STRESZCZENIE

Rozkład potęgowy naturalnych sekwencji nukleotydowych o różnych funkcjach jako cecha niektórych kodów zapisu informacji genetycznej

Rozkład potęgowy znajduje szerokie zastosowanie w opisie zarówno zjawisk naturalnych, jak i społecznych. W lingwistyce szczególnym przypadkiem tego rozkładu jest prawo Zipfa, które opisuje charakterystyczną właściwość tekstów języków naturalnych. Zgodnie z tą regułą, jeśli słowa w tekście uszeregować według częstości występowania – począwszy od najczęstszych (ranga = 1) do najrzadszych (o najwyższej randze) – to wykres zależności częstości od rangi, przedstawiony na podwójnie logarytmicznej skali, przyjmuje formę linii prostej o nachyleniu bliskim -1. Spełnienie przez tekst prawa Zipfa traktowane jest zarówno jako wyznacznik, jak i cecha rozpoznawcza struktur językowych.

Od momentu odkrycia podwójnej helisy DNA analogie między zapisem językowym a sekwencją nukleotydową genomu wydawały się naturalne i intuicyjne dla naukowców. W obu przypadkach informacja jest reprezentowana w jednowymiarowym, liniowym zapisie – w formie tekstu pisanego lub sekwencji genomu – który składa się z ciągu znaków lub symboli (liter alfabetu w języku naturalnym oraz nukleotydów w zapisie genetycznym) ułożonych w określonej kolejności. Podobnie jak w przypadku języka, odczyt tej informacji odbywa się poprzez łączenie poszczególnych symboli w większe jednostki znaczeniowe, takie jak słowa w tekstach językowych czy kodony w sekwencjach DNA. Podobieństwa między zapisem językowym a sekwencjami DNA zainspirowały naukowców do zastosowania narzędzi lingwistycznych w analizie genomowej. Pierwsze próby tego rodzaju przyniosły jednak wyniki niejednoznaczne. Możliwości rozwoju tego podejścia ograniczały doniesienia, że prawo Zipfa spełniają wyłącznie sekwencje niekodujące. Sceptycyzm pogłębiło odkrycie, że również losowe teksty mogą podlegać prawu potęgowemu. W kolejnych latach pojawiały się sprostowania sugerujące, że wcześniejsze wnioski były przedwczesne i opierały się na niewystarczająco precyzyjnych analizach. Mimo to entuzjazm wobec implementacji podejścia lingwistycznego w genomice osłabł, a dyskusje na temat użyteczności prawa potęgowego w badaniach genomowych pozostają nierozstrzygnięte do dnia dzisiejszego.

Celem niniejszej analizy nie jest ocena wcześniejszych badań ani ważności wysuwanej wobec nich krytyki, lecz spojrzenie na problem z odmiennej perspektywy.

Niektóre uwagi krytyczne dotyczące zastosowania narzędzi lingwistycznych w genomice mogą wynikać nie tyle z ograniczeń samej metody, ile z błędnych założeń przyjętych przez wcześniejszych badaczy. Warto zauważyć, że pionierzy lingwistyki genetycznej opierali swoje badania na dość intuicyjnych założeniach. Sekwencja genomowa zapisana jest za pomocą 4-literowego alfabetu {A, C, T, G}, odpowiadającego czterem nukleotydom: adeninie, cytozynie, tyminie i guaninie. Najmniejszymi znaczącymi blokami genomów są trójki nukleotydów, zwane kodonami, a zapis informacji genetycznej jest bezprzecinkowy, to znaczy pozbawiony znaków specjalnych. Rozprawa proponuje przebadanie alternatywnego scenariusza. Alfabet genetyczny zawiera 7 symboli; znaczące fragmenty genomów (słowa) nie muszą być trzyliterowe, ale mogą mieć różną wielkość; w zależności od kontekstu ten sam nukleotyd może pełnić odmienne role, a w zapisie informacji genetycznej istnieją znaki wyróżniające. Pomysł na wprowadzenie 7-symbolowego alfabetu ma uzasadnienie biologiczne i wynika z degeneracji kodu genetycznego. Wiemy, że w zależności od „kontekstu", to znaczy od układu dwóch wcześniejszych pozycji w kodonie, ostatnia pozycja może być chwiejna. Dokładnie pozycję tę może zająć dowolna puryna, dowolna pirymidyna, a w niektórych przypadkach odczyt informacji genetycznej nie zmienia się niezależnie od tego, jaki nukleotyd zajmuje to miejsce. W związku z powyższym zaproponowano rozszerzoną wersję alfabetu genomowego, składającą się z czterech nukleotydów oraz trzech dodatkowych symboli: „dowolnej puryny", „dowolnej pirymidyny" i „dowolnego nukleotydu". Transformacja sekwencji za pomocą tego 7-symbolowego alfabetu (tak zwana analiza kontekstowa) będzie polegała na odpowiedniej zamianie co trzeciego nukleotydu na jeden z nowych zaproponowanych znaków. Dodatkowo przyjęto, że symbol „dowolny nukleotyd" pełni w tekście funkcję spacji, co dostarcza naturalnego sposobu tokenizacji (czyli wydzielania tokenów/słów z ciągu liter). Sekwencje znajdujące się pomiędzy spacjami tworzą w tekście genomowym słowa o różnej długości.

