie jest to grupa monofiletyczna, a jedynie „zestaw taksonów mających z sobą coś wspólnego filogenetycznie”; taki „klad” jest elementem, który nie wystąpił na którymkolwiek z podstawowych drzew.



Ryc. 4.24. Choć podstawowe drzewa (a, b) różnią się jedynie wzajemnym przestawieniem taksonów A i E, to drzewo pełnej zgodności (c) jest jedną politomią, choć przecież układ taksonów B, C, D jest na drzewach podstawowych taki sam. Wówczas warto obliczyć drzewo zgodności Adamsa (d), choć jego interpretacja wymaga ostrożności

W przypadkach, gdy w drzewie pełnej zgodności dominują politomie, pomocne bywa też **drzewo zgodności Nelsona** (*Nelson consensus tree*). Bywa ono definiowane różnie; tutaj przedstawimy je w ujęciu Page’a (1989, 1993), za Kitchingiem i innymi (1998), posługując się ich przykładem (Ryc. 4.25a–e). W tym rozumieniu drzewo zgodności Nelsona nawiązuje do wspomnianej w Rozdziale 4.5 analizy kompatybilności kladów, poszukującej największego możliwego zestawu kladów kompatybilnych (*largest clique*); w bardziej złożonych przypadkach obliczenie drzewa zgodności Nel-

sona wymaga komputera, szczególnie pomocny będzie program COMPONENT (Page 1993). Technika polega na formalnej analizie kompatybilności elementów drzew podstawowych. Zestawiając fragmenty topologii, które nie zawierają powtarzalnych elementów niezgodnych z tymi fragmentami, może na obliczonym drzewie przedstawiać grupowania niezgodne z niektórymi z podstawowych drzew (podobnie, jak wiemy, jest na drzewie zgodności z większością). Takson B jest kładem siostrzanym dla (C,D) na drzewie z Ryc. 4.25a, dla kładu D na drzewie z Ryc. 4.25b i dla kładu F na Ryc. 4.25c. Kład (C,D) występuje na Ryc. 4.25a i c, a grupa (B,C,D) na Ryc. 4.25a–b, więc największy zestaw kładów kompatybilnych dla tych trzech kladogramów to B,(C,D). (E,F) występuje na Ryc. 4.25a–b, a politomia (G,H,I) z drzewa na Ryc. 4.25c jest niesprzeczna z kładem ((G,(H,I)) z Ryc. 4.25a–b. Tak więc drzewo zgodności Nelsona (Ryc. 4.25e) wyodrębnia te wszystkie grupy; warto zauważyć, że drzewo pełnej zgodności dla tych samych drzew (Ryc. 4.25a–d) to dwie politomie. Drzewo zgodności Nelsona jest podobne do drzewa zgodności z większością i niezbyt często bywa stosowane, zwłaszcza że jego użycie budzi podobne wątpliwości jak sama technika kompatybilności, zarzucona w rekonstrukcji filogenezy, a interpretacji takiego drzewa brak jasno określonych podstaw.



Ryc. 4.25. Dla podstawowych drzew (a, b, c) drzewo pełnej zgodności (d) stanowi dwie politomie. Drzewo zgodności Nelsona (e) wskazuje na wspólne elementy podstawowych topologii, choć jego interpretacja budzi wątpliwości. Gdy podstawowe drzewa różnią się znacznie, drzewo zgodnych fragmentów (f) wskazuje najbardziej stałe fragmenty topologii. Za Kitchingiem i innymi (1998)

Bywa, że w przypadkach dużych rozbieżności między drzewami – jak w powyższym przykładzie, gdy podstawowym drzewem (Ryc. 4.25a–c) odpowiada drzewo pełnej zgodności w formie jednej lub paru politomii (Ryc. 4.25d) – sensowne jest zrezygnowanie z uwzględniania tych taksonów, których pozycja najbardziej różni się na kolejnych drzewach. Obliczamy wówczas **drzewo cząstkowe** (*subtree*), uwzględniające jedynie część taksonów (Ryc. 4.25f). Technika znana jest jako **drzewo zgodnych fragmentów** (*agreement subtrees*, *common pruned trees*: Finden i Gorden 1985). Drzewo zgodnych fragmentów zawiera wyłącznie te taksony i klady, które występują na dwóch lub więcej z podstawowych drzew. Jest to więc technika wykluczająca skrajności – eliminująca grupowania „unikatowe”, pojawiające się jednorazowo. Oblicza się takie drzewo poprzez wycinanie jednej lub więcej gałęzi z każdego z podstawowych drzew, aż uzyska się zestaw identycznych topologii. Najlepsze z obliczonych drzew (*greatest agreement subtree* – GAS) to takie, dla którego obliczenia trzeba było odrzucić najmniejszą liczbę gałęzi. Page (1993) zaproponował dla dwóch podstawowych drzew t_1 , t_2 odległość $d_{GAS}(t_1, t_2)$, równą liczbie gałęzi odrzuconych dla uzyskania GAS, jako miarę odrębności tych drzew. Drzewo zgodnych fragmentów bywa użyteczne w tych samych przypadkach co drzewo zgodności Adamsa, dając podobne wyniki; niestety dzieli też jego wady: jak widzimy, zarówno (C,D), jak i (E,F) nie na każdym z drzew podstawowych są kładami, a politomia (G,H,I) pozostała politomią i na GAS.

Więcej niż jeden zestaw danych – analiza łączna i oddzielna

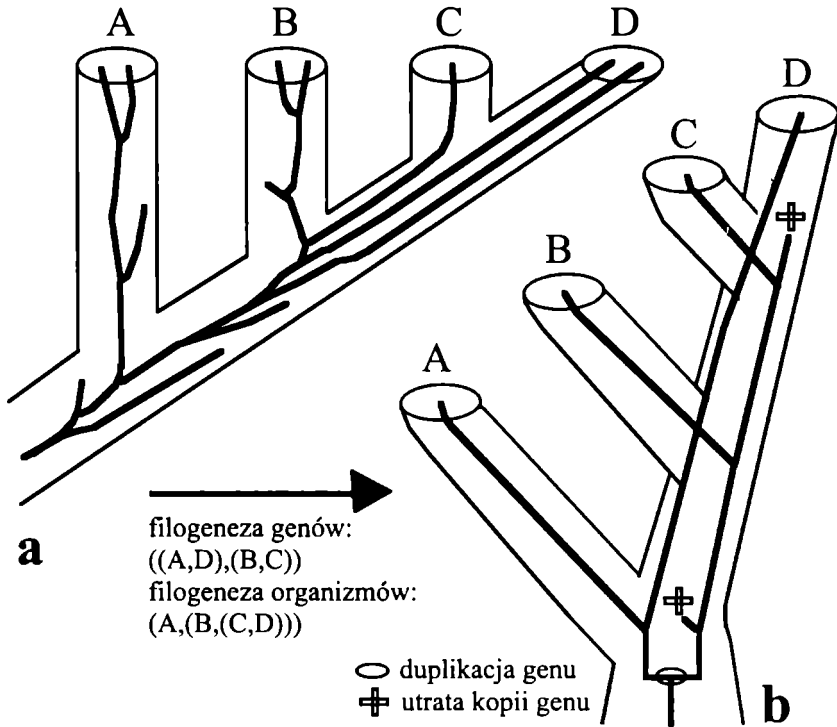
Jak już pisaliśmy, wynikiem analizy filogenetycznej jest często więcej niż jedno drzewo o takiej samej wartości kryterium optymalizacji, niejednokrotnie takich drzew jest wiele; wówczas jedynym sposobem redukcji danych jest obliczenie drzewa zgodności. W innych przypadkach, gdy mamy więcej niż jeden zestaw danych, postąpić możemy różnie: analizować każdy z zestawów oddzielnie, po czym obliczyć drzewo zgodności dla drzew podstawowych będących wynikiem tych analiz bądź też obliczyć drzewo dla całości danych, analizowanych łącznie. Często wyniki tych dwóch alternatywnych podejść będą różne, każde z nich ma zalety i wady, jak też zwolenników i przeciwników (Kluge 1989, Nixon i Carpenter 1996, Miyamoto i Fitch 1995). Niekiedy analiza oddzielna jest jedyną możliwością: nie da się poszukiwać drzewa na podstawie odległości genetycznych obliczonych dla częstości allozymów i jednocześnie odległości dla cech morfologicznych bądź odległości dla sekwencji DNA wraz z morfologicznymi albo odległości dla DNA z odległościami dla allozymów; tak samo niemożliwe jest obliczenie jednego rodzaju odległości dla obu lub trzech rodzajów danych łącznie, by na ich podstawie znaleźć drzewo. Podobnie nie możemy poszukiwać drzewa techniką maksymalizacji wiarygodności, przyjmując ten sam model ewolucji dla sekwencji i danych morfologicznych, zresztą – jak wiemy – techniki tej dla danych morfologicznych właściwie stosować się nie da. Większość rozważań o analizie łącznej przeciwstawianej oddzielnej dotyczy metody redukcjonistycznej (kladystycznej), gdzie formalnie nie ma ograniczeń dla analizy łącznej, zwłaszcza stosując różne wagi dla różnych cech i zróżnicowane koszty transformacji ich stanów.

Tam, gdzie to możliwe, wybieramy między trzema podejściami: **analizą łączną**, czyli **całościową** (*simultaneous analysis*, *total evidence analysis*), **analizą oddzielną** (*separate analysis*, *congruence approach*, *consensus approach*) i podejściem pośred-

nim: **warunkową kombinacją** (*conditional combination*). W tym ostatnim przypadku próbujemy analizować dane łącznie, jak długo nie stwierdzimy istotnej niezgodności sygnału filogenetycznego między różnymi rodzajami danych. Wyniki analizy łącznej – jak wspominaliśmy – mogą być inne niż drzewo zgodności obliczone dla podstawowych drzew, znalezionych w wyniku oddzielnych analiz kolejnych zestawów danych (Barrett i inni 1991), bowiem „najlepsze” drzewo dla wszystkich danych nie musi być identyczne z którymkolwiek z „najlepszych” drzew obliczonych dla cząstkowych danych.

Wybór między analizą łączną a oddzielną zależy po części od samych danych, ale też i od celu, jaki stawia sobie analiza. Najogólniej mówiąc, celem kladystyki jest maksymalizacja mocy wyjaśniającej – dąży się do wyjaśnienia obserwowanych rozkładów stanów wszystkich dostępnych cech, w sposób jak najpełniejszy (inna rzecz, że nie musi się to udawać – poza symulacjami określony przebieg filogenezy pozostaje niemal zawsze jedynie hipotezą). Zgodność ewolucji cech testuje filogenezę, a moc takiego testu rośnie z liczbą zgodnych cech, rośnie więc i wiarygodność rekonstrukcji filogenezy. Takie podejście przemawia za analizą łączną. Oczywiście jest tak wówczas, gdy technika rekonstrukcji jest spójna, czyli ze zwiększaniem liczby danych obraz ewolucji staje się coraz bliższy rzeczywistości – dodawanie nawet bardzo wielu danych wprowadzających w błąd jedynie lepiej zagłuszy filogenetyczny sygnał. Analiza oddzielna koncentruje się na fakcie, że różne, niezależne rodzaje danych w różnym stopniu odzwierciedlają sygnał filogenetyczny, więc należy je traktować odmiennie i osobno: skrajnym przykładem są dane morfologiczne przeciwstawiane molekularnym. Wyniki analiz jednych danych testują wyniki analiz innych danych, bowiem zgodność drzew obliczonych na podstawie danych niezależnych jest mocnym testem wiarygodności rekonstrukcji (Falniowski i Szarowska 1995).

Analiza oddzielna wymaga rzeczywistej niezależności kolejnych zestawów cech, a niezależność ta nie jest bynajmniej oczywista. Zewnętrzne skrzela u larwy salamandry to cecha czaszki, nie reszty szkieletu, ale też cecha części miękkich, nie szkieletowych, a zarazem cecha larwy, nie płaza dorosłego – jak więc dzielić? Miyamoto i Fitch (1995) sformułowali następujące kryteria odrębności danych: (1) geny nie są sprzężone, (2) produkty genów nie wchodzą w interakcje, (3) geny nie określają tej samej funkcji, (4) produkty genów nie uczestniczą w tym samym procesie fizjologicznym, (5) produkty genów z jednego zestawu nie regulują ekspresji genów z innego zestawu danych. Trudno nie zgodzić się z tymi kryteriami, a nasza wiedza o rzeczywistych mechanizmach dziedziczenia u niemal wszystkich organizmów jest na tyle mała, że właściwie nie wiemy nigdy, czy kryteria te są spełnione. Traktowanie kolejnych „genów” jako niezależnych jednostek również w znacznym stopniu odzwierciedla jedynie tradycję i takie czynniki, jak np. znajomość określonych primerów. Z drugiej strony, filogeneza różnych genów może być zdecydowanie różna i wówczas analiza łączna nie jest uprawniona. Ewolucja genu często nie odzwierciedla ewolucji organizmów (Ryc. 4.26a–b), w następstwie duplikacji, wygasania określonych linii, a także takich zjawisk, jak międzygatunkowa hybrydyzacja czy poziomy transfer. Dane molekularne można analizować oddzielnie od morfologicznych, mitochondrialne DNA oddzielnie od DNA jądrowego, a kodujące ocinki DNA oddzielnie od niekodujących. Ogólnie jednak dzielenie danych zawsze niesie z sobą ryzyko arbitralności, nieodzwierciedlenia podziałem odrębnych procesów ewolucyjnych.



Ryc. 4.26. Gen ulegać może duplikacji, a różne kopie genu są tracone, lecz te procesy często nie są skorelowane z procesami specjacji na poziomie organizmu (a), toteż rekonstrukcja filogenezy genu bywa różna od rekonstrukcji filogenezy organizmów mających ten gen (b). Jeżeli chcemy dopasować ewolucję genu do ewolucji na poziomie organizmów (b), musimy założyć występowanie duplikacji i utrat kopii

Bull i inni (1993) przeprowadzili symulację: dla znanej filogenezy utworzyli dwie grupy cech, ewoluujących szybko i wolno. Stwierdzili, że analiza łączna dała wyniki mniej zgodne z rzeczywistą filogenezą. Chippindale i Wiens (1994) wskazali, że – zgodnie z przewidywaniem – najgorsze wyniki dała analiza cech ewoluujących szybko, dla ewoluujących wolno była bliska rzeczywistej filogenezie, a analiza łączna dała dobre wyniki wówczas, gdy – inaczej niż Bull i inni (1993) wazący cechy tak samo – cechom ewoluującym szybko przypisano niższe wagi niż ewoluującym wolno, co jest zgodne z ogólnymi zasadami ważenia omówionymi w Rozdziale 4.5. Chippindale i Wiens (1994) wykazali, że odpowiednie ważenie cech i kosztów transformacji w analizie łącznej pozwala uniknąć większości problemów, wskazywanych przez przeciwników tego podejścia. Miyamoto i inni (1994) stwierdzili, że odpowiednie ważenie różnych pozycji przy analizie sekwencji paru genów mitochondrialnego DNA zwiększa spójność rekonstrukcji dla analizowanych oddzielnie genów. Sullivan (1996) przedstawił wyniki analiz dla dwóch genów: choć analiza oddzielna dała zdecydowanie różne drzewa, analiza łączna wskazała na drzewo bliskie rzeczywistej filogenezie.

Zwolennicy analizy oddzielnej wskazują, że pozwala ona na ocenę, w jakim stopniu drzewa obliczone dla kolejnych zestawów cech znajdują poparcie danymi, a więc jak

dobrze różne dane odzwierciedlają ewolucję. Jeżeli jeden zestaw danych przynosi – obok jednego drzewa najkrótszego – cały zestaw drzew krótszych zaledwie o jeden krok, a dla drugiego zestawu danych drzewo najlepsze jest o kilka kroków krótsze od jakiegokolwiek z suboptymalnych, to niewątpliwie drugi zestaw danych zawiera mniej homoplazji, jego sygnał filogenetyczny jest silniejszy. Z drugiej strony, w analizie łącznej można kolejno wyłączać cechy i całe grupy cech, uzyskując te same informacje. Innym argumentem za analizą oddzielną jest sytuacja, gdy np. mamy sekwencje o długości 2000 nukleotydów i, powiedzmy, 20 cech morfologicznych – tak wielka przewaga ilościowa cech molekularnych musi spowodować, że obliczone drzewo będzie odzwierciedlać wyłącznie cechy molekularne. W rzeczywistości jednak rzadko tak jest, bowiem tak wielka liczba cech molekularnych jest pozorna: większość pozycji będzie identyczna dla wszystkich sekwencji, w wielu wystąpią autapomorfie, więc ostateczna liczba cech filogenetycznie informacyjnych będzie podobna jak dla cech morfologicznych lub jedynie nieco większa.

Analiza łączna często daje dobre wyniki dla wielkich zestawów różnorodnych danych (Wheeler i inni 1993, Allard i inni 1999, Bininda-Emonds 1999). Jak wspomnieliśmy, jej wielką zaletą jest większa rozdzielczość, wyrażająca się znajdowaniem drzew o nielicznych politomiach, a więc większej zawartości informacyjnej. Politomie nieuchronnie występują w drzewach zgodności, podsumowujących analizę oddzielną. Inna rzecz, że takie drzewa zgodności mogą być bliższe rzeczywistości, przedstawiając konserwatywne estymaty filogenezy (Swofford 1991). Niewątpliwa utrata części informacji nie musi być jednak uzasadniona większą pewnością wyników – w rzeczywistości może być odwrotnie. Ponadto tracimy część informacji o ewolucji cech, bowiem politomie pociągają za sobą niepewność tych rekonstrukcji. Analiza oddzielna wiąże się też nieuchronnie z arbitralnym wyborem którejś techniki obliczeń drzewa zgodności (Kluge i Wolf 1993), choć w analizie łącznej też obliczamy drzewo zgodności, gdy wynikiem jest więcej niż jedno drzewo.

Teoretycznie celem systematyki jest rekonstrukcja kompletnej filogenezy całego świata żywego, a analiza wielkich zestawów danych, dla – powiedzmy – paruset taksonów to zadanie, z którym systematycy spotykają się często. Oczywiście dużym problemem jest wówczas niemożność pewnego znalezienia najlepszego drzewa, choć z tym w praktyce – jeżeli dane są hierarchiczne i odzwierciedlają rzeczywiście filogenetyczny sygnał – współcześnie stosowane techniki przybliżone radzą sobie nienajgorzej. Drugi problem to niekompletność danych: dla jednych taksonów mamy dane anatomiczne, gdy dla innych nie, dla taksonów kopalnych brak informacji o częściach miękkich, dane molekularne dostępne są wciąż jedynie dla małej części gatunków, a tam gdzie są – dotyczą różnych części sekwencji. We wszystkich tych sytuacjach warto jednak analizować wszystkie taksony. Można cechy morfologiczne jednego gatunku połączyć z molekularnymi innego, choć jest to ryzykowne. Czasem tworzy się „hybrydowy gatunek” – gdy np. mamy w obrębie rodzaju 15 gatunków i dla wszystkich z nich znamy stany cech morfologicznych, a jedynie dla jednego dysponujemy danymi molekularnymi, to możemy w obrębie rodzaju zrekonstruować filogenezę na podstawie morfologii, znaleźć plezjomorficzne stany wszystkich cech morfologicznych i utworzyć hipotetyczny takson, będący korzeniem drzewa dla rodzaju. Ten właśnie zestaw plezjomorficznych cech łączymy z posiadanym zestawem cech molekularnych i włączamy do większej grupy taksonów, dla której rekonstruujemy filogenezę. Finalnie „morfologiczne” drzewo dla rodzaju dołączamy do drzewa obliczonego dla całej

grupy. Technikę tworzenia takich „superdrzew” (*supertrees*) zaproponowali Sander-son i inni (1998), a Bininda i inni (1999) z powodzeniem użyli dla rekonstrukcji kompletnej filogenezy ssaków drapieżnych. Jedną z technik konstruowania superdrzewa łącząc drzewa cząstkowe jest metoda MRP, przedstawiona przez Bauma (1992) i Raganą (1992), zmodyfikowana przez Purvisa (1995); bardzo obiecująca, wymaga pełniejszego sprawdzenia na rzeczywistych danych.

Jak wiemy, zarówno polimorfizm jak i brakujące dane mogą spowodować błędną rekonstrukcję, zwłaszcza gdy rozsiane są w całej macierzy danych. To powoduje ostrożność we włączaniu taksonów, dla których brakuje znajomości stanów jednego rodzaju cech. Będzie tak, gdy do analizy prowadzonej na danych molekularnych i morfologicznych włączamy taksony, dla których informacji molekularnej brak, albo gdy włączamy taksony kopalne do analizy, prowadzonej dla cech szkieletów, części miękkich i cech molekularnych. Szczęśliwie przeprowadzone przez Wiensa i Reedera (1995) symulacje wskazują, że gdy brakujące informacje są skoncentrowane w niektórych częściach macierzy, to błędy rekonstrukcji filogenezy są niewielkie. Choć zalety analizy łącznej wydają się przeważać nad walorami analizy oddzielnej, bezkrytyczne użycie tego podejścia budzić musi niepokój. Warto więc przeprowadzić także analizę oddzielną, zanim przystąpimy do łącznej.