Celem pracy jest przetestowanie tego alternatywnego sposobu zapisu informacji genetycznej. Sformułowano następującą hipotezę badawczą: jeśli zaproponowany sposób zapisu informacji jest poprawny, wówczas teksty genomowe poddane analizie kontekstowej będą, podobnie jak teksty języków naturalnych, spełniały prawo Zipfa.

Badania przeprowadzono na sekwencjach różnych elementów strukturalnych genomu (3'UTR, 5'UTR, CDS, eksony, geny, introny, transkrypty i promotory) pochodzących z 60 organizmów reprezentujących wszystkie domeny życia. Przy doborze organizmów starano się zapewnić maksymalną możliwą różnorodność filogenetyczną.

Jako kontrolę pozytywną dla analizy kontekstowej zastosowano tokenizację metodą przesuwającej się ramki. Metoda ta jest uznawana w językoznawstwie za uniwersalne narzędzie, które znajduje zastosowanie w analizie tekstów językowych o nieznanych zasadach tokenizacji. Natomiast jako kontrolę negatywną zastosowano tradycyjną tokenizację trypletową, nawiązującą do zjawiska translacji i polegającą na podziale sekwencji na nienachodzące na siebie trójnukleotydowe słowa zapisane w standardowym czteronukleotydowym alfabecie. Przeanalizowano różne warianty eksperymentu, obejmujące standardową analizę nukleotydową, wykorzystującą czteroliterowy alfabet, oraz analizę kontekstową, opartą na 7-symbolowym alfabecie, a także różne metody tasowania tekstów – zarówno na poziomie nukleotydowym, jak i kontekstowym. Te same eksperymenty powtórzono dla tekstów losowych, wygenerowanych w sposób symulujący sekwencje nukleotydowe. Pseudo-genomy o identycznej długości i liczbie zostały poddane analizie statystycznej, analogicznej do tej stosowanej w przypadku tekstów genomowych.

Statystyczna analiza weryfikująca hipotezę badawczą została przeprowadzona w trzech etapach. W pierwszym etapie wykonano analizę wizualną, w której prawo Zipfa przedstawiono na trzy różne sposoby: za pomocą wykresów absolutnej częstości występowania słów, względnej zależności częstości słów od rangi oraz ich komplementarnej dystrybuanty. W przypadku dystrybuanty komplementarnej na dane empiryczne nałożono funkcje dopasowania odpowiadające teoretycznym rozkładom potęgowemu i logarytmiczno-normalnemu. W drugim etapie przeprowadzono dopasowanie rozkładu przy użyciu rygorystycznych metod statystycznych oraz oceniono to dopasowanie za pomocą testu Kołmogorowa-Smirnowa. Ze względu na trudności związane ze statystyczną oceną rozkładu potęgowego w trzecim etapie dodatkowo porównano badaną dystrybucję do alternatywnych rozkładów o najbardziej zbliżonym profilu, takich jak rozkład wykładniczy, logarytmiczno-normalny oraz potęgowy z obcięciem ciężkiego ogona.

Przeprowadzone badania wykazały, że naturalne fragmenty DNA poddane analizie kontekstowej wykazują wzorce zgodne z rozkładem Zipfa. Wyniki te zostały potwierdzone zarówno za pomocą metody tokenizacji kontekstowej, jak i kontroli z zastosowaniem tokenizacji ramkowej. Teksty genomowe zapisane w 7-symbolowym alfabecie charakteryzują się największym podobieństwem do rozkładu potęgowego, a parametry dopasowania częstości występowania słów są tu zbliżone do wartości

typowych dla prawa Zipfa w tekstach języków naturalnych. Jedynie niewielki odsetek tych tekstów wykazuje lepsze dopasowanie do alternatywnych modeli rozkładu. Uzyskane wyniki wspierają hipotezę o istnieniu alternatywnego sposobu zapisu informacji genetycznej.