Rozwiązaniem powyższych kontrowersji może być też warunkowa kombinacja (Huelsenbeck i inni 1996). Oczywiście warunkiem analizy łącznej jest odpowiednio niewielka niezgodność między zestawami łączonych danych. Teoretycznie część niezgodności to wynik błędów losowych – niepełnej reprezentatywności analizowanych danych – i takie niezgodności nie wykluczają analizy łącznej. Gdy niezgodności są większe, danych nie powinno się łączyć. Rozróżnienie wymaga oczywiście odpowiedniego testu, a brak jakiegoś powszechniej przyjmowanego. Huelsenbeck i Bull (1996) zaproponowali **test proporcji wiarygodności**, analogiczny do stosowanego przy wyborze najlepszego drzewa w technice maksymalizacji wiarygodności, aby ocenić, czy drzewa skonstruowane dla różnych danych różnią się istotnie od siebie (H_0 : drzewa są identyczne): $2\Delta = \ln L_1 - \ln L_2$, gdzie L_1 i L_2 to wiarygodności drzew 1 i 2. Oczywiście statystycznie istotna wartość 2Δ wyklucza analizę łączną. Inny test przedstawili Farris i inni (1994). De Queiroz (1993) proponował skorzystanie z wartości **poparcia** (*support*: patrz Rozdział 4.11) obliczonych techniką *bootstrap* dla kładów, różniących się między drzewami znalezionymi na podstawie oddzielnie analizowanych danych: gdy poparcie jest duże, to analizy łącznej prowadzić nie należy. Jest to dyskusyjne, bowiem technika *bootstrap* budzi zastrzeżenia (patrz Rozdział 4.11), a ponadto bywa, że te właśnie rejony drzewa obliczonego analizą łączną mają nawet większe poparcie niż na drzewach obliczonych oddzielnie.

Dla n genów możemy poszukiwać drzewa w sposób analogiczny w metodzie redukcyjnej, znajdując drzewo t o najniższym koszcie, przy czym koszt $C(t)$ danego drzewa definiujemy jako:

$$C(t) = \sum_{m+1}^n c(G_m, t), \text{ gdzie } c(G_m, t) \text{ to koszt „dopasowania” genu } m \text{ do drzewa } t,$$

czyli liczba duplikacji i delecji genu, które trzeba postulować, aby ewolucję genu m przedstawić na drzewie t (Ryc. 4.26b). Jest to więc ściśle technika redukcyjna.

Guigó i inni (1996) skonstruowali w ten sposób drzewo, przedstawiające ewolucję głównych grup Eukaryota, używając 53 genów. Filogeneza jedynie 17 z nich była w pełni spójna z preferowaną na podstawie wszystkich danych filogenezą na poziomie organizmów. Powtórna analiza tych samych danych (Page i Charleston 1997) podwyższyła ocenę liczby genów ewoluujących zgodnie z ewolucją organizmów do mniej więcej połowy, jednak wynik jest dalej niepokojący i każe odnosić się z wielką ostrożnością do rekonstrukcji filogenez opartych na jednym lub paru zsekwencjonowanych genach. Z drugiej strony, dane molekularne zwykle nie obalają filogenez opartych na badanych od stuleci cechach morfologicznych, natomiast pozwalają na wyjaśnianie tych części filogenez, dla których cechy morfologiczne okazały się niewystarczające (Patterson i inni 1993).

Templeton (1983) zaproponował nieparametryczny test porównujący dwa drzewa. Do porównania wykorzystujemy najkrótsze drzewa dla obu zestawów danych, a gdy drzew najkrótszych jest więcej, to z obu zestawów wybieramy po jednym tak, aby drzewa były do siebie najbardziej podobne. Następnie badamy zgodność rozmieszczenia stanów poszczególnych cech z obu topologiami: gdy dopasowanie cechy do drzewa obliczonego dla innego zestawu danych wymaga znacznie więcej lub mniej kroków niż na drzewie, będącym wynikiem analizy tego zestawu cech, do którego należy ta właśnie badana, to różnicę trudno przypisać niereprezentatywności danych, będącej następstwem przypadku. Tak więc dla każdej cechy jednego zestawu obliczamy długość na topologii utworzonej na podstawie zestawu drugiego i porównujemy z długością na topologii utworzonej na podstawie zestawu danych obejmującego tę cechę. Różnicę zapisujemy odpowiednio ze znakiem + lub -. Po obliczeniu różnic dla wszystkich cech obu zestawów różnice sumujemy. Z sum: dodatniej i ujemnej wybieramy tę, której bezwzględna wartość jest wyższa. Istotność różnicy znajdujemy dla tej bezwzględnej wartości w tablicach testu sum rang Wilcoxa (Sokal i Rohlf 1995).

Gdy wartość jest statystycznie istotna ($< 0,05$), drzewa różnią się istotnie i analizę łączną uznajemy za nieuprawnioną. Gdy się istotnie nie różnią, prowadzimy analizę łączną. Statystycznie istotna niezgodność drzew bywa spowodowana skrajnie różnym położeniem któregoś z taksonów na porównywanych drzewach, więc można rozważyć pominięcie tego taksonu w dalszej analizie albo prowadzenie analizy łącznej pomimo tej niezgodności. Gdy wagi wszystkich cech są takie same i różnice między drzewami nie przekraczają jednego kroku, można użyć prostszej wersji testu Templetona, zaproponowanej przez Pragera i Wilsona (1988). Polega ona na porównaniu liczby cech lepiej odwzorowanych na każdym z drzew i wykorzystaniu testu dla rozkładu dwumianowego. Gdy dwa drzewa obliczono na podstawie sekwencji kwasów nukleinowych, Kishino i Hasegawa (1989) przedstawili parametryczny test, wykorzystujący różnicę D długości drzew, gdzie $D = \sum D_{(i)}$, a $D_{(i)}$ to różnica minimalnych liczb substytucji nukleotydów na dwóch drzewach na i -tej informatywnej pozycji. Hipotezę $H_0: D = 0$ sprawdzamy testem Studenta o $n - 1$ stopniach swobody. Wariancja, dla n informatywnych pozycji, dana jest wzorem:

$$s_D^2 = \frac{n}{n-1} \sum_{i=1}^n \left[D_{(i)} - \frac{1}{n} \sum_{k=1}^n D_{(k)} \right]^2, \text{ a test Studenta zależnością: } t = \frac{D/n}{s_D/\sqrt{n}}.$$

Dla niezgodności danych, wchodzących w skład dwóch zestawów, zaproponowano indeksy niezgodności (*incongruence*) I , przyjmujące wartości od zera (doskonała zgodność) do 1 (brak jakiegokolwiek zgodności). Dwa z nich to I_{MF} (Mickevich i Farris 1981) oraz I_M (Miyamoto, za Omland 1994). Podobnie jak w analizie wariancji, opierają się na partycjonowaniu całkowitej niezgodności danych z drzewami i_T na niezgodność wewnątrz zestawów danych i_W i między zestawami i_B : $I = (i_T - i_W)/i_T = i_B/i_T$. W obu indeksach niezgodność wewnątrz zestawu danych i_W oblicza się tak samo: jako sumę dodatkowych kroków wymaganych na najkrótszym drzewie dla danej grupy cech (sumowanych na obu drzewach łącznie), przekraczających najmniejszą możliwą liczbę kroków dla tej liczby cech, występującą przy braku jakichkolwiek homoplazji (Swofford 1991). Ogólną niezgodność i_T oblicza się różnie: dla I_{MF} dane się łączy i oblicza na ich podstawie drzewo, a i_T to liczba dodatkowych kroków, ponad minimum możliwe przy zupełnym braku homoplazji, policzone na najkrótszym drzewie znalezionym dla połączonych danych; dla I_M sumuje się dodatkowe kroki danego zestawu danych na najkrótszym z drzew obliczonych dla drugiego zestawu (i znowu sumuje dane dla obu zestawów). Zgodnie z notacją Swofforda (1991) $F(A * b)$ oznaczamy liczbę dodatkowych kroków wymaganych dla danych A na drzewie b , najlepszym dla zestawu danych B : jest to zatem liczba kroków ponad minimum określone dla zestawu B na drzewie B . Tak więc dla I_M : $i_T = F(A * b) + F(B * a)$.

4.11. Błędy losowe i systematyczne, wiarygodność znalezionych drzew

Jak wielokrotnie mówiliśmy, wszystkie znane techniki analizy filogenetycznej znajdują jedno lub więcej drzew, mających odzwierciedlać filogenezę badanej grupy, lecz nigdy nie możemy być pewni, czy przedstawiany przez nie obraz historii ewolucyjnej jest prawdziwy. Wątpliwości dotyczą zarówno zrekonstruowanej topologii, jak i długości gałęzi. Jest więc istotne, by móc określać wiarygodność całych rekonstrukcji lub ich fragmentów. Problem nie jest trywialny, bowiem – jak wiemy – większość technik analizy filogenetycznej wykracza poza zakres parametrycznej statystyki, a istniejące modele statystyczne są za proste, aby móc je wykorzystywać w analizie filogenetycznej; pomimo to dysponujemy szeregiem kryteriów pozwalających lepiej lub gorzej oceniać wiarygodność zrekonstruowanych filogenez. Warto jednak pamiętać, że brak jakiegokolwiek metody pozwalającej na rutynowe, bezkrytyczne i pewne określanie, na ile możemy wierzyć naszym rekonstrukcjom.

Jak dla każdego danych empirycznych, wyróżniamy dwa rodzaje błędów: losowe i systematyczne. Błędy losowe są następstwem niereprezentatywności badanej próby, co zdarza się często, gdy próba jest za mała. Jeżeli przykładowo w populacji ślimaków o fioletowej barwie ciała trafiają się albinosy, z częstością – powiedzmy – 0,001, to zdarzyć się przecież może, że wśród trzech badanych osobników trafi się albinos albo wręcz wszystkie trzy będą albinotyczne. Podobnie częstość dwóch równoległych substytucji nukleotydów na 300 pozycji łańcucha DNA nie wyklucza, że w danych 200 pozycjach stwierdzimy pięć takich substytucji, choć oczywiście w obu przykładach prawdopodobieństwa takich zdarzeń będą niewielkie. Ze wzrostem wielkości próby prawdopodobieństwo błędu losowego maleje, by dla próby nieskończonej (lub całej

skończonej populacji) być równe zeru. Błąd systematyczny natomiast pojawia się wówczas, gdy niespełnione są założenia techniki.

Błąd systematyczny – ocena, sposoby redukcji

Zwiększanie liczebności badanych obiektów – a także liczby analizowanych cech – błędu systematycznego nie zlikwiduje, a może go nawet zwiększyć. Przypomnijmy choćby omawianą w Rozdziale 4.5 „strefę Felsensteina”, gdy kolejne dane jedynie utwierdzają naszą wiarę w prawdziwość błędnej rekonstrukcji. Oczywiście błąd, zależnie od znaku, może nas utwierdzać w błędnej, lecz może i zwiększać poparcie prawdziwej rekonstrukcji. Ponadto błąd systematyczny stanowi poważny problem dopiero wówczas, gdy jest duży, a filogenetyczny sygnał słaby. Tym niemniej błąd systematyczny w najlepszym przypadku zmniejsza czułość stosowanej techniki. Jak pamiętamy, wszystkie metody rekonstrukcji filogenezy zakładają przekazywanie cech w „pionowym” następstwie dziedziczenia przodek – potomek, przy braku (lub niezmiernie rzadkości) przekazywania „poziomego”, w następstwie takich zjawisk, jak międzygatunkowa hybrydyzacja (bynajmniej nie tak rzadka jak niegdyś sądzono: Templeton 1989), introgresja czy poziomy transport fragmentów genomu przez wirusy. Inne założenie, jak wiemy niejednokrotnie niespełniane a zawsze niemożliwe do jednoznaczne go potwierdzenia, to wzajemna niezależność cech.

Źródła błędów dla kolejnych technik omawialiśmy w rozdziałach im poświęconych. Dla analizy skupisk wszystkie odległości muszą być znane, a dane ultrametryczne (obecny zegar molekularny), co rzadko w pełni ma miejsce. Dla technik drzew addytywnych również nie może być brakujących odległości, a same odległości muszą być addytywne; ponadto zwykle nie potrafimy prawidłowo korygować nieobserwowalnych zmian (wymaga to znajomości procesu ewolucji i istnienia opisującego go modelu matematycznego), a odległości nie są wzajemnie niezależne; kolejnym źródłem błędu systematycznego są same odległości, zwykle niedoskonale odzwierciedlające ewolucyjne różnicowanie w obrębie badanej grupy. Metoda redukcjonistyczna działa dobrze, gdy danych jest wiele (np. długie sekwencje), a różnice między taksonami nieliczne; ponadto drzewo musi być gęste, pozbawione długich gałęzi występujących obok krótkich. Wreszcie technika maksymalizacji wiarygodności wymaga spełniania warunków modelu (choć jest dość odporna na ich złamanie), w dodatku im model bardziej złożony, tym więcej potencjalnych źródeł błędu; metoda nienajlepiej też radzi sobie z długimi gałęziami. We wszystkich wymienionych przypadkach czułość technik spada, bowiem spodziewać się możemy systematycznych błędów.

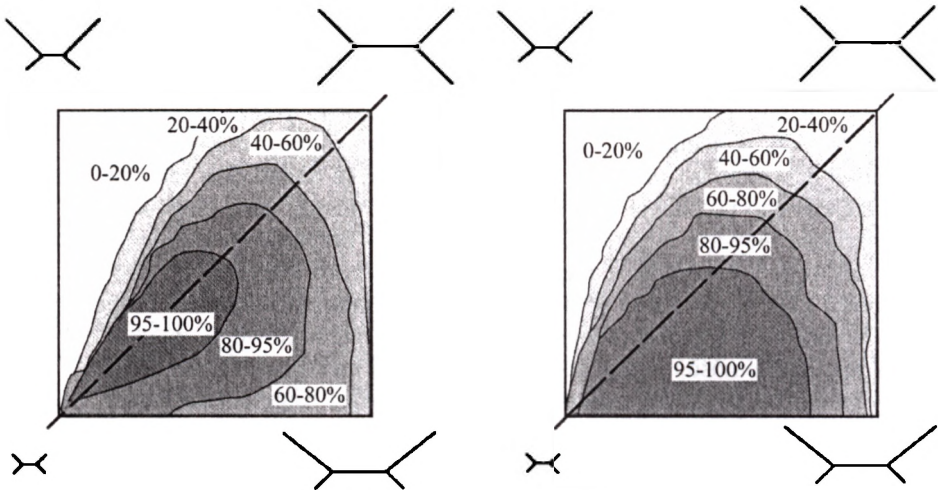
Możemy też podjąć pewne kroki w celu redukcji błędu systematycznego. Jednym z nich jest zmiana założeń; gdy przykładowo mamy przesłanki dla uznania, że między taksonami zmieniają się istotnie częstości poszczególnych nukleotydów, to celowe jest użycie dla odległości transformacji LogDet. Najczęściej jednak jakichś konkretniejszych przesłanek brak. Warto też pozbywać się długich gałęzi. Niezależność poszczególnych odległości zakładana przez techniki na nich oparte nie jest do końca spełniona, a obecność długich gałęzi w takim drzewie (dużych odległości) spowoduje tendencję do sumowania błędów, co ostatecznie może spowodować wystąpienie dużego błędu systematycznego. Dlatego też w tych metodach nie powinno się włączać do analizy więcej jak jednego – dwóch taksonów grupy zewnętrznej, choć możemy oczywiście wykorzystać więcej takich taksonów, lecz kolejno. Dla odmiany użycie wielu takso-

nów nienależących do grupy wewnętrznej jest pożądane w technice redukcjonistycznej, bowiem pozwala dzielić długie gałęzie; oczywiście nie jest tak, gdy użyjemy np. dwóch blisko z sobą spokrewnionych, a odległych od grupy wewnętrznej taksonów. Taksony zewnętrzne powinny w miarę równo porozcinać długie gałęzie. Oczywiście długie gałęzie wewnątrz grupy wewnętrznej tak samo mogą być źródłem błędu.

Paradoksalnie, choć czas obliczeń rośnie gwałtownie z liczbą taksonów i już dla mniej więcej 20 stosować musimy techniki przybliżone, to chcąc w sposób jak najpewniejszy ustalić pokrewieństwa między, powiedzmy, pięciu taksonami, najlepiej dołączyć do nich, powiedzmy, 15 jak najbliżej z nimi spokrewnionych, znaleźć (choćby techniką przybliżoną) MPR, po czym usunąć zeń te dodatkowe taksony, nie zmieniając wzajemnego położenia tych pięciu. Warto też usunąć dane niepewne lub wyraźnie zawierające niewiele sygnału filogenetycznego – przykładem szczególnie zmienne odcinki sekwencji. Oczywiście nieuniknione jest ryzyko arbitralnych decyzji, jednak arbitralności i tak nie jesteśmy w stanie całkowicie uniknąć w analizie filogenetycznej. Choćby wybór sekwencjonowanego odcinka kwasu nukleinowego, zwykle dobranego tak, aby jego zmienność między taksonami najlepiej umożliwiała rekonstrukcję pokrewieństw na danym poziomie uniwersalności. Jeżeli więc fragment sekwencji, wybranej np. dla rekonstrukcji pokrewieństw między rodzinami, wykazuje zmienność między najbliższymi sobie gatunkami, to zalety jego usunięcia z analizy niewątpliwie przeważają nad ryzykiem następstw arbitralności. Także omówione w Rozdziale 4.5 ważenie cech powinno zmniejszać błąd systematyczny.

Analityczne podejście do techniki jest najprostszą i precyzyjną metodą oceny wiarygodności wyników, jeżeli tylko (jak w przypadku analizy skupisk) model jest na tyle prosty, że możemy zwyczajnie sprawdzić spełnianie jego założeń. Dla analizy skupisk prowadzonej techniką UPGMA Nei i inni (1985) przedstawili analityczną metodę estymacji błędu znalezionych długości wewnętrznych gałęzi, a Rzhetsky i Nei (1992a, 1993) stworzyli wydajny algorytm obliczania błędu standardowego długości drzewa znalezionej metodą minimalnej ewolucji, wykorzystując technikę najmniejszych kwadratów. Niestety, możliwe jest znalezienie statystycznie istotnych wartości P_C dla długości gałęzi, pomimo że badane drzewo jest błędne (Sitnikova 1995, Sitnikova i inni 1996), ponadto test jest odpowiedni jedynie dla drzewa przedstawiającego pokrewieństwa między blisko spokrewnionymi taksonami (Nei i Kumar 2000). Teoretycznie wiarygodność drzew obliczonych różnymi technikami sprawdzać możemy, **porównując wyniki analiz ze znaną filogenezą**: problemem jest to, że w praktyce takimi „znanymi filogenezami” dysponujemy jedynie wyjątkowo – dla organizmów hodowanych w laboratorium. Hillis i inni (1992) hodowali bakteriofaga T7 w obecności mutagenu, po czym zrekonstruowali tę znaną z laboratorium filogenezę różnymi metodami (UPGMA, Fitch-Margoliash, Cavalli-Sforza, technika redukcjonistyczna). Wszystkie techniki prawidłowo zrekonstruowały rzeczywiste drzewo, a metoda redukcjonistyczna wiernie odtworzyła 97,3% ancestralnych stanów cech, jednak wyniki te są niezbyt miarodajne, bowiem drzewo było łatwe do rekonstrukcji. Leitner i inni (1996) podobnie wykorzystali znaną filogenezę wirusa HIV. Bardziej realistyczne oceny wiarygodności różnych technik rekonstrukcji filogenezy można uzyskać, rekonstruując drzewa dla organizmów, których pokrewieństwa są ustalone w sposób niewątpliwý, np. na podstawie rozległych danych morfologicznych, embriologicznych, paleontologicznych itp. (Kumazawa i Nishida 1995, Nei i Kumar 2000). Omówione w poprzednim

rozdziale przypadki **zgodności filogenezy** znalezionych różnymi technikami lub na podstawie różnych danych to kolejna metoda sprawdzania ich wiarygodności.



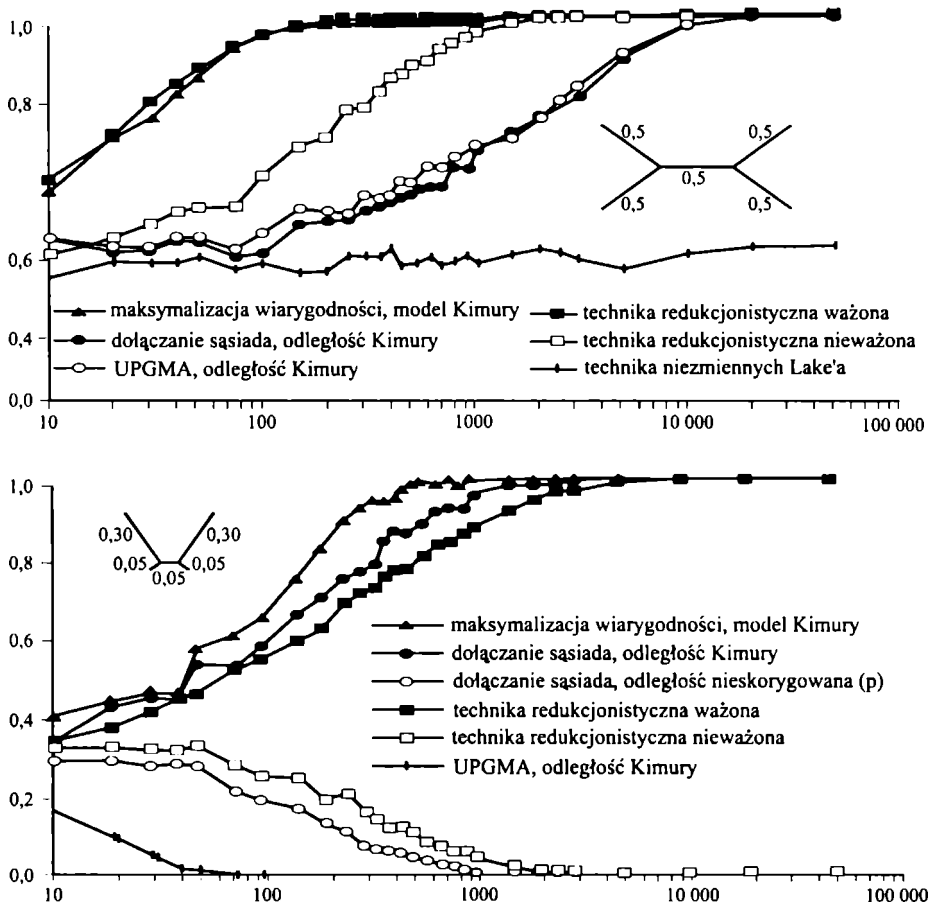
Ryc. 4.27. Zdolność do znajdowania prawdziwej filogenezy dla UPGMA (po lewej) i techniki kladystycznej (po prawej), przy różnych długościach kolejnych gałęzi, jak na rozmieszczonych wokół rogów kwadratów rysunkach (dokładniejsze wyjaśnienia w tekście). Przedstawione dane to procent przypadków odnalezienia prawdziwego drzewa dla określonej kombinacji długości gałęzi. W strefie Felsensteina (lewy górny róg) obie techniki nie są w stanie znajdować prawidłowych rekonstrukcji. Z Huelsenbecka i Hillisa (1993)

Inną metodą oceny spójności różnych technik są **komputerowe symulacje**, przeprowadzane najczęściej na czterotaksonowych nieukorzenionych drzewach. Najpierw tworzymy drzewo, o określonej topologii i długości gałęzi, po czym odpowiednio dobieramy do tego drzewa model ewolucji. Wykorzystując ten model, generujemy symulowane dane (najczęściej są to sekwencje kwasów nukleinowych), generowanie przeprowadzamy wielokrotnie, dla otrzymania wielu (zwykle kilkuset) zestawów. Następnie dla każdego zestawu znajdujemy drzewo najlepiej spełniające badane kryterium optymalizacji i w końcu liczymy, jaka część obliczonych drzew była zgodna z drzewem, dla którego dobraliśmy model. Takie symulacje mają bogatą literaturę (Nei 1991: omówienie wcześniejszych studiów, Hasegawa i inni 1991, Rzhetsky i Nei 1992a, Hasegawa i Fujiwara 1993, Kim i inni 1993, Kuhner i Felsenstein 1994, Gaut i Lewis 1995, Huelsenbeck 1995, Nei i inni 1995). Huelsenbeck i Hillis (1993) badali zachowanie różnych technik dla czterotaksonowych drzew o różnych długościach gałęzi. Rycina 4.27 przedstawia porównanie spójności rekonstrukcji dla analizy skupisk metodą UPGMA i nieważonej metody redukcyjnej. Kwadraty wyznaczają „przestrzeń topologiczną” drzew – skrajne przypadki to drzewo o równych krótkich gałęziach (lewy dolny róg), równych długich gałęziach (prawy górny róg: wzdłuż przekątnej lokuje się drzewa o równych długościach gałęzi), drzewo o dwóch gałęziach terminalnych długich, a pozostałych dwóch terminalnych oraz wewnętrznej krótkich (lewy górny róg: jest to, jak wiemy, „strefa Felsensteina”), a w prawym dolnym rogu drzewo o dwóch gałęziach terminalnych krótkich, dwóch długich i tak samo długiej

gałęzi wewnętrznej. Jak widzimy, obie techniki zupełnie nie radzą sobie w „strefie Felsensteina”, metoda redukcjonistyczna daje spójne wyniki w szerszym zakresie niż UPGMA, a zachowanie obu zależy od proporcji i długości gałęzi. Spójność rekonstrukcji jest największa dla gałęzi krótkich i o równych długościach. Stosunkowo najlepiej zachowywały się techniki maksymalizacji wiarygodności i ważona redukcjonistyczna.

Warto jednak zauważyć że, jak zwrócili na to uwagę Nei i Kumar (2000), spójność technik dla biologicznie realistycznych sytuacji jest daleko lepsza: Huelsenbeck i Hillis (1993) przeprowadzili symulacje dla skrajnie małych i skrajnie wielkich odległości p między sekwencjami, gdy w rzeczywistości tak wielkich odległości niemal nigdy nie spotykamy. Przy tak dalekich taksonach, dla których mogłyby wystąpić, napotykamy trudności współosiowania, a także wysoki poziom szumu informacyjnego. Dla danych rzeczywistych „przestrzeń topologiczna” jest daleko mniejsza, ograniczając się do lewej dolnej części kwadratu (mniej więcej ćwiartki albo i mniej: Nei i inni 1995, Rzhetsky i Sitnikova 1996), a tam różne techniki zachowują się w sumie podobnie (Nei i Kumar 2000). Dla testowanych czterotaksonowych drzew oznaczmy długości gałęzi – mierzone liczbą spodziewanych substytucji na pozycję sekwencji – jako a , b i c (Ryc. 4.27). Symulacje wykazały, że gdy $a = b = c$ i a zawiera się w przedziale $<0,1, 0,5>$, praktycznie wszystkie metody zapewniają poprawną rekonstrukcję, jeżeli tylko liczba nukleotydów $n > 100$. Gdy $0,1 < a < 0,5$, a b i c są zbliżone do 0,05, technika maksymalizacji wiarygodności jest lepsza od dołączania sąsiada, a najgorzej zachowuje się metoda redukcjonistyczna (Hasegawa i inni 1991, Hasegawa i Fujiwara 1993, Tateno i inni 1994, Huelsenbeck 1995). Natomiast gdy a , b i c zawierają się w przedziale $<0,01, 0,025>$ i n zbliża się do 1000, to wszystkie trzy metody dobrze znajdują prawidłową rekonstrukcję (Tateno i inni 1994). Jeżeli jednak para długich gałęzi terminalnych znajduje się na jednym końcu gałęzi wewnętrznej, a para krótkich na długim, relatywna spójność technik się zmienia (Nei 1996, Russo i inni 1996): metoda maksymalizacji wiarygodności jest najmniej spójna, znacznie lepiej zachowują się technika redukcjonistyczna i dołączania sąsiada, ta ostatnia najlepiej. Wynika to z tendencji do przyciągania krótkich/długich gałęzi, cechującej zarówno metodę redukcjonistyczną, jak i dołączania sąsiada.

Nawet przy takiej samej częstości zmian na wszystkich gałęziach nieważona technika redukcjonistyczna może być bardziej spójna niż maksymalizacji wiarygodności. Widać więc, że już dla czterotaksonowego drzewa nie jesteśmy w stanie ocenić, która z metod jest najbardziej spójna. Symulacje dla większej liczby taksonów były nieliczne, a jeżeli odległości między taksonami mają się mieścić w biologicznie realistycznych granicach, to długość wewnętrznych gałęzi maleje na tyle, że prawdopodobieństwo znalezienia właściwego drzewa gwałtownie się zmniejsza, chyba że użyjemy długich sekwencji, co zwiększa czas obliczeń ponad akceptowalną miarę. Ogólnie maksymalizacja wiarygodności jest lepsza niż dołączanie sąsiada, a to ostatnie niż nieważona technika redukcjonistyczna, lecz różnice nie są wielkie (Nei i Kumar 2000). Znów też musimy przypomnieć, że wszystkie modele substytucji są niezbyt realistyczne. Gdy Håstad i Björklund (1988) do generacji zestawów danych w symulacjach użyli parametrów obserwowanych dla mitochondrialnego genu cytochromu b , wszystkie trzy metody były równie dobre. Dopóki nie ma skrajnego zróżnicowania tempa ewolucji między gałęziami, techniki wydają się podobnie spójne.



Ryc. 4.28. Prawdopodobieństwa znalezienia prawdziwego drzewa (oś rzędnych) w zależności od długości użytych sekwencji (oś odciętych) dla różnych technik rekonstrukcji filogenezy, obliczone dla dwóch czterotaksonowych nieukorzenionych drzew, z których jedno ma wszystkie gałęzie tej samej długości (górny wykres), a drugie dwie dłuższe od pozostałych (dolny wykres). Dokładniejsze wyjaśnienia w tekście. Z Huelsenbecka i innych (1996)

Jak wiemy, metoda rekonstrukcji jest spójna, gdy w miarę wzrostu liczby danych rośnie prawdopodobieństwo znalezienia prawdziwej filogenezy, a gdy danych jest niewiele, najczęściej rozpatruje się przykłady krótkich sekwencji – praktycznie każda technika zawodzi tym bardziej, im wyższe jest T , czyli liczba taksonów terminalnych. Prawdopodobieństwo rośnie jednak różnie dla różnych metod i różnych długości gałęzi. Warto przytoczyć przykład symulacji (Ryc. 4.28) przeprowadzonych przez Huelsenbecka i innych (1996). Jak widać, dla wszystkich gałęzi tej samej długości – choć długich – zarówno ważona technika redukcyjistyczna, jak i maksymalizacji wiarygodności już dla sekwencji długości około 150 nukleotydów osiągały prawdopodobieństwo znalezienia prawidłowej rekonstrukcji bliskie 1, takie prawdopodobieństwo dla nieważonej techniki redukcyjistycznej wymagało już ponad tysiąca nukleotydów,

dla UPGMA i dołączania sąsiada – ponad 10 000, a technika niezmiennych Lake'a niezmiennie nie przekraczała prawdopodobieństwa 0,4. Dla drzewa o dwóch gałęziach terminalnych znacznie dłuższych od pozostałych gałęzi UPGMA dołączanie sąsiada i nieważona technika redukcjonistyczna ze wzrostem liczby nukleotydów wykazywały spadek prawdopodobieństwa znalezienia prawidłowej rekonstrukcji (czyli były niespójne), gdy techniki maksymalizacji wiarygodności (najlepsza w tym przypadku), dołączania sąsiada i ważona redukcjonistyczna osiągały wszystkie p bliskie 1 przy około 1000 nukleotydach. Dla bardzo krótkich sekwencji rekonstrukcja prawidłowa jest mało prawdopodobna, jakiegokolwiek techniki byśmy nie użyli.

Nei i inni (1998) stworzyli symulowane drzewo, dla którego przyjęli model ewolucji Jukesa i Cantora i wygenerowali 500 zestawów sekwencji, po czym rekonstruowali drzewa technikami minimalnej ewolucji, redukcjonistyczną i maksymalizacji wiarygodności, dla każdego z obliczonych drzew zapisując wartość kryterium optymalizacji. Symulacje przeprowadzili dla sekwencji o długości n równej 100, 300 i 600. Dla $n = 100$ wiele drzew miało niższą wartość kryterium niż drzewo modelowe, więc prawdopodobieństwo znalezienia prawdziwej rekonstrukcji było niskie, lecz już dla $n = 600$ przekraczało 0,8. Co istotne, poszukiwanie najlepszego drzewa prowadzono zarówno metodą pewną, jak i heurystyczną i wyniki dla tej ostatniej były praktycznie równie dobre jak dla pewnej (Nei i Kumar 2000).

W sumie wydaje się, że podczas gdy UPGMA często zawodzi, a nieważona technika redukcjonistyczna jest podatna na błędy systematyczne, to w większości przypadków wszystkie pozostałe metody są podobnie spójne, dając podobnie wiarygodne drzewa. Wątpliwości dotyczą krótkich, słabo popartych gałęzi, ale zwykle pozostaną one niepewne, bez względu na użytą technikę. Tym niemniej warto zawsze spróbować różnych metod, a obliczone drzewa poddać jak największej liczbie różnorodnych testów wiarygodności, mimo że wiele z nich budzi szereg zastrzeżeń i ma małą moc. Ponadto znów musimy podkreślić, że każda technika da drzewa błędne, gdy niespełnione są jej warunki, np. model korekcji nieobserwowalnych substytucji nukleotydów jest nieodpowiedni (Penny i inni 1993).

Ocena wiarygodności całego drzewa

Mając obliczone drzewo, możemy oceniać wiarygodność zarówno całego drzewa, jak i poszczególnych tworzących je kładów. Zajmijmy się najpierw metodami oceny wiarygodności całego drzewa. Większość rozważań dotyczyć będzie kladogramów obliczonych techniką redukcjonistyczną, bowiem najwięcej wiemy o tej właśnie metodzie. Dane słabiej lub silniej popierają MPR, czyli różnią się intensywnością sygnału filogenetycznego. Gdyby wszystkie potencjalnie informatywne filogenetyczne stany cech występowały w takiej samej liczbie, to wszystkie w pełni dychotomiczne topologie miałyby tę samą długość, więc dane byłyby filogenetycznie nieinformatywne: Goloboff (1991) określił je jako **niedecyzyjne**. Dane decyzyjne natomiast są podstawą rekonstrukcji drzew, różniących się długością: im bardziej, tym dane są lepsze, bardziej decyzyjne. **Decyzyjność danych** (*data decisiveness* – DD) definiuje się jako:

$$DD = \frac{\bar{S} - S}{S - M},$$

gdzie \bar{S} to średnia długość wszystkich możliwych dychotomicznych kladogramów, S to stwierdzona długość kladogramu MPR, a M to najmniejsza możliwa długość kladogramu, przy zupełnym braku homoplazji. Oczywiście im wyższa wartość DD, tym bardziej obliczone na podstawie danych kladogramy różnią się długością. Przybiera wartości w przedziale $\langle 0, 1 \rangle$, gdzie 1 to zupełny brak konfliktów w danych, a 0 to dane całkowicie niedecyzyjne; jego wartość nie zależy od obecności cech nieinformatywnych. DD nie musi korelować z liczbą MPR ani nie ma wyraźniejszego związku z tym, w jakim stopniu MPR są lepsze od pozostałych drzew. Niska wartość DD wskazuje, że podstawy uznania MPR za lepszy od pozostałych drzew są słabe, lecz wysoka nie upewnia nas w wyborze MPR, bowiem rozkład wartości DD jest nieznanym – teoretycznie można by oceniać istotność DD próbkowaniem numerycznym, np. permutując dane.

Hillis (1991) oraz Hillis i Huelsenbeck (1992) zaproponowali **badanie kształtu rozkładu długości drzew** obliczonych dla analizowanych danych (*distribution of cladogram lengths* – DCL), wszystkich lub losowej próby (Rozdział 4.9), gdy taksonów jest więcej. Już Fitch (1979, 1984) zauważył, że dane niehierarchiczne bądź słabo hierarchiczne dają drzewa, których długość ma rozkład symetryczny, mniej więcej normalny; dla danych hierarchicznych spodziewać się możemy rozkładu **lewoskośnego**, tzn. z wyraźnym ogonem po lewej stronie: jedynie wówczas nieliczne drzewa są niewiele krótsze od MPR. Hillis i Huelsenbeck (1992) wykazali, że miarą hierarchiczności danych jest stopień lewoskośności rozkładu, dany standardową statystyką g_1 , dla n drzew długości L i odchylenia standardowego s długości drzew (Sokal i Rohlf 1995):

$$g_1 = \frac{\sum_{i=1}^n (L_i - \bar{L})^3}{ns^3},$$

Huelsenbeck (1991) porównał rozkłady DCL dla drzew obliczonych na podstawie rzeczywistych danych i danych losowych, wykorzystując g_1 . Gdy dane odpowiadały jednemu MPR, DCL było silnie lewoskośne, gdy MPR było wiele, to rozkład był nieodróżnialny od otrzymanego dla danych losowych. Z drugiej strony, przypadkowe zgodności w losowych danych również dawały hierarchiczny obraz – lewoskośny DCL, więc g_1 zwykle było ujemne – należało więc poszukiwać niższej („bardziej negatywnej”) wartości dla danych hierarchicznych, w porównaniu z losowymi. Test DCL budzi wiele zastrzeżeń. Lewoskośność wystąpi wówczas, gdy określony stan cechy występuje u nielicznych taksonów, w niewielkiej części drzewa: można to np. osiągnąć dla danych losowych, duplikując któryś z taksonów. Hillis (1991) zaproponował znajdowanie takich taksonów/kładów, poprzez obliczanie wartości g_1 dla danych, w których kolejno eliminuje się po taksonie, a także porównywania rozkładu z rozkładami generowanymi losowo, dla ustalenia poziomu istotności wartości parametru. Losowe generowanie danych jest obliczeniowo znacznie szybsze od omówionych dalej permutacji danych rzeczywistych.

Brak lewoskośności stwierdzimy w przypadkach, gdy stan cechy dzieli taksony na dwie grupy podobnie liczne. Symulacje Huelsenbecka (1991) zakładały stałe prawdopodobieństwo zmian wzdłuż wszystkich gałęzi drzewa. Tempo ewolucyjnych zmian wpływa zarówno na wartość g_1 , jak i rzeczywiście na liczbę MPR i drzew niewiele od

MPR dłuższych. Najwyższe wartości g_1 i nieliczne MPR znajdujemy wówczas, gdy tempo jest umiarkowane. Gdy jest niskie, to większość cech jest niezmienna i występuje szereg autapomorfii ($g_1 = 0$); gdy jest wysokie, to rozkład stanów cech staje się praktycznie niezależny od topologii drzewa (skośność jest niewielka). Skośność nie zależy od liczby cech, co nie jest pożądane dla miary wiarygodności rekonstrukcji, ponadto skośność wystąpi zawsze, gdy mediana jest większa od średniej, choć „ogona” może nie być. Do tego obliczenia skośności są proste dla niewielu taksonów: dla ich większej liczby musimy użyć losowej próby drzew, a obecność w niej nielicznych w końcu drzew z „ogona” jest mało prawdopodobna. W sumie więc rozkład DCL jest słabym i mało wiarygodnym testem hierarchiczności danych. Ponadto rzeczywista hierarchiczność może też być następstwem nie filogenetycznego sygnału, a obciążenia danych różnymi frekwencjami nukleotydów w sekwencjach, albo wynikać z konwergencji.

Faith i Cranston (1991) zaproponowali **test permutacyjny** (*permutation tail probability* – PTP). Zwrócili oni uwagę na to, że przy licznych homoplazjach sygnał filogenetyczny bywa tak słaby, iż rekonstrukcja przeprowadzona na porównywalnym zestawie danych, w którym kladystyczne związki (kowariancje) między cechami zastąpiono związkami powstałymi w wyniku przypadku, dać może MPR o podobnej lub nawet niższej długości, a wówczas wiarygodność rekonstrukcji przeprowadzonej na oryginalnych danych jest oczywiście bardzo niska. PTP polega na permutacji macierzy danych w ten sposób, że częstości poszczególnych stanów cech pozostają niezmienione, ponadto nie permutuje się cech dla taksonów grupy zewnętrznej. Permutacje przeprowadza się niezależnie, kolejno dla każdej z cech. Dla takiej macierzy znajduje się MPR i odnotowuje jego długość. Permutacje powtarza się, powiedzmy, 99 razy (oczywiście im więcej permutacji, tym lepiej), otrzymując w ten sposób $99 + 1$ macierzy i z nich $99 + 1$ MPR. Ostatecznie liczymy te kladogramy, które były tej samej długości bądź krótsze niż MPR obliczony dla oryginalnych danych. Hipoteza H_0 zakłada brak kladystycznej struktury danych; jeżeli więc kladogramów krótszych od obliczonego dla oryginalnych danych było mniej niż pięć, to $PTP \leq 0,05$, a więc hipotezę H_0 możemy odrzucić z prawdopodobieństwem 0,05. Gdy brak kladogramów o długości równej lub mniejszej od MPR obliczonego dla danych oryginalnych, Faith i Cranston (1991) zaproponowali umowną wartość $PTP = 0,01$.