W przeciwieństwie do wyników uzyskanych w analizie kontekstowej rozkłady częstości względem rangi w tekstach poddanych tokenizacji trypletowej wyraźnie nie odpowiadają prawu potęgowemu. Brak liniowości na wykresach, w połączeniu z niskimi ocenami statystycznymi, wskazuje, że tradycyjnie stosowany trypletowy sposób zapisu nie stanowi odpowiedniego modelu do reprezentacji informacji semiotycznej w genomie.

Tasowanie sekwencji genomowych na poziomie nukleotydów zakłóca informację zawartą w genomie, co objawia się pogorszeniem jakości dopasowania do prawa potęgowego. Z kolei tasowanie sekwencji genomowych na poziomie kontekstowym (po przekształceniu na 7-symbolowy alfabet) poprawia jakość tego dopasowania. Wynik zdaje się wskazywać na istniejące różnice między tekstem języka naturalnego a tekstem genomowym. Zapis informacji genomowej zdaje się być mniej zoptymalizowany i mniej równomierny w porównaniu z tekstami języka naturalnego, co wiąże się prawdopodobnie z odmiennym procesem ewolucji obu zapisów. W przypadku genomów jest to proces wolniejszy i bardziej kosztowny pod względem optymalizacji.

Analiza kontekstowa tekstów reprezentujących różne elementy strukturalne genomu wykazała, że rozkład częstości słów zarówno w sekwencjach kodujących, jak i niekodujących podlega prawu Zipfa. Jednocześnie charakterystyki tych rozkładów różnią się między sobą. Narzędzia lingwistyczne zastosowane do analizy statystycznej genomu umożliwiają rozróżnienie tekstów pochodzących z różnych elementów funkcjonalnych genomu. Podobnych wniosków nie można wyciągnąć w odniesieniu do pochodzenia ewolucyjnego tekstów. Czynnik ten wydaje się nie mieć wpływu na rozkład częstości słów w tekstach lub nie stanowi wystarczająco istotnego elementu w analizie statystycznej.

Przeprowadzona analiza wykazała istotne różnice między rozkładem częstość-ranga w losowych tekstach pseudo-genomów a dystrybucją tekstów genomowych. Rozkład tekstów losowych ma inną charakterystykę, jest bardziej dyskretny i jednorodny. Parametry opisujące funkcję dopasowania wyraźnie różnią się między tekstami genomowymi a losowymi, a ogólna jakość dopasowania rozkładu potęgowego do danych jest znacznie lepsza dla tekstów genomowych w porównaniu z tekstami losowymi. Analiza statystyczna wykazuje ponadto, że teksty losowe częściej preferują alternatywne

rozkłady w stosunku do dystrybucji potęgowej. Tasowanie tekstów losowych na poziomie nukleotydów (w odróżnieniu od genomów) nie zmienia ich rozkładu, co jest zgodne z bronioną hipotezą. Skoro teksty losowe nie przenoszą sensownej informacji genetycznej, tasowanie nukleotydów nie wpływa na ich strukturę. Tasowanie na poziomie kontekstowym tekstów losowych najbardziej upodabnia je do tekstów genomowych, jednak nawet w tym przypadku istnieją wyraźne różnice w jakości dopasowania obu typów tekstów. Analiza tekstów losowych wspiera tezę, że struktura informacji wykrywana przez analizę kontekstową w tekstach genomowych nie jest artefaktem stochastycznych właściwości prawa potęgowego ani wynikiem zastosowanych metod manipulacji tekstem, ale niesie sensowną treść, której źródłem jest ewolucja biologiczna.

W centrum niniejszej pracy znajduje się zagadnienie zapisu informacji w sekwencji genomowej. Praca ma więc charakter teoretyczny, dotykający kluczowego problemu kodowania informacji w biologii. Właśnie ze względu na ten fundamentalny charakter poruszanego zagadnienia jej wyniki mogą potencjalnie mieć dalekosiężne konsekwencje praktyczne. Alternatywny sposób kodowania informacji genetycznej może znaleźć zastosowanie w różnych dziedzinach biologii, takich jak analiza strukturalna biomolekuł, rozwój algorytmów dopasowania sekwencji, optymalizacja metod kompilacji i przechowywania danych genomowych, badania filogenetyczne czy analizy w ramach biologii systemowej. Zaprezentowana w pracy analiza kontekstowa może również stanowić nowy impuls do ponownego zbliżenia dwóch tradycji badawczych: lingwistycznej i genomowej.