Oczywiście przy większej liczbie taksonów stukrotne znalezienie najkrótszego drzewa będzie bardzo czasochłonne, a dla jeszcze większej liczby – zwyczajnie niemożliwe. Proponowano obliczenie drzewa dla danych oryginalnych techniką pewną, a dla permutowanych – metodami przybliżonymi, jednak wówczas należy spodziewać się zaniżonych wartości PTP, czyli zawyżonego estymatu wiarygodności rekonstrukcji. Inna możliwość, to użycie tej samej przybliżonej techniki; wtedy zaniżanie wartości PTP może być jeszcze większe, bowiem dla danych losowych spodziewać się należy szczególnie licznych lokalnych optimów długości drzewa. Kitching i inni (1998) omawiają jeszcze inne próby estymacji wartości PTP dla większej liczby taksonów, również niezbyt pewne. Test PTP budzi jednak poważniejsze zastrzeżenia (Bryant 1992, Carpenter 1992, Kitching i inni 1998). Jeżeli nawet filogenetyczny sygnał w naszych danych jest słaby, to i tak macierz danych zestawia starannie wybrane cechy. Cechy te mają określone biologiczne znaczenie, gdyż starannie rozważono – zgodnie z całą wiedzą – homologie, synapomorfie, serie transformacyjne, itd., a kladogram odzwierciedla biologiczne związki między cechami. Porównywanie go więc z drzewami obliczonymi

dla losowo przemieszanych danych (a zdarzyć się może, że po takim przemieszaniu rekonstrukcje będą bardziej spójne, choć przecież pozbawione jakiegokolwiek biologicznego sensu) budzić musi poważne wątpliwości. Ponadto „statystyka PTP” założeń parametrycznej statystyki nie spełnia: technika permutacji została zaproponowana bez jakiegokolwiek uzasadnienia, czemu akurat permutacje i akurat w ten sposób przeprowadzane, ponadto badane dane nie są losową próbą, a wręcz odwrotnie – starannie wybranym zestawem. W sumie więc obliczanie poziomu istotności jest nieuprawnione. PTP może być co najwyżej wskazówką, na ile intensywny jest sygnał filogenetyczny. Jeśli długość MPR dla danych oryginalnych o wiele niższa od długości dla permutowanych wskazuje na hierarchiczną strukturę danych i zwiększa nasze zaufanie do obliczonego drzewa, to występowanie drzew o podobnej długości obliczonych dla danych permutowanych takiej struktury nie wyklucza. Fu i Murphy (1999) uznali jednak PTP, na podstawie zachowania dla danych symulacyjnych i rzeczywistych, za dobrą metodę wykazywania kowariancji cech i wartościowania poziomu sygnału filogenetycznego w danych, zwłaszcza dla sekwencji DNA zawierających liczne homoplazje. Inne techniki permutacyjnych testów obecności struktury filogenetycznej w danych przedstawia Alroy (1994), wymagają one jednak sprawdzenia.

Goldman (1993) zaproponował użycie omówionego już w Rozdziale 4.6 testu proporcji wiarygodności do oceny, czy bardziej złożony model ewolucji lepiej wyjaśnia dane – daje lepsze drzewo. Jak pamiętamy, bardziej złożony model może lepiej wyjaśniać procesy ewolucyjne, lecz wymaga dłuższych obliczeń i jest bardziej podatny na błędy losowe. Zaczynamy więc od najprostszego modelu, a gdy bardziej złożony statystycznie istotnie jest lepszy, odrzucamy ten prostszy. Nie znaczy to oczywiście, że ten bardziej złożony jest właściwy, a jedynie że prostszy jest za prosty. Teraz uznajemy ten bardziej złożony za prostszy i porównujemy z kolejnym, jeszcze bardziej złożonym. Oczywiście test wymaga znajomości rozkładu wartości Δ – jak wiemy (Rozdział 4.6), zastosowanie rozkładu χ^2 budzi zastrzeżenia. Goldman (1993) zaproponował więc użycie symulacji dla zbadania rozkładu statystyki Δ . Drzewo i parametry modelu są estymowane przy założeniu, że prostszy model jest właściwy, po czym te estymaty drzewa i procesu służą do generacji wielu zestawów danych, każdy tej samej wielkości co dane oryginalne. Na ich podstawie oblicza się drzewa, ich wiarygodności tworzą rozkład dla hipotezy H_0 . Jedną z wad techniki jest to, że bardziej zgodny z rzeczywistością model ewolucji niekoniecznie musi umożliwiać rekonstrukcję filogenezy bardziej zgodną z prawdziwą (patrz Rozdział 4.6). Nie można natomiast użyć testu proporcji wiarygodności dla porównania dwóch alternatywnych topologii obliczonych techniką ML, bowiem miałby on zero stopni swobody (Felsenstein 1988), a funkcja wiarygodności nie zachowuje ciągłości i różniczkowalności (Yang i inni 1995).

Ocena wiarygodności poszczególnych kładów

Przy braku homoplazji, im więcej synapomorfii określa dany kład, tym pewniejsze jest jego wyodrębnianie. Gdy występują homoplazje – jak niemal zawsze ma miejsce – to proste kryterium przestaje być użyteczne. Jedną z metod określania stopnia poparcia danego kładu analizowanymi danymi jest miara **poparcia Bremera** (*Bremer support*: Bremer 1988, 1994, Kitching i inni 1988: używa się też terminów *decay index* lub *support index*). Poparcie Bremera obliczamy jako liczbę dodatkowych kroków na drzewie konieczną, aby kład znikł z drzewa pełnej zgodności – a więc im więcej takich

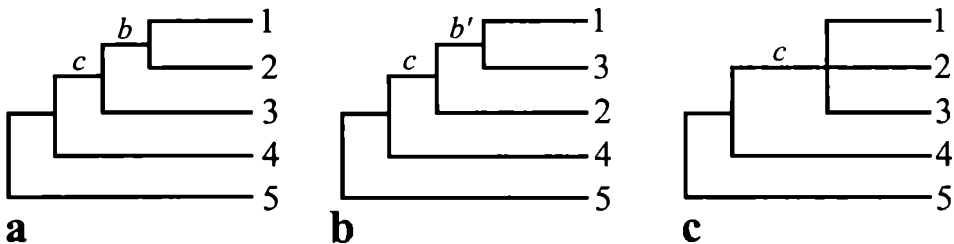
kroków trzeba, tym większa wartość poparcia Bremera, tym większa wiarygodność rozpatrywanego kladu. Gdy mamy jeden MPR, poparcie Bremera dla wszystkich kladów musi być ≥ 1 . Gdy MPR jest więcej, to nieuchronnie niektóre z kladów znikną już z drzewa pełnej zgodności obliczonego dla wszystkich MPR, a więc ich poparcie Bremera będzie równe zero. Dla obliczenia poparcia Bremera znajdujemy w pierw drzewo pełnej zgodności dla wszystkich MPR, następnie dla tego drzewa pełnej zgodności lub MPR (gdy jest jeden) i wszystkich drzew dłuższych o jeden krok od MPR znów obliczamy drzewo pełnej zgodności: te klady, które wówczas znikną (wejdą w skład politonii), będą miały poparcie równe 1. Następnie obliczamy drzewo pełnej zgodności dla tego drzewa pełnej zgodności i wszystkich drzew dłuższych od MPR o dwa kroki, i tak dalej, aż wszystkie klady znikną z drzewa pełnej zgodności – w ten sposób dla każdego z kladów określiliśmy miary poparcia Bremera. Oczywiście w przypadku całkowitego braku homoplazji poparcie Bremera będzie równe liczbie kroków na danej gałęzi, czyli długości tej gałęzi. Suma poparcia dla kolejnych gałęzi to **ogólne poparcie** (*total support* – TI) dla całego kladogramu. Wartość TI nie może przekroczyć całkowitej długości kladogramu, a więc TI możemy skalować: jako proporcję TI do długości kladogramu: $ti = TI/L$. Gdy drzewo pełnej zgodności dla wszystkich MPR jest jedną politonią, to $ti = 0$; gdy brak homoplazji, $ti = 1$. Wysoka wartość ti wskazuje na dużą wiarygodność drzewa i tworzących go kladów, lecz niska nie wyklucza istnienia kladów, dla których poparcie Bremera jest wysokie. W sumie poparcie Bremera to użyteczna miara wiarygodności kolejnych fragmentów drzewa, choć nie jest jasne, jak duża wartość poparcia musi charakteryzować dobrze poparte klady.

Davis (1993) zaproponował **indeks stabilności kladu** (*clade stability index* – CSI). Cechy stopniowo usuwa się z macierzy, najpierw pojedynczo, później coraz większymi grupami, aż kład znika z drzewa pełnej zgodności obliczonego dla wszystkich MPR z tak zredukowanej macierzy danych. Wówczas określamy CSI jako proporcję liczby cech, które musiały zostać usunięte z macierzy, aby kład znikł z drzewa pełnej zgodności, do całkowitej liczby cech filogenetycznie informatywnych w wyjściowej, pełnej macierzy. Przy większej liczbie cech obliczenia CSI mogą być zbyt czasochłonne. Davis (1993) zaproponował przybliżone obliczenia CSI w takich przypadkach. Wartość CSI silnie zależy od liczby cech w macierzy, ponadto nie daje się sensownie obliczać dla cech wielostanowych uporządkowanych. Poparcie Bremera wydaje się więc znacznie lepszym parametrem.

Faith (1991) zaproponował **test permutacyjny zależny od topologii** (*topology-dependent permutation tail probability test* – T-PTP) dla badania monofiletyczności danej grupy taksonów. Parametrem jest różnica długości najkrótszych drzew, z których na jednym badana grupa jest monofiletyczna, a na drugim nie. Rozkład parametru określamy techniką permutacji danych, z których następnie oblicza się oba najkrótsze drzewa. Wartość parametru zależy od struktury danych, niemającej związku z monofiletycznością testowanej grupy; T-PTP jest mało użytecznym sprawdzianem tej monofiletyczności.

Li i Gouy (1991) przedstawiają szereg metod testowania, czy długość określonej gałęzi wewnętrznej drzewa obliczonego na podstawie addytywnych odległości jest statystycznie istotnie różna od zera. Gdy nie jest, to topologia drzewa, na którym występuje ta gałąź, jest raczej błędna (Sitnikova i inni 1995). Drzewo (a) (Ryc. 4.29a) jest identyczne z drzewem (b) (Ryc. 4.29b) jedynie wówczas, gdy długości gałęzi są $b = b' = 0$, i wtedy rzeczywistą topologię przedstawia drzewo (c) (Ryc. 4.29c). Dla

drzew opartych na odległościach możemy więc przedstawić hipotezę zerową H_0 : długość gałęzi ≤ 0 , a w naszym przypadku odrzucenie H_0 : $b \leq 0$ jest testem wiarygodności topologii. Dla dostatecznie dużej liczby taksonów T rozkład długości gałęzi zbliża się do normalnego (Rzhetsky i Nei 1992a), można więc wykorzystać standardową statystykę Z do testowania istotności różnicy od zera (Sokal i Rohlf 1995). Możemy np. sprawdzić monofiletyczność grupy taksonów 1 i 2 (Ryc. 4.29): odrzucić musimy hipotezę H_0 : $b \leq 0$, natomiast długość c nie ma tu znaczenia. Dla sprawdzenia identyczności drzew (a) i (b) drzewo (c) jest „drzewem zerowym” – jego odrzucenie wyklucza tę identyczność. Takie testy są odpowiednie jedynie dla drzew opartych na odległościach, gdzie długości prawdziwych topologii muszą być dodatnie, a obliczonych niekoniecznie, natomiast nie nadają się dla technik redukcjonistycznej i maksymalizacji wiarygodności, gdzie długości gałęzi zawsze są dodatnie, toteż nie jesteśmy w stanie sformułować hipotezy zerowej.



Ryc. 4.29. Drzewo (a) jest identyczne z drzewem (b) jedynie wówczas, gdy $b = b' = 0$ (c), bez względu na wartość c . Drzewo (c) jest więc „drzewem zerowym” hipotezy o identyczności drzew (a) i (b)

Felsenstein (1988b) zaproponował testowanie hipotezy H_0 : $b = 0$ znanym nam testem proporcji wiarygodności, bowiem gdy dla jednego z drzew $b = 0$, a dla drugiego $b \geq 0$, to pierwsze z drzew jest szczególnym (prostszym) przypadkiem drugiego, a więc warunki testu są spełnione. Z drugiej strony, w przypadku analizy filogenetycznej nie wydaje się to uzasadnione, gdyż – jak już pisaliśmy w Rozdziale 4.6 – przy porównaniach topologii funkcja wiarygodności nie spełnia warunków ciągłości i różniczkowalności (Goldman 1993, Yang 1994c, Yang i inni 1995). Symulacje wykazały, że choć test wydaje się poprawny, gdy używa się przy rekonstrukcji odpowiedniego modelu substytucji, to może też wskazać na wysoką istotność statystyczną długości gałęzi błędnego drzewa, gdy przyjęto błędny model (Tateno i inni 1994, Gaut i Lewis 1995, Zhang 1999).

Poparcie dla poszczególnych kładów ocenia się też – powszechnie, wręcz rutynowo – stosując rozmaite techniki randomizacji, omówione w Rozdziale 2.8: permutacje, *jackknifing*, jak też techniki Monte Carlo, z których najczęściej używa się *bootstrap*. Na wstępie jednak warto zauważyć, że choć zachowanie tych technik najlepiej poznano dla kladogramów obliczonych techniką redukcjonistyczną, to wszystkie zastrzeżenia sformułowane wyżej dla testu PTP i tutaj pozostają w mocy: dane oryginalne nie są losowe, a przeciwstawianie starannie dobranym kladystycznym cechom losowych zestawów danych wydaje się już w swej istocie nieuzasadnione; ponadto założenia statystycznych testów istotności dla kladystycznych danych nie są spełniane, więc formalne poziomy istotności nie mogą być w ten sposób obliczane. Oryginalne dane powinny

być losową próbą wszystkich możliwych cech, co jest wręcz odwrotnością kładystycznego wyboru cech. Zwolennicy technik randomizacji twierdzą, że założenie to jest spełniane dzięki wysiłkom taksonomów, wybierających cechy wzajemnie niezależne, jednak takie starania nie zapewniają przecież postulowanej niezależności między cechami.

Jak wiemy, wielkość błędów losowych, będących następstwem ograniczonej reprezentatywności analizowanych danych, ocenić można, pobierając szereg prób, co umożliwi następnie obliczenie wariancji estymowanych parametrów – jak długość gałęzi rekonstruowanego drzewa. W praktyce jednak dysponujemy jednym, ograniczonym zestawem danych i więcej nie ma skąd wziąć. W takiej sytuacji pozostaje próbkowanie numeryczne, czyli stworzenie z posiadanej próby – zestawu danych – dużej liczby niyprób i użycie ich następnie dla estymacji wariancji. **Bootstrap** jako technika oceny wiarygodności drzew i/lub kładów, zaproponowany przez Felsensteina (1985), doczekał się bogatej literatury (Hedges 1992, Felsenstein i Kishino 1993, Hillis i Bull 1992, Zharkikh i Li 1992a, b, Li i Zharkikh 1994). **Bootstrap** w analizie redukcjonistycznej tworzy niypróby tej samej wielkości co dane oryginalne, poprzez losowanie z powtórzeniami n elementów (patrz Rozdział 2.8). Zwykle takich niyprób generuje się 1000 lub więcej (znów: im więcej, tym lepiej), dla każdej niypróby znajduje MPR, po czym oblicza drzewo zgodności z większością, zwykle przyjmując poziom 50%. Procent drzew, na których występuje dany kład, określa stopień poparcia tego kładu, np. 80%. Takie drzewo zgodności w pewnych warunkach może być bliższe rzeczywistej filogenezie niż drzewo obliczone dla danych oryginalnych (Berry i Gascuel 1996). Felsenstein i Kishino (1993) zaproponowali, by poparcie interpretować jako miarę prawdopodobieństwa, że określona gałąź występuje na drzewie rzeczywiście odwierciedlającym odtwarzaną filogenezę.

Bootstrap wymaga spełnienia szeregu warunków. Po pierwsze klady, których monofiletyczność testujemy, powinny być określone *a priori*, jeżeli poziomy istotności mają mieć jakiś formalny sens (Swofford i Olsen 1990). To ten sam problem co w testach *post hoc* w analizie wariancji: gwałtowny wzrost błędu I rodzaju, ponad wszelkie akceptowalne poziomy. Po drugie, estymaty poziomów ufności są zaledwie szacunkami, jak długo próba – czyli liczba cech – nie jest duża. W tym miejscu nieuchronnie spotkamy się z odmiennym rozumieniem dużych danych przez taksonomów i statystyków: dla taksonoma ustalenie stanów stu cech to wielka praca, a często o wiele mniej danych w ogóle da się zebrać; dla statystyka cech powinno być nie mniej jak 10 000, ostatecznie 1000. Pozornie lepiej jest w przypadku danych molekularnych – sekwencje mogą być długie – jednak w rzeczywistości liczą się jedynie pozycje filogenetycznie informatywne (niektórzy dla **bootstrap** temu przeczą), a tych nigdy nie ma wiele.

Inny wariant techniki **bootstrap** (Nei i Kumar 2000) nie oblicza drzewa zgodności, a jedynie zlicza węzły drzew, dzielących drzewo na takie same partycje jak na drzewie obliczonym dla danych oryginalnych. Tak estymowane poparcie nazywa się **wartością ufności bootstrap** P_B (*bootstrap confidence value*) bądź po prostu **wartością bootstrap**. Dla drzew addytywnych obliczonych dla bliskich sobie taksonów P_B ma podobne wartości jak P_C obliczone metodą analityczną, choć estymaty techniką **bootstrap** są konserwatywne, więc P_B ma tendencję do przyjmowania niższych wartości niż P_C (Sitnikova i inni 1995, Sitnikova 1996). Przewagą tego wariantu jest możliwość bardziej precyzyjnego sformułowania hipotezy H_0 . Nie stosuje się go raczej dla drzew znalezo-

nych metodą redukcjonistyczną, bowiem na ogół jest więcej niż jeden MPR, więc drzewo zgodności staje się nieodzowne. Gdy niektóre wewnętrzne gałęzie drzewa mają niskie wartości poparcia, warto zastanowić się, czy nie należy przedstawić **skondensowanego drzewa** (*condensed tree*), na którym długość wszystkich gałęzi o niskim poparciu (wartość wymaganego minimalnego poparcia P_B , albo poparcia w rozumieniu Felsensteina można odpowiednio dobrać) przyjmuje się za równą zero, czyli tworzy politomię: w ten sposób uzyskujemy drzewo, które przedstawia jedynie tę część rekonstrukcji filogenezy, która jest bardziej uzasadniona. Jakkolwiek skondensowane drzewo może przypominać drzewo zgodności, to jednak nim nie jest.

Chociaż formalne poziomy istotności w praktyce nie są estymowane, to porównanie poparcia dla różnych kładów jest użyteczne. Trzeba jednak pamiętać, że „poparcie kładu” nie jest wielkością mającą jakieś konkretne znaczenie w statystyce. Ponadto jest to znów test jednostronny: klady pojawiające się na zdecydowanej większości MPR rzeczywiście są stabilne, poparte wielu synapomorfiami. Z drugiej strony niski procent poparcia nie wyklucza wiarygodności kładu: jak wiemy, wystarcza jedna pewna synapomorfia, aby kład wyodrębnić, choć oczywiście taki kład będzie miał bardzo niskie poparcie obliczone techniką *bootstrap*. Wartości poparcia zależą też od wielkości zestawu danych, wydajności przybliżonych technik znajdowania najkrótszego drzewa, topologii kladogramu, zróżnicowania częstości zmian między gałęziami. Gdy jakiś takson/klad w kolejnych drzewach *bootstrap* przemieszcza się od kładu do kładu, to obniża on poparcie tym kładom, nawet gdy poza tym są one stabilne. Dla sekwencji *bootstrap* zakłada m.in., że pozycje sekwencji są niezależne i mają takie same rozkłady częstości (*independent and identically distributed: i.d.d.*). Cummings i inni (1995) sprawdzili to założenie: zrekonstruowali drzewo dla 10 kompletnych mitochondrialnych genomów, po czym przeprowadzili takie same rekonstrukcje dla ciągłych fragmentów genomów (co przypomina sytuację, gdy dysponujemy sekwencjami dla jednego lub paru genów mtDNA) oraz dla losowo wybranych pozycji genomu. Okazało się, że drzewa obliczone dla losowo wybranych pozycji były bliższe drzewom zrekonstruowanym dla całych mitochondrialnych genomów niż te obliczone dla ciągłych fragmentów. Wskazuje to, że założenie i.d.d. dla mtDNA nie jest spełnione, ponadto każde ostrożnie interpretować zarówno same filogenezy zrekonstruowane na podstawie jednego lub paru mitochondrialnych genów, jak i estymaty ich wiarygodności otrzymane techniką *bootstrap*. W każdym razie, gdy dane są niereprezentatywne lub metoda rekonstrukcji niespójna, *bootstrap* nie usunie tego obciążenia wyników, tak jak nie wykaże niskich wartości poparcia dla błędnych topologii, znalezionych w następstwie przyciągania długich gałęzi. W technikach opartych na odległościach *bootstrap* wykonujemy na wyjściowych danych i dla tak powstałych niyprób obliczamy odległości, używane następnie do znalezienia drzew.

Hillis i Bull (1993) badali własności estymatów otrzymanych techniką *bootstrap*, wykorzystując zarówno dane pochodzące z symulacji, jak i znanych filogenez organizmów laboratoryjnych. Stwierdzili, że wartości były nieobciążonymi lecz bardzo niedokładnymi estymatami powtarzalności gałęzi, a obciążonymi estymatami dokładności, czyli zgodności z rzeczywistą filogenezą. Jeżeli technika rekonstrukcji jest spójna, *bootstrap* daje zanizone estymaty dokładności, gdy osiąga wartości wysokie, a zawyżone, gdy ma wartości małe. Aby poprawić dokładność estymatów, zaproponowano dwa ulepszenia techniki. Rodrigo (1993) wprowadził iteratywny *bootstrap* (Hall i Martin 1988), czyli *bootstrap* dla wszystkich niyprób, obliczonych w pierwszej fazie

bootstrap oryginalnych danych; technika jest komputerowo bardzo intensywna. Zharkikh i Li (1995) wykazali, że podobną poprawkę obliczyć można dwoma cyklami *bootstrap* na oryginalnych danych: jeden przeprowadza się na kompletnych danych, drugi na mniejszej liczbie cech. Oba estymaty, z poprawką na wielkość próby, pozwalają obliczyć skorygowane estymaty dokładności; symulacje tych badaczy wykazały, że technika zmniejsza obciążenie estymatów, przynajmniej gdy cech informatywnych w oryginalnych danych jest co najmniej 100.

Statystyczne własności *bootstrap* są złożone i niejasne, choć były wielokrotnie badane (Zharkikh i Li 1992a, Li i Zharkikh 1994, Felsenstein i Kishino 1993, Hillis i Bull 1993, Sitnikova i inni 1995, Efron i inni 1996). Najprościej określić je dla techniki dołączania sąsiada. Gdy każda pozycja sekwencji ewoluuje niezależnie, odległość użyta jest nieobciążonym estymatorem liczby substytucji nukleotydów, zarówno T , jak i n są dostatecznie duże, a hipotezę H_0 formułujemy, zakładając, że długości każdej wewnętrznej gałęzi równe są zeru, to wartość P_B dla określonej gałęzi powinny odpowiadać P_C . Raczej tak jest dla dużego T (Efron i inni 1996), jak też dla małych wartości T i długiej gałęzi wewnętrznej (Sitnikova i inni 1995), lecz gdy T ma niską wartość i rzeczywista długość wewnętrznej gałęzi jest bliska zeru, P_B ma tendencję do bycia zaniżonym estymatem P_C , gdy P_C jest bliskie 1, a zawyżonym, gdy P_C ma wartość niską (Sitnikova i inni 1996). Dla technik redukcjonistycznej i maksymalizacji wiarygodności zachowanie P_B wydaje się podobne (Nei i Kumar 2000). W każdym przypadku *bootstrap* daje konserwatywne estymaty poparcia, co przy niekompletności naszej wiedzy o rzeczywistych procesach ewolucyjnych wydaje się do przyjęcia. *Bootstrap* pozwala też np. uniknąć długich obliczeń metodą minimalnej ewolucji, gdy obliczone techniką dołączania sąsiada pojedyncze drzewo ma gałęzie, dla których poparcie jest wysokie. Oczywiście może być tak i dla błędnego drzewa, gdy użyto niewłaściwej odległości, lecz tego błędu również nie unikniemy, stosując technikę minimalnej ewolucji. Z kolei gałęzie o dużym poparciu będą obecne i na najlepszym drzewie znalezionym intensywną obliczeniowo metodą minimalnej ewolucji, a gałęzie o poparciu niskim i tak pozostaną mało wiarygodne pomimo długich obliczeń.

Jak wspomnieliśmy, dla technik redukcjonistycznej i maksymalizacji wiarygodności długości gałęzi są zawsze dodatnie, więc nie potrafimy sformułować takiego testu jak dla metod opartych na odległościach. Możemy jednak użyć **testu *bootstrap* dla wewnętrznych gałęzi** (*bootstrap interior branch test*), zaproponowanego przez Dopazo (1994). *Bootstrap* macierzy danych przeprowadza się tak samo jak opisaliśmy, uzyskując kilkaset nibyprób, po czym dla każdej oblicza się długość gałęzi, nie zmieniając jednak topologii drzewa obliczonego z danych oryginalnych. W ten sposób dla każdej z wewnętrznych gałęzi uzyskujemy kilkaset wartości, w tym również ujemnych, możemy więc sformułować hipotezę H_0 : długość danej gałęzi ≤ 0 , po czym użyć standardowej statystyki Z , jak opisaliśmy to dla drzew opartych na odległościach. I ta technika jednakże zawodzi, gdy badanych cech (np. nukleotydów w sekwencji) jest niewiele.

W ocenie wiarygodności drzew obliczonych techniką ML stosuje się też **parametryczny *bootstrap***, będący w pewnym sensie połączeniem próbkowania numerycznego i symulacji. Dla rzeczywistych danych opisujących zestaw taksonów przyjmujemy różne parametry modelu ewolucji, dla każdego z nich generujemy po, powiedzmy, 1000 symulowanych zestawów danych tworzonych zgodnie z danym modelem, dla każdego z zestawów znajdujemy drzewo, po czym określamy częstości różnych topologii. Jeżeli np. pomimo dobrania parametrów modelu odpowiadających określonej

topologii ta topologia występuje rzadko wśród drzew obliczonych dla nibyprób, to dane nie popierają tej topologii – którą na podstawie innych danych uważamy za najbliższą rzeczywistości – i nie jest to wynikiem błędów losowych, a raczej błędu systematycznego, jak np. przyciąganie długich gałęzi. Podobnie testować możemy np. obciążenie rekonstrukcji różnymi częstościami nukleotydów, jak zdecydowana przewaga A i T u nietoperzy, potencjalnie będąca następstwem adaptacji, a więc konwergencją (Pettigrew 1994). Van Den Bussche i inni (1996; cytowani za Swofford i inni 1996) wykazali, że to obciążenie samo nie wystarcza dla wytłumaczenia obserwowanej monofiletyczności nietoperzy, która i bez tych wysokich frekwencji A i T wynikałaby z analizowanych sekwencji. Huelsenbeck i Crandall (1997) zaproponowali dla drzew obliczonych techniką maksymalizacji wiarygodności testowanie metodą parametrycznego *bootstrap* poziomu istotności zdefiniowanych wcześniej grup monofiletycznych, choć nie jest to test rygorystyczny statystycznie (Nei i Kumar 2000), zwłaszcza że hipotezę H_0 trudno zdefiniować dla parametrycznego *bootstrap* użytego do filogenetycznych drzew, a proponowany test może okazać się zbyt konserwatywny bądź zbyt liberalny, zależnie od rozpatrywanej grupy taksonomicznej, zwłaszcza że możemy błędnie dobrać model substytucji, a wszystkie istniejące modele są niedoskonałe. Lepiej więc przeprowadzić standardowy *bootstrap*.

Inna technika to *jackknife*, czyli tworzenie z próby o liczebności n zestawu n nibyprób o liczebności $n - 1$, poprzez losowanie bez powtórzeń (patrz Rozdział 2.8). Zestaw nibyprób służy do obliczeń MPR. *Jackknife* stosować możemy do cech lub do taksonów. Lanyon (1985) zaproponował *jackknife* taksonów: gdy w danych brak homoplazji, MPR dla $T - 1$ taksonów będzie identyczny z MPR dla T taksonów, jedynie gałęzi kończącej się usuniętym taksonem będzie brakować; w takim przypadku MPR znajdowałibyśmy już za pierwszym dołączeniem do drzewa kolejnych taksonów. Tak jednak właściwie nigdy nie jest, bowiem dane zawierają homoplazje: wówczas usunięcie jednego taksonu może zmieniać topologię drzewa pozostałego po usunięciu tego taksonu, często drastycznie. Stąd jedną z technik oceny wiarygodności rekonstrukcji jest kolejne usuwanie taksonów (lub zawierających więcej niż jeden takson kładów): wrażliwość topologii na usunięcie kolejnych taksonów jest wskazówką, które części drzewa są lepiej uzasadnione danymi. Oczywiście wrażliwość topologii na usuwanie taksonów może być następstwem zarówno losowych, jak i systematycznych błędów, lecz – zwłaszcza gdy topologia zmienia się zdecydowanie po usunięciu taksonu kończącego długą gałąź – znaczny błąd systematyczny jest bardzo prawdopodobny. Formalnie dla T rekonstrukcji o $T - 1$ taksonach obliczyć należy drzewo pełnej zgodności i zobaczyć, które klady są na nim obecne. Wiemy jednak, że drzewo pełnej zgodności byłoby wówczas zawsze politomią, jedynie grupa zewnętrzna – której się nie pomija w *jackknife* – byłaby odrębna, więc Lanyon (1985) zaproponował modyfikację techniki obliczeń drzewa pełnej zgodności, tak aby obok wspólnych uwzględniała też węzły niesprzeczne, czyli brakujące kolejno klady. Problemem jest jednak to, że drzewo pełnej zgodności nie uwzględnia częstości różnych kładów w nibypróbach: pojedyncza topologia skrajnie różna od pozostałych, nawet całkowicie między sobą zgodnych, przekształca drzewo zgodności w jedną politomię. Dla ominięcia tej niedogodności, Siddall (1996) zaproponował *jackknife* indeks monofiletyczności (*jackknife monophyly index* – JMI):

$$\text{JMI}_c = \frac{\sum_{t=1}^T \rho(c_t)}{T},$$

gdzie T to liczba taksonów grupy wewnętrznej, a $\rho(c_t)$ to proporcja kladogramów MPR niabypróby t , które popierają kład c . JMI oblicza się dla ukorzenionych kladogramów, toteż *jackknife* też nie obejmuje grupy zewnętrznej. JMI to znów test jednostronny: wskazuje na klady dobrze poparte danymi, lecz nie wyklucza poparcia pozostałych kładów. Jest też pomocny we wskazywaniu kładów krytycznych i problematycznych: usunięcie kładów krytycznych gwałtownie zwiększa liczbę MPR, gdy usunięcie kładów problematycznych zmniejsza tę liczbę. Tak więc klady krytyczne zwiększają jednoznaczność rekonstrukcji, gdy problematyczne – zmniejszają. *Jackknife* można też przeprowadzać, zamiast dla pojedynczych taksonów, dla całych ich grup, jednak możliwych kombinacji jest tu wiele (dla T taksonów liczba możliwych wydzieleni grup, o liczebności do $T - 1$, równa jest $2^T - 2$), co ogranicza możliwości obliczeń. *Jackknife* dla cech wprowadzili Mueller i Ayala (1982), estymując wariancję odległości genetycznej Nei. W kladystyce takie użycie *jackknife* nie daje jednak nic więcej ponad wskazanie, czy więcej niż jedna synapomorfia popiera dany kład – a to można stwierdzić na samym drzewie, można też obserwować zmiany topologii po kolejnym usuwaniu cech i ich różnych kombinacji (Davis i inni 1993).

Wydaje się, że pomimo wszystkich sformułowanych wyżej zastrzeżeń najlepszą techniką oceny wiarygodności jest nieparametryczny *bootstrap*, toteż warto podać wartości poparcia dla każdej gałęzi. Niska wartość poparcia nie oznacza, że istnienie danej gałęzi musimy odrzucić, prowadząc badania nad procesami ewolucyjnymi. Jeżeli jednak istnienie określonej gałęzi ma być podstawą do decyzji taksonomicznych, zwłaszcza radykalnie zmieniających dotychczasową systematykę, wartość poparcia musi być wysoka, co najmniej około 95%. Wiąże się to z tym, że systematyka – jakkolwiek nowatorskich technik by nie używała – ze swej natury musi być konserwatywna, bowiem korzystają z niej wszystkie nauki biologiczne i nie wolno do niej wprowadzać chaosu.

5. Programy komputerowe

Współczesna analiza filogenetyczna – a także fenetyczna – wymaga użycia programów komputerowych. Jest ich obecnie paręset i wciąż powstają nowe. Felsenstein (<http://evolution.genetics.washington.edu/phylip/software.html>) wymienia i krótko omawia przeszło 190 programów w dokumentacji do pakietu PHYLIP (main.doc), wskazując źródła ich uzyskania. Takie informacje znaleźć można też na stronach: <http://www.nhm.ac.uk/hennig/software.html>, <http://mep.bio.psu.edu>, <http://www.oup-usa.org/sc/0195135857>. W każdym przypadku podane są odsyłacze do innych stron z oprogramowaniem lub informacją o nim, toteż łatwo dotrzeć do najnowszych wersji programów, dlatego tutaj zamieścimy informację o zaledwie kilku najczęściej używanych.

Wiele programów można po prostu ściągnąć z sieci, bez jakichkolwiek opłat, inne trzeba kupić. Ceny nie są wysokie, zwykle najwyżej nieco ponad 100 dolarów amerykańskich. W zamian otrzymujemy zwykle program bardziej dopracowany, lepiej sprawdzony i możliwość z korzystania pomocy w razie problemów technicznych. To nie znaczy oczywiście, że dostępne bezpłatnie programy są gorsze.

COMPONENT (Page 1993); Macintosh OS, PC MS Windows; porównywanie drzew i obliczanie drzew zgodności dla analiz koewolucji i biogeograficznych; dostępny bez opłat: <http://evolve.zps.ox.ac.uk/Rod/cpw.html>.

FREQPARS (Swofford i Berlocher 1987); PC MS DOS i kod źródłowy; analiza filogenetyczna częstości techniką *frequency parsimony*; dostępny bez opłat przez anonymous ftp z onyx.si.edu.

HENNIG86 (Farris 1988); PC MS DOS; choć dość stary i niezbyt wygodny w użyciu, jest to szybki i wydajny program znajdujący najkrótsze drzewa techniką redukcjonistyczną, dostępne rozszerzenia; aby kupić, należy skontaktować się z A. Kluge (arnold.g.kluge@um.cc.umich.edu) lub Department of Biological Sciences, George Washington University, Washington DC (biodl@gwuvvm.gwu.edu).

MACCLADE (Maddison i Maddison 1992); Macintosh OS; doskonały program do interaktywnej manipulacji drzewami i cechami, do badania ewolucji cech i kładów techniką redukcjonistyczną, natomiast funkcje poszukiwania najkrótszego drzewa mocno uproszczone, współpracuje z PAUP (ten sam format plików) i doskonale go uzupełnia; można go kupić w Sinauer Associates, Sunderland, Massachusetts (orders@sinauer.com).

- MEGA2 (Kumar i inni 2000); PC MS Windows; odległości dla kwasów nukleinowych i białek, UPGMA, drzewa addytywne, technika redukcjonistyczna, analizuje cechy molekularne; dostępny bez opłat: <http://www.oup-usa.org/sc/0195135857>.
- NTSYSpc (Rohlf 1998); PC MS Windows; wygodny w użyciu i wydajny program do analiz fenetycznych, obejmujący wszystkie metody opisane w Rozdziale 3 oraz dołączanie sąsiada i obliczanie drzew zgodności; można go kupić w Exeter Software, Seaton, New York (ExeterSftw@aol.com).
- PAML (Yang 1999); kod źródłowy w języku C; zestaw programów do obliczeń techniką maksymalizacji wiarygodności dla protein i kwasów nukleinowych, rekonstruuje ancestralne sekwencje, analizy dla wielu genów jednocześnie, symulacje drzew, estymuje matryce substytucji, zmienność częstości substytucji między pozycjami i ancestralne stany techniką redukcjonistyczną; dostępny bez opłat przez anonymous ftp z: <ftp.bio.indiana.edu/molbio/evolve>.
- PAUP (Swofford 1998); Macintosh OS, PC MS DOS, UNIX, VAX, VMS; znajduje drzewa techniką redukcjonistyczną, minimalnej ewolucji, a dla DNA także maksymalizacji wiarygodności, oblicza wiele testów wiarygodności drzew i ich gałęzi, drzewa zgodności i stany ancestralne cech, zapewne najczęściej używany program do analiz filogenetycznych i najlepiej nadający się do analiz danych molekularnych, choć jest wygodny i dla danych morfologicznych, współpracuje z MACCLADE; można go kupić (niestety nadal jedynie wciąż testowaną wersję beta, bez podręcznika) w Sinauer Associates, Sunderland, Massachusetts (orders@sinauer.com).
- PHYLIP (Felsenstein 2000); Macintosh OS, PC MS DOS, PC Windows, kod źródłowy w języku C; zestaw 30 programów obliczających drzewa techniką redukcjonistyczną, maksymalizacji wiarygodności (dla nukleotydów, protein, miejsc restrykcyjnych, frekwencji genotypów i innych cech ciągłych), niezmiennych Lake'a, technik opartych na odległościach, oblicza rozmaite statystyczne testy dla drzew, drzewa zgodności, odległości genetyczne; zapewne jest to pakiet obejmujący najwięcej technik, zorientowanych raczej na dane molekularne, choć niezbyt wygodny w użyciu i dość powolny; dostępny bez opłat przez anonymous ftp z: evolution.genetics.washington.edu/pub/phylip, albo z World Wide Web: <http://evolution.genetics.washington.edu/phylip/html>.
- RANDOMCLADISTICS (Siddall 1996b); PC MS DOS; testy permutacyjne, *bootstrap* i *jackknifing* dla drzew obliczonych za pomocą HENNIG86, dokumentacja zawiera jasny podręcznik dla tamtego programu; dostępny bez opłat przez anonymous ftp z: zoo.utoronto.ca/pub.

6. Wybrana literatura

- Adachi J. i Hasegawa M. 1996. *MOLPHY: Programs for molecular phylogenetics*. Institute of Statistical Mathematics, Tokyo.
- Adams E.N. 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology* **21**: 390–397.
- Akaike H. 1974. A new look at the statistical model identifications. *IEEE Transactions Automatic Control* **AC-19**: 716–723.
- Allard M.W., Farris J.S. i Carpenter J.M. 1999. Congruence among Mammalian Mitochondrial Genes. *Cladistics* **15**: 75–84.
- Almeida M.D. i Bisby F.A. 1984. A simple method for establishing taxonomic characters from measurement data. *Taxon* **22**: 405–409.
- Alroy J. 1994. Four Permutation Tests for the Presence of Phylogenetic Structure. *Systematic Biology* **43**: 430–437.
- Anderberg M.R. 1973. *Cluster Analysis for Applications*. Academic Press, New York.
- Anderberg A. i Tehler A. 1990. Consensus trees, a necessity in taxonomic practice. *Cladistics* **6**: 399–402.
- Archie J. 1985. Methods for coding variable morphological features for numerical taxonomic analysis. *Systematic Zoology* **34**: 326–345.
- Archie J. 1989. Homoplasy excess ratios: New indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. *Systematic Zoology* **38**: 239–252.
- Avise J. C. 2000. *Phylogeography. The history and formation of species*. Harvard University Press, Cambridge, Massachusetts–London, England.
- Balakrishnan V. i Sanghvi L. D. 1968. Distance between populations on the basis of attribute data. *Biometrics* **24**: 859–865.
- Bandelt H.-J. i Dress A.W.M. 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* **1**: 242–252.
- Barrett M., Donoghue M.J. i Sober E. 1991. Against consensus. *Systematic Zoology* **40**: 486–493.
- Barrodale I i Roberts F.D.K. 1973. An improved algorithm for discrete l_1 linear approximation. *SIAM Journal of Numerical Analysis* **10**: 839–848.
- Barry D. i Hartigan J.A. 1987. Asynchronous distance between homologous DNA sequences. *Biometrics* **43**: 261–276.
- Barthélemy J.-P. i McMorris F. R. 1986. The median procedure for n-trees. *Journal of Classification* **3**: 329–334.
- Barthélemy J.-P. i Monjardet B. 1981. The median procedure in cluster analysis and social choice theory. *Mathematical Social Sciences* **1**: 235–267.
- Baum B.R. 1988. A simple procedure for establishing discrete characters from measurement data, applicable to cladistics. *Taxon* **37**: 63–70.

- Baum B.R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41**: 3–10.
- Berry V. i Gascuel O. 1996. On the interpretation of bootstrap trees: Appropriate threshold of clade selection and induced gain. *Journal of Molecular Evolution* **13**: 999–1011.
- Bininda-Emonds O.R.P., Gittleman J.L. i Purvis A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Review* **74**: 143–175.
- Bookstein F., Chernoff B., Elder R., Humphries J., Smith G. i Strauss R. 1985. *Morphometrics in Evolutionary Biology. The Geometry of Size and Shape Change, With Examples from Fishes. Special Publication 15*, The Academy of Natural Sciences of Philadelphia, Philadelphia.
- Bremer K. 1988. The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42**: 795–803.
- Bremer K. 1990. Combinable component consensus. *Cladistics* **6**: 369–372.
- Bremer K. 1994. Branch support and tree stability. *Cladistics* **10**: 295–304.
- Britten R.J., Baron W.F., Stout D.B. i Davidson E.H. 1988. Sources and evolution of human *Alu* repeated sequences. *Proceedings of the National Academy of Sciences of the USA* **85**: 4770–4774.
- Brooks D.R. 1988. Macroevolutionary comparisons of host and parasite phylogenies. *Annual Review of Ecology and Systematics* **19**: 235–259.
- Brooks D.R. 1990. Parsimony analysis in historical biogeography and coevolution: Methodological and theoretical update. *Systematic Zoology* **39**: 14–30.
- Brooks D.R. i McLennan D.A. 1991. *Phylogeny, Ecology and Behavior: A Research Program in Comparative Biology*. University of Chicago Press, Chicago.
- Brooks D.R. i McLennan D.A. 1993. *Parascript: Parasites and the Language of Evolution*. Smithsonian Institution Press, Washington.
- Brown W.M., Prager E.M., Wang A. i Wilson A.C. 1982. Mitochondrial DNA sequences of primates: Tempo and mode of evolution. *Journal of Molecular Evolution* **18**: 225–239.
- Bryant H.N. 1992. The role of permutation tail probability tests in phylogenetic systematics. *Systematic Biology* **41**: 258–263.
- Bull J.J., Huelsenbeck J.P., Cunningham C.W., Swofford D.L. i Waddell P.J. 1993. Partitioning and combining data in phylogenetic analysis. *Systematic Biology* **42**: 384–397.
- Buth D.G. 1984. The application of electrophoretic data in systematic studies. *Annual Review of Ecology and Systematics* **15**: 501–522.
- Cain A.J. i Harrison G.A. 1958. An analysis of the taxonomist's judgement of affinity. *Proceedings of the Zoological Society of London* **131**: 85–98.
- Camin J.H. i Sokal R.R. 1965. A method for deducing branching sequences in phylogeny. *Evolution* **19**: 311–326.
- Carpenter J.M. 1988. Choosing among multiple equally parsimonious cladograms. *Cladistics* **4**: 291–296.
- Carpenter J.M. 1992. Random cladistics. *Cladistics* **8**: 147–153.
- Cavalli-Sforza L.L. i Edwards A.W.F. 1967. Phylogenetic analysis: Models and estimation procedures. *Evolution* **32**: 550–570 i *American Journal of Human Genetics* **19**: 233–257.
- Cavender J.A. 1989. Mechanized derivation of linear invariants. *Molecular Biology and Evolution* **6**: 301–316.
- Cavender J.A. i Felsenstein J. 1987. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification* **4**: 57–71.
- Chappill J.A. 1989. Quantitative characters in phylogenetic analysis. *Cladistics* **5**: 217–234.
- Charleston M.A. 1994. *Factors affecting the performance of phylogenetic methods*. Ph. D. dissertation, Massey University; cytowane za Swofford i inni (1996).
- Chippindale P.T. i Wiens J.J. 1994. Weighting, Partitioning, and Combining Characters in Phylogenetic Analysis. *Systematic Biology* **43**: 278–287.
- Clark P.J. 1952. An extension of the coefficient of divergence for use with multiple characters. *Copeia* **2**: 61–64.
- Coddington J.A. 1988. Cladistic tests of adaptational hypotheses. *Cladistics* **4**: 3–22.
- Colless D.H. 1980. Congruence between morphological and allozyme data for *Menidia* species: a reappraisal. *Systematic Zoology* **29**: 288–299.

- Cooper A. i Penny D. 1997. Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. *Science* **275**: 1109–1113.
- Cummings M.P., Otto S.P. i Wakely J. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. *Molecular Biology and Evolution* **12**: 814–822.
- Czekanowski J. 1909. Zur Differentialdiagnose der Neandertalgruppe. *Korrespondenzblatt Deutsche Gesellschaft Anthropologie Ethnologie Urgesch.* **40**: 44–47.
- Czekanowski J. 1932. „Coefficient of racial likeness“ und „durchschnittliche Differenz“. *Anthropologisches Anzeiger* **9**: 227–249.
- Davis J.I. 1993. Character removal as a means for assessing the stability of clades. *Cladistics* **9**: 201–210.
- Davis J.I., Frohlich M.W. i Soreng R.J. 1993. Cladistic Characters and Cladogram Stability. *Systematic Botany* **18**: 188–196.
- Dayhoff M.O., Schwartz R.M. i Orcutt B.C. 1978. A model of evolutionary change in proteins. W: Dayhoff M.O. (red.), *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Silver Spring, Maryland: 345–352.
- deQueiroz A. 1993. For consensus (sometimes). *Systematic Zoology* **42**: 368–372.
- Dice L.R. 1945. Measures of the amount of ecologic association between species. *Ecology* **26**: 297–302.
- Dillon R.T., Jr. 1984. What shall I measure on my snails? Allozyme data and multivariate analysis used to reduce the non-genetic component of morphological variance in *Goniobasis proxima*. *Malacologia* **25**: 503–511.
- Dopazo J. 1994. Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. *Journal of Molecular Evolution* **38**: 300–304.
- Dunn G. i Everitt B.S. 1982. *An Introduction to Mathematical Taxonomy*. Cambridge University Press, New York.
- Eck R.V. i Dayhoff M.O. 1966. *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Silver Springs, Maryland.
- Edwards A.W.F. 1972. *Likelihood*. Cambridge University Press, Cambridge, UK.
- Edwards A.W.F. i Cavalli-Sforza L.L. 1964. Reconstruction of evolutionary trees. W: Heywood V.H. i McNesil J. (red.), *Phenetic and phylogenetic classification. Systematics Association Volume No. 6*. Systematics Association, London.
- Efron B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**: 1–26.
- Efron B., Halloran E. i Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the USA* **93**: 7085–7090.
- Emberton K.C. 1994. Allozyme Cladistics in Malacology: Why and How? *Nautilus*, Supplement **2**: 44–50.
- Estabrook G.F. 1983. The causes of character incompatibility. W: Felsenstein J. (red.), *Numerical Taxonomy. NATO ASI Series, Volume G1*, Springer Verlag, Berlin: 279–295.
- Estabrook G.F. i Landrum L. 1975. A simple test for the possible simultaneous evolutionary divergence of two amino acid positions. *Journal of the Mathematical Biology* **4**: 195–200.
- Everitt B.S. 1993. *Cluster Analysis*. 3rd edition. Edward Arnold, London.
- Everitt B.S. i Dunn G. 1992. *Applied Multivariate Data Analysis*. Oxford University Press, New York.
- Faith D.P. 1991. Cladistic permutation tests for monophyly and nonmonophyly. *Systematic Zoology* **40**: 366–375.
- Faith D.P. i Cranston P.S. 1991. Could a cladogram this short have arisen by chance alone? – On permutation tests for cladistic structure. *Cladistics* **7**: 1–28.
- Falniowski A. 1993. Gastropod Phylogenetic Torsion – Arising of a Class. *Folia Malacologica* **5**: 25–60.
- Falniowski A., Heller J., Szarowska M. i Mazan-Mamczarz K. 2002. Allozymic taxonomy within the genus *Melanopsis* (Gastropoda: Cerithiacea) in Israel: a case in which slight differences are congruent. *Malacologia* **44**: 307–324.
- Falniowski A., Kozik A., Szarowska M., Fiałkowski W. i Mazan K. 1996. Allozyme and morphology evolution in European Viviparidae (Mollusca: Gastropoda: Architaenioglossa). *Journal of Zoological Systematics and Evolutionary Research* **34**: 49–62.

- Falniowski A., Kozik A., Szarowska M., Rapala-Kozik M. i Turyna I. 1993. Morphological and allozymic polymorphism and differences among local populations in *Bradybaena fruticum* (O.F. Müller, 1777) (Gastropoda: Stylommatophora: Helicoidea). *Malacologia* **35**: 371–388.
- Falniowski A., Mazan K. i Szarowska M. 1999. Homozygote excess and gene flow in the spring snail *Bythinella* (Gastropoda: Prosobranchia). *Journal of Zoological Systematics and Evolutionary Research* **37**: 165–175.
- Falniowski A., Mazan K., Szarowska M. i Kozik A. 1997. Tracing the viviparid evolution: soft part morphology and opercular characters (Gastropoda: Architaenioglossa: Viviparidae). *Malakologische Abhandlungen, Staatliches Museum für Tierkunde Dresden* **18**: 193–211.
- Falniowski A. i Szarowska M. 1995. Can poorly understood new characters support a poorly understood phylogeny? Shell-structure data in Hydrobiid systematics (Mollusca: Gastropoda: Prosobranchia: Hydrobiidae). *Journal of Zoological Systematics and Evolutionary Research* **33**: 133–144.
- Farrell B. i Mitter C. 1990. Phylogenies of insect/plant interactions: Have *Phyllobrotica* leaf beetles (Chrysomelidae) and the Lamiales diversified in parallel? *Evolution* **44**: 1389–1403.
- Farris J.S. 1969a. On the cophenetic correlation coefficient. *Systematic Zoology* **18**: 279–285.
- Farris J.S. 1969b. A successive approximation approach to character weighting. *Systematic Zoology* **18**: 374–385.
- Farris J.S. 1970. Methods for computing Wagner Trees. *Systematic Zoology* **19**: 83–92.
- Farris J.S. 1972. Estimating phylogenetic trees from distance matrices. *American Naturalist* **106**: 645–668.
- Farris J.S. 1977. Phylogenetic analysis under Dollo's Law. *Systematic Zoology* **26**: 77–88.
- Farris J.S. 1988. *HENNIG86, Version 1.5*. New York: Port Jefferson Station.
- Farris J.S. 1989. The retention index and the rescaled consistency index. *Cladistics* **5**: 417–419.
- Farris J.S. 2000. Corroboration versus „Strongest Evidence“. *Cladistics* **16**: 385–393.
- Farris J.S., Källersjö M., Kluge A.G. i Bult C. 1994. Testing significance of incongruence. *Cladistics* **10**: 315–319.
- Farris J.S., Källersjö M., Albert V.A., Allard M., Anderberg A., Bowditch B., Bult C., Carpenter J.M., Crow T.M., De Laet J., Fitzhugh K., Frost D., Goloboff P.A., Humphries C.J., Jondelius U., Judd D., Karis P.O., Lipscomb D., Luckow M., Mindell D., Muona J., Nixon K.C., Presch W., Seberg O., Siddall M.E., Struwe L., Tehler A., Wenzel J., Wheeler Q.D. i Wheeler W. 1995. Explanation. *Cladistics* **11**: 211–218.
- Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401–410.
- Felsenstein J. 1981a. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**: 368–376.
- Felsenstein J. 1981b. Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. *Evolution* **35**: 1229–1242.
- Felsenstein J. 1984. Distance methods for inferring phylogenies: A justification. *Evolution* **38**: 16–24.
- Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**: 783–791.
- Felsenstein J. 1988a. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* **19**: 445–471.
- Felsenstein J. 1988b. Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* **22**: 521–565.
- Felsenstein J. 1997. An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology* **46**: 101–111.
- Felsenstein J. 2000. *PHYLIP (Phylogeny Inference Package), version 3.6*. Department of Genetics, University of Washington, Seattle.
- Felsenstein J. i Churchill G.A. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* **13**: 93–104.
- Felsenstein J. i Kishino H. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Systematic Biology* **42**: 193–200.
- Finden C.R. i Gorden A.D. 1985. Obtaining common pruned trees. *Journal of Classification* **2**: 255–276.
- Fisher R.A. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* **7**: 179–188.
- Fisher R.A. 1938. The Statistical Utilization of Multiple Measurements. *Annals of Eugenics* **8**: 376–386.

- Fitch W.M. 1970. Distinguishing homologous from analogous proteins. *Systematic Zoology* 19: 99–113.
- Fitch W.M. 1971. Towards defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology* 20: 406–416.
- Fitch W.M. 1977. On the problem of discovering the most parsimonious tree. *American Naturalist* 111: 223–257.
- Fitch W.M. 1979. Cautionary remarks on using gene expression events in parsimony procedures. *Systematic Zoology* 28: 375–379.
- Fitch W.M. 1981. A non-sequential method for constructing trees and hierarchical classifications. *Journal of Molecular Evolution* 18: 30–37.
- Fitch W.M. 1984. Cladistic and other methods: Problems, pitfalls, and potentials. W: Duncan T. i Stuessy T. G. (red.), *Cladistic Perspectives on the Reconstruction of Evolutionary History*. Columbia University Press, New York: 221–252.
- Fitch W.M. i Farris J.S. 1974. Evolutionary trees with minimum nucleotide replacements from amino acid sequences. *Journal of Molecular Evolution* 3: 263–278.
- Fitch W.M. i Margoliash E. 1967. Construction of phylogenetic trees. *Science* 155: 270–284.
- Fitch W.M. i Ye J. 1991. Weighted parsimony: Does it work? W: Miyamoto M.M. i Cracraft J. (red.), *Phylogenetic analysis of DNA sequences*. Oxford University Press, New York: 147–154.
- Florek K., Łukaszewicz J., Perkal J., Steinhaus H. i Zubrzycki S. 1951a. Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum* 2: 282–285.
- Florek K., Łukaszewicz J., Perkal J., Steinhaus H. i Zubrzycki S. 1951b. Taksonomia wrocławska. *Przegląd Antropologiczny* 17: 193–211.
- Fu J. i Murphy R.W. 1999. Discriminating and Locating Character Covariance: An Application of Permutation Tail Probability (PTP) Analyses. *Systematic Biology* 48: 380–395.
- Fukami K. i Tateno Y. 1989. On the maximum likelihood method for estimating molecular trees: Uniqueness of the likelihood point. *Journal of Molecular Evolution* 28: 460–464.
- Futuyama D.J. i McCafferty S.S. 1991. Phylogeny and the evolution of host plant associations in the leaf beetle genus *Ophraella* (Coleoptera, Chrysomelidae). *Evolution* 44: 1885–1913.
- Gascuel O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* 14: 685–695.
- Gaut B.S. i Lewis P.O. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. *Molecular Biology and Evolution* 12: 152–162.
- Goldman N. 1988. Methods for discrete coding of morphological characters in numerical analysis. *Cladistics* 4: 59–71.
- Goldman N. 1990. Maximum likelihood of phylogenetic trees, with special reference to Poisson process models of DNA substitution and to parsimony analysis. *Systematic Zoology* 39: 345–361.
- Goldman N. 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* 36: 182–198.
- Goldman N. i Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11: 725–736.
- Goloboff P.A. 1991. Homoplasy and the choice among cladograms. *Cladistics* 7: 215–232.
- Goloboff P.A. 1993. Estimating character weights during tree search. *Cladistics* 9: 83–91.
- Goloboff P.A. 1996a. Methods for faster parsimony analysis. *Cladistics* 12: 199–220.
- Goloboff P.A. 1996b. *PIWE version 2.51. MS-DOS program*. Opublikowane przez autora, San Miguel de Tucumán, Argentina.
- Gould S.J. 1977. *Ontogeny and phylogeny*. The Belknap Press of Harvard University Press, Cambridge, Massachusetts–London, England.
- Gower J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53: 325–338.
- Greenacre M.J. 1984. *Theory and Applications of Correspondence Analysis*. Academic Press, London.
- Gu X., Fu Y.-X. i Li W.-H. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology and Evolution* 12: 546–557.

- Gu X. i Zhang J. 1997. A simple method for estimating the parameter of substitution rate variation among sites. *Molecular Biology and Evolution* **15**: 1106–1113.
- Guigó R., Muchnik I. i Smith T.F. 1996. Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution* **6**: 189–213.
- Hafner M.S. i Nadler S.A. 1990. Cospeciation in host-parasite assemblages: Comparative analysis of rates of evolution and timing of cospeciation events. *Systematic Zoology* **39**: 192–204.
- Hafner M.S. i Page R.D.M. 1995. Molecular phylogenies and host-parasite cospeciation: gophers and lice as a model system. *Philosophical Transactions of the Royal Society of London B* **349**: 77–83.
- Hamann U. 1961. Merkmalbestandt und Verwandtschaftsbeziehungen der Farinosae. Ein Beitrag zum System der Monokotyledonen. *Willdenowia* **2**: 639–768.
- Hand D.J. 1981. *Discrimination and classification*. Wiley Interscience, New York.
- Harding E.F. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances of Applied Probability* **3**: 44–77.
- Hartl D.L. i Clark A.G. 1997. *Principles of Population Genetics*. Third Edition. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts.
- Harvey A.W. 1992. Three-taxon statements: more precisely, an abuse of parsimony? *Cladistics* **8**: 345–354.
- Hasegawa M. i Fujiwara M. 1993. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Molecular Phylogeny and Evolution* **2**: 1–5.
- Hasegawa M., Kishino H. i Saitou N. 1991. On the maximum likelihood method in molecular phylogenetics. *Journal of Molecular Evolution* **32**: 443–445.
- Hasegawa M., Kishino H. i Yano T.-A. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**: 160–174.
- Hall P. i Martin M. A. 1988. On bootstrap resampling and iteration. *Biometrika* **75**: 661–671.
- Hlstä O. i Björklund M. 1988. Nucleotide substitution models and estimation of phylogeny. *Molecular Biology and Evolution* **15**: 1381–1389.
- Hedges S.B. 1992. The number of replications needed for accurate estimation of the bootstrap *P* value in phylogenetic studies. *Molecular Biology and Evolution* **9**: 366–369.
- Hedges S.B., Parker P.H., Sibley C.G. i Kumar S. 1996. Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**: 226–229.
- Hendy M.D. 1991. A combinatorial description of the closest tree algorithm for finding evolutionary trees. *Discrete Mathematics* **96**: 51–58.
- Hendy M.D. i Penny D. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* **59**: 277–290.
- Hendy M.D. i Penny D. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology* **38**: 297–309.
- Hendy M.D. i Penny D. 1993. Spectral analysis of phylogenetic data. *Journal of Classification* **10**: 5–24.
- Hennig W. 1966. *Phylogenetic systematics*. University of Illinois Press, Urbana.
- Hillis D.M. 1984. Misuse and modification of Nei's genetic distance. *Systematic Zoology* **33**: 238–240.
- Hillis D.M. 1991. Discriminating between phylogenetic signal and random noise in DNA sequences. W: Miyamoto M.M. i Cracraft J. (red.), *Phylogenetic Analysis of DNA Sequences*. Oxford University Press, New York: 278–294.
- Hillis D.M., Allard M.W. i Miyamoto M.M. 1993. Analysis of DNA sequence data: phylogenetic inference. W: Zimmer E.A., White T.J., Cann R.L. i Wilson A.C. (red.), *Molecular Evolution: producing the biochemical data. Methods in Enzymology* **224**. Academic Press, San Diego, California: 456–487.
- Hillis D.M. i Bull J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**: 182–192.
- Hillis D.M., Bull J.J., White M.E., Badgett M.R. i Molineux I.J. 1992. Experimental phylogenetics: generation of known phylogeny. *Science* **255**: 589–592.
- Hillis D.M. i Huelsenbeck J.P. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. *Journal of Heredity* **83**: 189–195.
- Hillis D.M., Huelsenbeck J.P. i Swofford D.L. 1994. Hobgoblin of phylogenetics? *Nature* **369**: 363–364.

- Hillis D.M., Mable B.K., Larson A., Davis S.K. i Zimmer E.A. 1996. Nucleic Acids IV: Sequencing and Cloning. W: Hillis D.M., Moritz C. i Mable B.K. (red.), *Molecular Systematics*. Second Edition. Sinauer Associates, Inc., Sunderland, Massachusetts: 321–381.
- Hoberg E.P., Brooks D.R. i Seigel-Causey D. 1997. Host-parasite co-speciation: history, principles, and prospects. W: Clayton D.H. i Moore J. (red.), *Host-Parasite Evolution: General Principles and Avian Models*. Oxford University Press, Oxford: 212–235.
- Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24.
- Huelsenbeck J.P. 1991. Tree-length distribution skewness: An indicator of phylogenetic information. *Systematic Zoology* 40: 257–270.
- Huelsenbeck J.P. 1995. The performance of phylogenetic methods in simulation. *Systematic Biology* 44: 17–48.
- Huelsenbeck J.P. i Bull J.J. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. *Systematic Biology* 45: 92–98.
- Huelsenbeck J.P., Bull J.J. i Cunningham C.W. 1996. Combining data in phylogenetic analysis. *Trends in Ecology and Systematics* 11: 152–158.
- Huelsenbeck J.P. i Crandall K.A. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics* 28: 437–466.
- Huelsenbeck J.P. i Hillis D.M. 1993. Success of phylogenetic methods in the four-taxon case. *Systematic Biology* 42: 247–264.
- Huelsenbeck J.P., Hillis D.M. i Jones R. 1996. Parametric bootstrapping in molecular phylogenetics: applications and performance. W: Ferraris J.D. i Palumbi S.R. (red.), *Molecular Zoology: Advances, Strategies, and Protocols*. Wiley-Liss, New York: 19–45.
- Huey R.B. i Bennett A.F. 1987. Phylogenetic studies of coadaptation: Preferred temperatures versus optimal performance temperatures of lizards. *Evolution* 41: 1098–1115.
- Hull D.L. 1967. Certainty and circularity in evolutionary taxonomy. *Evolution* 21: 174–189.
- Jaccard P. 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* 44: 223–270.
- Jackson J.E. 1991. *A User's Guide to Principal Components*. Wiley Interscience, New York.
- Jajuga K. 1993. *Statystyczna analiza wielowymiarowa*. Biblioteka ekonometryczna. Wydawnictwo Naukowe PWN, Warszawa.
- Jin L. i Nei M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution* 7: 82–102.
- Johnson R.A. i Wichern D.W. 1998. *Applied Multivariate Statistical Analysis*. Fourth edition. Prentice Hall, Upper Saddle River, New Jersey.
- Jones D.T., Taylor W.R. i Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applied Biosciences* 8: 275–282.
- Jurka J. i Smith T. 1988. A fundamental division in the *Alu* family of repeated sequences. *Proceedings of the National Academy of Sciences of the USA* 85: 475–478.
- Kachigan S.K. 1991. *Multivariate Statistical Analysis. A Conceptual Introduction*. Second edition. Radius Press, New York.
- Kidd K.K., Astolfi P. i Cavalli-Sforza L.L. 1974. Error in the reconstruction of evolutionary trees. W: Crow J.F. i Denniston C. (red.), *Genetic Distance*. Plenum Press, New York: 121–136.
- Kidd K.K. i Sgaramella-Zonta L.A. 1971. Phylogenetic analysis: Concepts and methods. *American Journal of Human Genetics* 23: 235–252.
- Kim J. 1996. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing number of taxa. *Systematic Biology* 45: 363–374.
- Kim J., Rohlf F.J. i Sokal R.R. 1993. The accuracy of phylogenetic estimation using the neighbor-joining method. *Evolution* 47: 471–486.
- Kimura M. 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* 11: 247–269.
- Kimura M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16: 111–120.

- Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- Kimura M. i Ohta T. 1971. Protein polymorphism as a phase of molecular evolution. *Nature* **229**: 467–469.
- Kishino H. i Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution* **29**: 170–179.
- Kishino H., Miyata T. i Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution* **31**: 151–160.
- Kitching I.J., Forey P.L., Humphries C.J. i Williams D.M. 1998. *Cladistics. The Theory and Practice of Parsimony Analysis*. Second edition. The Systematics Association Publication No. 11. Oxford University Press, Oxford–New York–Tokyo.
- Kluge A.G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boiidae: Serpentes). *Systematic Zoology* **38**: 7–25.
- Kluge A.G. 1993. Three-taxon transformation in phylogenetic inference: ambiguity and distortion as regards explanatory power. *Cladistics* **9**: 246–259.
- Kluge A.G. 1994. Moving targets and shell games. *Cladistics* **10**: 403–413.
- Kluge A.G. i Farris J.S. 1969. Quantitative phyletics and the evolution of anurans. *Systematic Zoology* **18**: 1–32.
- Kluge A.G. i Wolf J. 1993. Cladistics: what's in a word? *Cladistics* **9**: 1–25.
- Kocher T.D. i Wilson A.C. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees: Control region and a protein-coding region. W: Osawa S. i Honjo T. (red.), *Evolution of Life*. Springer Verlag, New York: 391–413.
- Kruskal J.B. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**: 1–27.
- Kruskal J.B. 1964b. Nonmetric multidimensional scaling: a numerical method. *Psychometrika* **29**: 115–129.
- Krzanowski W.J. 1988. *Principles of Multivariate Analysis*. Clarendon Press, Oxford.
- Kuhner M.K. i Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* **11**: 459–468.
- Kumar S. 1996. A stepwise algorithm for finding minimum evolution trees. *Molecular Biology and Evolution* **13**: 584–593.
- Kumar S., Tamura K., Jakobsen I. i Nei M. 2000. *MEGA: Molecular evolutionary genetics analysis ver. 2*. Pennsylvania State University, University Park, Pennsylvania and Arizona State University, Tempe, Arizona.
- Kumar S., Tamura K. i Nei M. 1993. *MEGA: Molecular Evolutionary Genetics Analysis. Version 1.0*. Pennsylvania State University, University Park, Pennsylvania.
- Kumazawa Y. i Nishida M. 1995. Phylogenetic utility of mitochondrial transfer RNA genes for deep divergence in animals. W: Nei M. i Takahata N. (red.), *Current topics in molecular evolution*. Pennsylvania State University, University Park, Pennsylvania: 23–35.
- Lake J.A. 1987. Rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Molecular Biology and Evolution* **4**: 167–191.
- Lance G.N. i Williams W.T. 1967a. Mixed-data classificatory programs. I. Agglomerative systems. *Australian Computer Journal* **1**: 15–20.
- Lance G.N. i Williams W.T. 1967b. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer Journal* **9**: 373–380.
- Lanyon S. 1985. Detecting internal inconsistencies in distance data. *Systematic Zoology* **34**: 397–403.
- Lebart L., Morineau A. i Warwick K.M. 1984. *Multivariate descriptive statistical analysis*. Wiley Interscience, New York.
- Leitner T., Escanilla D., Franzen C., Uhlen M. i Albert J. 1996. Accurate reconstruction of a known HIV-1 transmission history. *Proceedings of the National Academy of Sciences of the USA* **93**: 10864–10869.
- Lento G.M., Hickson R.E., Chambers G.K. i Penny D. 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. *Molecular Biology and Evolution* **12**: 28–52.
- LeQuesne W.J. 1982. Compatibility analysis and its applications. *Zoological Journal of the Linnean Society* **74**: 267–275.

- Lewis P., Huelsenbeck J.P. i Swofford D.L. 1996. Maximum likelihood. W: Swofford D.L., *PAUP**, Version 4.0. Sinauer Associates, Sunderland, Massachusetts.
- Li W.H. i Gouy M. 1991. Statistical methods for testing phylogenies. W: Miyamoto M.M. i Cracraft J. (red.), *Phylogenetic Analysis of DNA Sequences*. Oxford University Press, New York: 249–277.
- Li W.-H. i Zharkikh A. 1994. What Is the Bootstrap Technique? *Systematic Biology* 43: 424–430.
- Li W.-H. i Graur D. 1991. *Fundamentals of molecular evolution*. Sinauer Associates, Inc., Sunderland, Massachusetts.
- Li W.-H., Wolfe K.H., Sourdiss J. i Sharp P.M. 1987. Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. *Cold Spring Harbor Symposium on Quantitative Biology* 52: 847–856.
- Lockhart P.J., Steel M.A., Hendy M.D. i Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution* 11: 605–612.
- Lyons-Weiler J., Hoelzer G.A. i Tausch R.J. 1996. Relative apparent synapomorphy analysis (RASA) I: the statistical measurement of phylogenetic signal. *Molecular Biology and Evolution* 13: 749–757.
- Łomnicki A. 1995. *Wprowadzenie do statystyki dla przyrodników*. Wydawnictwo Naukowe PWN, Warszawa.
- Mabee P.M. i Humphries J. 1993. Coding polymorphic data: Examples from allozymes and ontogeny. *Systematic Biology* 42: 166–181.
- MacQueen J.B. 1967. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* 1. University of California Press, Berkeley, CA: 281–297.
- Maddison D.R. 1990. *Phylogenetic inference of historical pathways and models of evolutionary change*. Ph. D. dissertation, Harvard University; cytowane za Swofford i inni (1996).
- Maddison D.R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Biology* 40: 315–328.
- Maddison D.R. 1994. Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters. *Annual Review of Entomology* 39: 267–292.
- Maddison W.P. 1990. A method for testing the correlated evolution of two binary characters: Are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution* 44: 539–557.
- Maddison W.P., Donoghue M.J. i Maddison D.R. 1984. Outgroup analysis and parsimony. *Systematic Zoology* 33: 83–103.
- Maddison W.P. i Maddison D.R. 1992. *MacClade: analysis of phylogeny and character evolution*, Version 3.0. Sinauer Associates, Sunderland, Massachusetts.
- Maddison W.P. i Slatkin M. 1991. Null models for the number of evolutionary steps in a character on a phylogenetic tree. *Evolution* 45: 1184–1197.
- Mahalanobis P.C. 1936. On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India*, 2: 49–55.
- Manly B.F.J. 1998. *Randomization, bootstrap and Monte Carlo methods in biology. Second edition*. Text in Statistical Science. Chapman & Hall, London–Weinheim–New York–Tokyo–Melbourne–Madras.
- Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27: 209–220.
- Mardulyn P. i Pasteels J.M. 1994. Coding Allozyme Data Using Step Matrices: Defining New original States for the Ancestral Taxa. *Systematic Biology* 43: 567–572.
- Margush T. i McMorris F.R. 1981. Consensus n-trees. *Bulletin of Mathematical Biology* 43: 239–244.
- Martins E. i Garland T. 1991. Phylogenetic analyses of the evolution of continuous characters: A simulation study. *Evolution* 45: 534–557.
- Matile L., Tassy P. i Goujet D. 1993. *Wstęp do systematyki zoologicznej. Koncepcje, zasady, metody*. Wydawnictwo Naukowe PWN, Warszawa.
- Mayr E. 1969. *Principles of systematic zoology*. McGraw-Hill, New York.
- Mayr E. 1974. *Populacje, gatunki i ewolucja*. Wiedza Powszechna, Warszawa.
- McArthur A.G. i Koop B.F. 1999. Partial 28S rDNA sequences and the antiquity of hydrothermal vent endemic gastropods. *Molecular Phylogeny and Evolution* 13: 255–274.

- Meier R. 1994. On the inappropriateness of presence / absence recoding for nonadditive multistate characters in computerized cladistic analyses. *Zoologischer Anzeiger* **232**: 201–212.
- Mickevich M.F. 1982. Transformation series analysis. *Systematic Zoology* **31**: 461–478.
- Mickevich M.F. i Farris J.S. 1981. The implications of congruence in *Menidia*. *Systematic Zoology* **30**: 351–370.
- Mickevich M.F. i Johnson M.F. 1976. Congruence between morphological and allozyme data in evolutionary inference and character evolution. *Systematic Zoology* **25**: 260–270.
- Mickevich M.F. i Mitter C. 1981. Treating polymorphic characters in systematics: A phylogenetic treatment of electrophoretic data. W: Funk V.A. i Brooks D.R. (red.), *Advances in cladistics*. Volume 1. New York Botanical Garden, New York: 45–60.
- Mickevich M.F. i Mitter C. 1983. Evolutionary patterns in allozyme data: A systematic approach. W: Platnick N.I. i Funk V.A. (red.), *Advances in cladistics*. Volume 2. Columbia University Press, New York: 169–176.
- Mickevich M.F. i Weller S.J. 1990. Evolutionary character analysis: Tracing character change on a cladogram. *Cladistics* **6**: 137–170.
- Mitter C. i Brooks D.R. 1983. Phylogenetic aspects of coevolution. W: Futuyma D. J. i Slatkin M. (red.), *Coevolution*. Sinauer Associates, Sunderland, Massachusetts: 65–98.
- Miyamoto M.M. 1985. Consensus cladograms and general classification. *Cladistics* **1**: 186–189.
- Miyamoto M.M., Allard M.W., Adkins R.M., Janecek L.L. i Honeycutt R.L. 1994. A congruence test of reliability using linked mitochondrial DNA sequences. *Systematic Biology* **43**: 236–249.
- Miyamoto M.M. i Fitch W.M. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology* **44**: 64–67.
- Moore G.W., Barnabas J. i Goodman M. 1973. A method for constructing maximum parsimony ancestral amino acid sequences on a given network. *Journal of Theoretical Biology* **38**: 459–485.
- Moritz C. i Hillis D.M. 1996. Molecular Systematics: Context and Controversies. W: Hillis D.M., Moritz C. i Mable B.K. (red.), *Molecular Systematics. Second Edition*. Sinauer Associates, Inc., Sunderland, Massachusetts: 1–13.
- Morrison D.F. 1990. *Multivariate Statistical Methods. Third edition*. McGraw-Hill, New York.
- Mueller L.D. i Ayala F.J. 1982. Estimation and interpretation of genetic distance in empirical studies. *Genetical Research* **40**: 127–137.
- Murphy R.W. 1993. The phylogenetic analysis of allozyme data: Invalidity of coding alleles by presence/absence and recommended procedures. *Biochemical Systematics and Ecology* **21**: 25–38.
- Murphy R.W., Sites J.W., Jr., Buth D.G. i Haufler H. 1996. Proteins: Isozyme Electrophoresis. W: Hillis D.M., Moritz C. i Mable B.K. (red.), *Molecular Systematics. Second Edition*. Sinauer Associates, Inc., Sunderland, Massachusetts: 51–120.
- Muse S.V. i Gaut B.S. 1994. A likelihood approach for comparing synonymous and nonsynonymous substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**: 715–724.
- Navidi W.C., Churchill G.A. i Haeseler A.V., von 1991. Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Molecular Biology and Evolution* **8**: 128–143.
- Neff N. 1986. A rational basis for a priori character weighting. *Systematic Zoology* **35**: 110–123.
- Nei M. 1972. Genetic distance between populations. *American Naturalist* **106**: 283–292.
- Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**: 583–590.
- Nei M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.
- Nei M. 1991. Relative efficiencies of different tree making methods for molecular data. W: Miyamoto M.M. i Cracraft J.L. (red.), *Recent advances in phylogenetic studies of DNA sequences*. Oxford University Press, Oxford, UK: 133–147.
- Nei M. 1996. Phylogenetic analysis in molecular evolutionary genetics. *Annual Review of Genetics* **30**: 371–403.
- Nei M. i Kumar S. 2000. *Molecular Evolution and Phylogenetics*. Oxford University press, Oxford, UK – New York.

- Nei M., Kumar S. i Takahashi K. 1998. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. *Proceedings of the National Academy of Sciences of the USA* **76**: 5269–5273.
- Nei M., Stephens J.C. i Saitou N. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. *Molecular Biology and Evolution* **2**: 66–85.
- Nei M., Takezaki N. i Sitnikova T. 1995. Assessing molecular phylogenies. *Science* **267**: 253–255.
- Nelson G.J. 1979. Cladistic analysis and synthesis: principles and definitions, with a historical note on Adanson's *Famille des Plantes* (1763–1764). *Systematic Zoology* **28**: 1–21.
- Nelson G.J. 1992. Reply to Harvey. *Cladistics* **8**: 355–360.
- Nelson G.J. 1993. Reply. *Cladistics* **9**: 261–265.
- Nelson G.J. 1996. Nullius in verba. *Journal of Comparative Biology* **1**: 141–152.
- Nelson G.J. i Ladiges P.Y. 1992. Information content and fractional weight of three-taxon statements. *Systematic Biology* **41**: 490–494.
- Nelson G.J. i Ladiges P.Y. 1993. Missing data and three-item analysis. *Cladistics* **9**: 111–113.
- Nelson G.J. i Ladiges P.Y. 1994. Three-item consensus: empirical test of fractional weighting. W: Scotland R.W., Siebert D.J. i Williams D.M. (red.), *Models in phylogeny reconstruction. Systematics Association Special Volume No. 52*. Clarendon Press, Oxford: 193–207.
- Nelson G.J. i Platnick N.I. 1991. Three-taxon statements: a more precise use of parsimony? *Cladistics* **7**: 351–366.
- Nixon K.C. i Carpenter J.M. 1996. On simultaneous analysis. *Cladistics* **9**: 413–426.
- Nixon K.C. i Davis J.I. 1991. Polymorphic taxa, missing values and cladistic analysis. *Cladistics* **7**: 233–241.
- Olsen G.J. 1988. Phylogenetic analysis using ribosomal RNA. *Methods in Enzymology* **164**: 793–838.
- Olsen G.J., Matsuda H., Hagstrom R. i Overbeek R. 1994. Fast DNAmI: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Computer Applied Biosciences* **10**: 41–48.
- Olsen G.J., Overbeek R., Larsen N., Marsh T.L., McCaughey M.J., Maciukenas M.A., Kuan W.-M., Macke T.J., Xing Y. i Woese C.R. 1992. The ribosomal database project. *Nucleic Acids Research, Supplement*: 2100–2200.
- Omland K.E. 1994. Character congruence between a molecular and a morphological phylogeny for dabbling ducks (*Anas*). *Systematic Biology* **43**: 369–386.
- Page R.D.M. 1988. Quantitative cladistic biogeography: Constructing and comparing area cladograms. *Systematic Zoology* **37**: 254–270.
- Page R.D.M. 1989. Comments on component-compatibility in historical biogeography. *Cladistics* **5**: 167–182.
- Page R.D.M. 1993. *COMPONENT version 2.0. MS-DOS program for Windows®*. The Natural History Museum, London.
- Page R.D.M. i Charleston M.A. 1997. Reconciled trees and incongruent gene and species trees. W: Mirkin B., McMorris F.R., Roberts F.S. i Rzhetsky A. (red.), *Mathematical Hierarchies in Biology*, Volume 37. American Mathematical Society, Providence, Richmond.
- Page R.D.M. i Holmes E.C. 1998. *Molecular Evolution. A Phylogenetic Approach*. Blackwell Science, Oxford, UK.
- Patterson C., Williams D.M. i Humphries C.J. 1993. Congruence between molecular and morphological phylogenies. *Annual Review of Ecology and Systematics* **24**: 153–188.
- Paterson A.M. i Gray R.D. 1997. Host-parasite cospeciation, host switching, and missing the boat. W: Clayton D.H. i Moore J. (red.), *Host-Parasite Evolution: General Principles and Avian Models*. Oxford University Press, Oxford, UK: 236–250.
- Pearson K. 1926. On the coefficient of racial likeness. *Biometrika* **18**: 105–117.
- Pearson W.R., Robins G. i Zhang T. 1999. Generalized neighbor-joining: More reliable phylogenetic tree reconstruction. *Molecular Biology and Evolution* **16**: 806–816.
- Penny D. i Hendy M.D. 1985a. Testing methods of evolutionary tree construction. *Cladistics* **1**: 266–272.
- Penny D. i Hendy M.D. 1985b. The use of tree comparison metrics. *Systematic Zoology* **34**: 75–82.

- Penny D. i Hendy M.D. 1986. Estimating the reliability of evolutionary trees. *Molecular Biology and Evolution* **3**: 403–417.
- Penny D., Watson E.E., Hickson R.E. i Lockhart P.J. 1993. Some recent progress with methods for evolutionary trees. *New Zealand Journal of Botany* **31**: 275–288.
- Pettigrew J.D. 1994. Flying DNA. *Current Biology* **4**: 277–280.
- Pimentel R.A. i Riggins R. 1987. The nature of cladistic data. *Cladistics* **3**: 275–289.
- Pinna M.C.C., de 1996. Comparative biology and systematics: some controversies in retrospective. *Journal of Comparative Biology* **1**: 3–16.
- Platnick N.I. 1993. Character optimization and weighting: differences between the standard and three-taxon approaches to phylogenetic inference. *Cladistics* **9**: 267–272.
- Platnick N.I., Griswold C.E. i Coddington J.A. 1991. On missing entries in cladistic analysis. *Cladistics* **7**: 337–343.
- Platnick N.I., Humphries C.J., Nelson G.J. i Williams D.M. 1996. Is Farris optimization perfect? Three-taxon statements and multiple branching. *Cladistics* **11**: 309–315.
- Pociecha J., Podolec B., Sokołowski A. i Zajac K. 1988. *Metody taksonomiczne w badaniach społeczno-ekonomicznych*. PWN, Warszawa.
- Posada D. i Crandall K.A. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.
- Prager E.M. i Wilson A.C. 1988. Ancient origin of lactalbumin from lysozyme: Analysis of DNA and amino acid sequences. *Journal of Molecular Evolution* **27**: 326–335.
- Prim R.C. 1957. Shortest connection networks and some generalizations. *Bell System Technical Journal* **36**: 1389–1401.
- Purvis A. 1995. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* **44**: 251–255.
- Ragan M.A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* **1**: 53–58.
- Rannala B. i Yang Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* **43**: 304–311.
- Rao C.R. 1948. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society B*, **10**: 159–193.
- Rao C.R. 1994. *Statystyka i prawda*. Wydawnictwo Naukowe PWN, Warszawa.
- Reeves J.H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *Journal of Molecular Evolution* **35**: 17–31.
- Remane A. 1952. *Die Grundlagen der natürlichen Systems, der vergleichenden Anatomie, und der Phylogenetik*. Akademische Verlag, Leipzig.
- Reyment R.A. 1991. *Multidimensional Paleobiology*. Pergamon Press, New York.
- Reynolds J., Weir B.E. i Cockerham C. 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* **105**: 767–779.
- Richardson B.J., Baverstock P.R. i Adams M. 1986. *Allozyme Electrophoresis: a handbook for animal systematics and population studies*. Academic Press, Sydney.
- Riedl R.J. 1978. *Order in Living Organisms: a System Analysis of Evolution*. Wiley and Sons, New York.
- Rodrigo A.G. 1993. Calibrating the bootstrap test of monophyly. *International Journal of Parasitology* **23**: 507–514.
- Rodriguez F., Oliver J.L., Marin A. i Medina J.R. 1990. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology* **142**: 485–501.
- Rogers D.J. i Tanimoto T.T. 1960. A computer program for classifying plants. *Science* **132**: 1115–1118.
- Rogers J.S. 1972. Measures of genetic similarity and genetic distance. *Studies in Genetics VII. University of Texas Publications* **7213**: 145–153.
- Rogers J.S. 1984. Deriving phylogenetic trees from allele frequencies. *Systematic Zoology* **33**: 52–63.
- Rogers J.S. i Swofford D.L. 1999. Multiple local maxima for likelihoods of phylogenetic trees: A simulation study. *Molecular Biology and Evolution* **16**: 1079–1085.

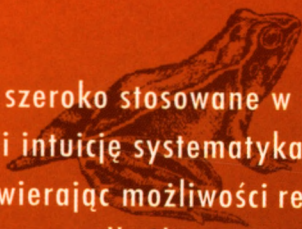
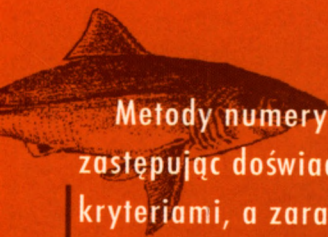
- Rohlf F.J. 1994. *NTSYS-pc. Numerical Taxonomy and Multivariate Analysis System. Version 1.80*. Exeter Software, Seatuket, New York.
- Rohlf F.J. 1998. *NTSYS-pc. Numerical Taxonomy and Multivariate Analysis System. Version 2.0*. Exeter Software, Seatuket, New York.
- Rohlf F.J. i Fisher D.L. 1968. Test for hierarchical structure in random data sets. *Systematic Zoology* 17: 407–412.
- Rohlf F.J. i Sokal R.R. 1981. Comparing numerical taxonomic studies. *Systematic Zoology* 30: 459–490.
- Ronquist F. i Nylin S. 1990. Process and pattern in the evolution of species associations. *Systematic Zoology* 39: 323–344.
- Russo C.A.M., Takezaki N. i Nei M. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Molecular Biology and Evolution* 13: 525–536.
- Rzhetsky A. i Nei M. 1992a. A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution* 9: 945–967.
- Rzhetsky A. i Nei M. 1992b. Statistical properties of the ordinary least-squares, generalized least-squares and minimum-evolution methods of phylogenetic inference. *Journal of Molecular Evolution* 35: 367–375.
- Rzhetsky A. i Nei M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution* 10: 1073–1095.
- Rzhetsky A. i Nei M. 1995. Tests of applicability of several substitution models for DNA sequence data. *Molecular Biology and Evolution* 12: 131–151.
- Rzhetsky A. i Sitnikova T. 1996. When is it safe to use an oversimplified substitution model in tree-making? *Molecular Biology and Evolution* 13: 1255–1265.
- Saitou N. 1988. Property and efficiency of the maximum likelihood method for molecular phylogeny. *Journal of Molecular Evolution* 27: 261–273.
- Saitou N. i Imanishi M. 1989. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic reconstructions in obtaining the correct tree. *Molecular Biology and Evolution* 6: 514–525.
- Saitou N. i Nei M. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. *Journal of Molecular Evolution* 24: 189–204.
- Saitou N. i Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 14: 406–425.
- Salisbury B.A. 1999. Strongest Evidence: Maximum Apparent Phylogenetic Signal as a New Cladistic Optimality Criterion. *Cladistics* 15: 137–149.
- Salisbury B.A. 2000. Strongest Evidence Revisited. *Cladistics* 16: 394–402.
- Sanderson M.J., Purvis A. i Henze C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13: 105–109.
- Sankoff D. 1990. Designer invariants for large phylogenies. *Molecular Biology and Evolution* 7: 255–269.
- Sankoff D. i Cedergren R.J. 1983. Simultaneous comparison of three or more sequences related by a tree. W: Sankoff D. i Kruskal J.B. (red.), *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Addison-Wesley, Reading, Massachusetts: 253–263.
- Sankoff D. i Rousseau P. 1975. Locating the vertices of a Steiner tree in arbitrary space. *Mathematical Programming* 9: 240–246.
- Sattath S. i Tversky A. 1977. Additive similarity trees. *Psychometrika* 42: 319–345.
- Schuh R.T. i Farris J.S. 1981. Methods for investigating taxonomic congruence and their application to the Lep-topodomorpha. *Systematic Zoology* 30: 331–351.
- Schuh R.T. i Polhemus J.T. 1981. Analysis of taxonomic congruence among morphological, ecological, and biogeographic data sets for the Lep-topodomorpha (Hemiptera). *Systematic Zoology* 29: 1–26.
- Sharkey M.J. 1989. A hypothesis-independent method of character weighting for cladistic analysis. *Cladistics* 5: 63–86.
- Sharkey M.J. 1993. Exact indices, criteria to select from minimum length trees. *Cladistics* 9: 211–222.

- Shepard R.N. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. I i II. *Psychometrika* **27**: 125–140 i 219–246.
- Shepard R.N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology* **3**: 287–315.
- Shepard R.N. 1980. Multidimensional Scaling, Tree-Fitting, and Clustering. *Science* **210**: 390–398.
- Siddall M.E. 1996a. Another monophyly index: revisiting the jackknife. *Cladistics* **11**: 33–56.
- Siddall M.E. 1996b. *Random Cladistics*, Version 4.0.3, Ohio edition. *MS-DOS program*. Virginia Institute of Marine Sciences, Gloucester Point.
- Sidow A. i Wilson A.C. 1990. Compositional statistics: An improvement of evolutionary parsimony and its application to deep branches in the tree of life. *Journal of Molecular Evolution* **31**: 51–68.
- Sidow A. i Wilson A.C. 1992. Compositional statistics evaluated by computer simulations. W: Miyamoto M.M. i Cracraft J. (red.), *Phylogenetic Analysis of DNA Sequences*. Oxford University Press, Oxford, UK: 129–146.
- Silberman J.D., Clark C.G., Diamond L.S. i Sogin M.L. 1999. Phylogeny of the Genera *Entamoeba* and *Endolimax* as deduced from small-subunit ribosomal RNA sequences. *Molecular Biology and Evolution* **16**: 1740–1751.
- Simberloff D., Heck K.L., McCoy E.D. i Connor E.F. 1981. There have been no statistical tests of cladistic biogeographical hypotheses. W: Nelson G. i Rosen D.E. (red.), *Vicariance Biogeography: A Critique*. Columbia University Press, New York: 40–63.
- Simon C., Frati F., Beckenbach A., Crespi B., Liu H. i Flook P. 1994. Evolution, weighting and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of Entomological Society of America* **87**: 651–701.
- Simpson G.G. 1961. *Principles of Animal Taxonomy*. Columbia University Press, New York.
- Singer M.F. 1982. SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. *Cell* **28**: 433–434.
- Sitnikova T. 1996. Bootstrap method of inferior-branch test for phylogenetic trees. *Molecular Biology and Evolution* **13**: 605–611.
- Sitnikova T., Rzhetsky A. i Nei M. 1995. Interior-branch and bootstrap tests of phylogenetic trees. *Molecular Biology and Evolution* **12**: 319–333.
- Slatkin M. i Maddison W.P. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**: 603–613.
- Sneath P.H.A. i Sokal R.R. 1973. *Numerical taxonomy. The principles and practice of numerical classification*. W.H. Freeman and Company, San Francisco.
- Sneath P.H.A., Sackin M.J. i Amber R.P. 1975. Detecting evolutionary incompatibilities from protein sequences. *Systematic Zoology* **24**: 311–332.
- Sober E. 1988. *Reconstructing the Past: Parsimony, Evolution, and Inference*. MIT Press, Cambridge, MA.
- Sokal R.R. i Michener C.D. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* **38**: 1409–1438.
- Sokal R.R. i Sneath P.H.A. 1963. *Principles of Numerical Taxonomy*. W.H. Freeman and Company, San Francisco.
- Sokal R.R. i Rohlf F.J. 1962. The comparison of dendrograms by objective methods. *Taxon* **11**: 33–40.
- Sokal R.R. i Rohlf F.J. 1981. Taxonomic congruence in the Leptopodomorpha re-examined. *Systematic Zoology* **30**: 309–325.
- Sokal R.R. i Rohlf F.J. 1987. *Introduction to Biostatistics*. Second edition. W.H. Freeman and Company, New York.
- Sokal R.R. i Rohlf F.J. 1995. *Biometry. The Principles and Practice of Statistics in Biological Research. Third edition*. W.H. Freeman and Company, New York.
- Stanley S.M. 1998. *Macroevolution. Pattern and Process*. The Johns Hopkins University Press, Baltimore–London, UK.
- Steel M.A., Hendy M.D. i Penny D. 1993. Parsimony can be consistent! *Systematic Biology* **42**: 581–587.
- Steel M. 1994a. Recovering a tree from the Markov leaf colourations it generates under a Markov model. *Applied Mathematics Letters* **7**: 19–23.

- Steel M. 1994b. The maximum likelihood point for a phylogenetic tree is not unique. *Systematic Biology* 43: 560–564.
- Stinebrickner R. 1984. Consensus trees and indices. *Bulletin of Mathematical Biology* 46: 923–935.
- Studier J.A. i Keppeler K.J. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution* 5: 729–731.
- Sullivan J. 1996. Combining data with different distributions of among-site variation. *Systematic Biology* 45: 375–380.
- Sullivan J., Holsinger K.E. i Simon C. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA in Sigmodontine rodents. *Molecular Biology and Evolution* 12: 988–1001.
- Swofford D.L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? W: Miyamoto M.M. i Cracraft J. (red.), *Phylogenetic analysis of DNA sequences*. Oxford University Press, Oxford: 295–333.
- Swofford D.L. 1998. *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)*. Version 4.0. Sinauer Associates, Sunderland, Massachusetts.
- Swofford D.L. i Begle D.P. 1993. *PAUP: Phylogenetic analysis using parsimony, ver. 3.1 user's manual*. Illinois Natural History Survey, Champaign, IL.
- Swofford D.L. i Berlocher S.H. 1987. Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. *Systematic Zoology* 36: 293–325.
- Swofford D.L. i Maddison W.P. 1987. Reconstructing ancestral character states under Wagner parsimony. *Mathematical Bioscience* 87: 199–229.
- Swofford D.L. i Olsen G.J. 1990. Phylogeny Reconstruction. W: Hillis D.M. i Moritz C. (red.), *Molecular Systematics*. Sinauer Associates, Inc., Sunderland, Massachusetts: 411–501.
- Swofford D.L., Olsen G.J., Waddell P.J. i Hillis D.M. 1996. Phylogenetic Inference. W: Hillis D.M., Moritz C. i Mable B.K. (red.), *Molecular Systematics. Second Edition*. Sinauer Associates, Inc., Sunderland, Massachusetts: 407–514.
- Tajima F. i Nei M. 1982. Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. *Journal of Molecular Evolution* 18: 115–120.
- Takahashi K. i Nei M. 2000. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution and maximum likelihood when a large number of sequences are used. *Molecular Biology and Evolution* 17:
- Takane Y.F., Young W. i DeLeeuw J. 1977. Non-metric Individual Differences Multidimensional Scaling: Alternating Least Squares with Optimal Scaling Features. *Psychometrika* 42: 7–67.
- Takezaki N. 1998. Tie trees generated by distance methods of phylogenetic reconstruction. *Molecular Biology and Evolution* 15: 727–737.
- Takezaki N. i Nei M. 1994. Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. *Journal of Molecular Evolution* 39: 210–218.
- Tamura K. 1992. The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. *Molecular Biology and Evolution* 9: 814–825.
- Tamura K. i Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* 10: 512–526.
- Tateno Y., Takezaki N. i Nei M. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Molecular Biology and Evolution* 11: 261–277.
- Templeton A.R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the humans and apes. *Evolution* 37: 221–244.
- Templeton A.R. 1989. The Meaning of Species and Speciation: A Genetic Perspective. W: Otte D. i Endler J.A. (red.), *Speciation and Its Consequences*. Sinauer Associates, Inc., Sunderland, Massachusetts: 3–27.
- Thiele K. 1993. The holy grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics* 9: 275–304.
- Tillier E.R.M. 1994. Maximum likelihood with multiparameter models of substitution. *Journal of Molecular Evolution* 39: 409–417.

- Tukey J.W. 1958. Bias and confidence in not quite large samples (Abstract). *Annals of Mathematical Statistics* **29**: 614.
- Thorpe R.S. 1984. Coding morphometric characters for constructing distance Wagner networks. *Evolution* **38**: 244–255.
- Verneau O., Catzeflis F. i Furano A.V. 1997. Determination of the evolutionary relationships i *Rattus sensu lato* (Rodentia: Muridae) using L1 (LINE-1) amplification events. *Journal of Molecular Evolution* **45**: 424–436.
- Waddell P.J. 1995. *Statistical methods of phylogenetic analysis, including Hadamard conjugations, LogDet transforms, and maximum likelihood*. Ph. D. dissertation, Massey University; cytowane za Swofford i inni (1996).
- Waddell P.J. i Penny D. 1996. Evolutionary trees of apes and humans from DNA sequences. W: Lock A.J. i Peters C.R. (red.), *Handbook of Symbolic Evolution*. Clarendon Press, Oxford.
- Waddell P.J., Penny D., Hendy M.D. i Arnold G. 1994. The sampling distributions and covariance matrix of phylogenetic spectra. *Molecular Biology and Evolution* **11**: 630–642.
- Wagner G.P. 1989. The origin of morphological characters and the biological basis of homology. *Evolution* **43**: 1157–1171.
- Wagner W.H. 1961. Problems in the classification of ferns. *Recent Advances in Botany* **1**: 841–844.
- Watrous L.E. i Wheeler Q.D. 1981. The outgroup comparison method of character analysis. *Systematic Zoology* **30**: 1–11.
- Weir B.S. 1990. *Genetic Data Analysis. Methods for Discrete Population Genetic Data*. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts.
- Wheeler W.C. 1990. Combinatorial weights in phylogenetic analysis: A statistical parsimony procedure. *Cladistics* **6**: 269–275.
- Wheeler W.C., Cartwright P. i Hayashi C.Y. 1993. Arthropod phylogeny: a combined approach. *Cladistics* **9**: 1–39.
- Whelan S. i Goldman N. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Molecular Biology and Evolution* **16**: 1292–1299.
- Wiens J.J. i Reeder T.W. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. *Systematic Biology* **44**: 548–558.
- Wiley E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. Wiley Interscience, New York.
- Wiley E.O. 1988. Parsimony analysis and vicariance biogeography. *Systematic Zoology* **37**: 271–290.
- Wilke T., Davis G.M., Falniowski A., Giusti F., Bodon M. i Szarowska M. 2001. Molecular systematics of Hydrobiidae (Mollusca: Gastropoda: Rissooidea): testing monophyly and phylogenetic relationships. *Proceedings of the Academy of Natural Sciences of Philadelphia* **151**: 1–21.
- Wilks S.S. 1932. Certain Generalizations in the Analysis of Variance. *Biometrika* **24**: 471–494.
- Williams P.L. i Fitch W.M. 1989. Finding the minimum change in a given tree. W: Fernholm B., Bremer K. i Jönrvall H. (red.), *The Hierarchy of Life*. Elsevier Press, Amsterdam: 453–470.
- Williams P.L. i Fitch W.M. 1990. Phylogeny determination using dynamically weighted parsimony method. W: Doolittle R.F. (red.), *Methods in enzymology*. Academic Press, San Diego, California: 615–626.
- Wolstenholme D.R. 1992. Animal mitochondrial DNA: Structure and evolution. W: Wolstenholme D.R. i Jeon K.W. (red.), *International review of cytology: Mitochondrial genomes*. Academic Press, San Diego, California: 173–216.
- Wright S. 1978. *Evolution and the genetics of populations*. Vol. 4. *Variability within and among natural populations*. University Chicago Press, Chicago.
- Yang Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10**: 1396–1402.
- Yang Z. 1994a. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**: 105–111.
- Yang Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* **39**: 306–314.
- Yang Z. 1994c. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic Biology* **43**: 329–342.

- Yang Z. 1996. Among-site variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution* **11**: 367–371.
- Yang Z. 1997. How often do wrong models produce better phylogenies? *Molecular Biology and Evolution* **14**: 105–108.
- Yang Z. 1999. *PAML: Phylogenetic analysis by maximum likelihood*, Version 2.0. University College London, London.
- Yang Z., Goldman N. i Friday A. 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Molecular Biology and Evolution* **11**: 316–324.
- Yang Z., Goldman N. i Friday A. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Systematic Biology* **44**: 384–399.
- Yang Z. i Kumar S. 1996. Approximate methods for estimating the pattern of nucleotide substitution rates among sites. *Molecular Biology and Evolution* **13**: 650–659.
- Yang Z. i Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo Method. *Molecular Biology and Evolution* **14**: 717–724.
- Zadeh L.A. 1975. Fuzzy Sets – Notation, Terminology and Basic Properties. W: Zadeh L.A., Fu K.S., Tanaka K. i Shimura M. (red.), *Fuzzy Sets and their Applications to Cognitive and Decision Processes*. Academic Press, New York: 27–39.
- Zardoya R., Economidis P.S. i Doadrio I. 1999. Phylogenetic relationships of Greek Cyprinidae: Molecular evidence for at least two origins of the Greek Cyprinid fauna. *Molecular Phylogeny and Evolution* **13**: 122–131.
- Zhang J. 1999. Performances of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Molecular Biology and Evolution* **16**: 868–875.
- Zharkikh A. i Li W.-H. 1992a. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molecular Biology and Evolution* **9**: 1119–1147.
- Zharkikh A. i Li W.-H. 1992b. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. *Journal of Molecular Evolution* **35**: 356–366.
- Zharkikh A. i Li W.-H. 1993. Inconsistency of the maximum-parsimony method: The case of five taxa with a molecular clock. *Systematic Biology* **42**: 113–125.
- Zharkikh A. i Li W.-H. 1995. Estimation of confidence in phylogeny: The full-and-partial bootstrap technique. *Molecular Phylogenetics and Evolution* **4**: 44–63.



Metody numeryczne są szeroko stosowane w taksonomii biologicznej, zastępując doświadczenie i intuicję systematyka bardziej obiektywnymi kryteriami, a zarazem otwierając możliwości rekonstruowania procesów makroewolucji. W wielu przypadkach są pomocne, w innych — przede wszystkim dla danych molekularnych — niezbędne. Poza taksonomią są użyteczne w biologii porównawczej i ewolucyjnej. Współczesny taksonom czy biolog ewolucyjny powinien znać przynajmniej ich podstawy, które przedstawiono w tej książce. Choć szereg technik nadaje się do każdego rodzaju danych, książka koncentruje się na analizie danych molekularnych, bowiem ich analiza rozwija się najszybciej, a wiele technik tylko tu znajduje zastosowanie. Książka zaczyna się od danych, zaś kończy na drzewach obliczonych na podstawie danych wyjściowych. Nie zajmuje się ani teorią klasyfikacji, ani interpretacją uzyskanych drzew. Przedstawiono też podstawowe techniki fenetyczne, doskonale nadające się do wstępnej analizy danych — czyli właśnie badania różnorodności. Stronę matematyczną ograniczono do minimum, kładąc nacisk na intuicyjne zrozumienie przedstawianych technik.

www.wuj.pl

ISBN 83-233-1745-3



9 788323 317456