Uniwersytet Jagielloński

Wydział Biochemii, Biofizyki i Biotechnologii

Rozprawa doktorska

Reprezentacja gry chaosu w badaniach sekwencji nukleotydowych i aminokwasowych

Adrian Kania

Promotor: dr hab. Krzysztof Murzyn, prof. UJ

Zakład Biofizyki Obliczeniowej i Bioinformatyki

Kraków, grudzień 2024

1:1069115522

Podziękowanie

dr hab. Krzysztof Murzyn, prof. UJ za cierpliwość oraz cenne rady które przyczyniły się do ostatecznej wersji tej pracy

2:3328989382

Spis treści

Słowa kluczowe **5**

Wykaz najważniejszych skrótów **5**

Streszczenie **6**

Abstract **7**

Rozdział 1. Wprowadzenie **8**

Rozdział 2. Hipotezy badawcze i cele pracy **11**

2.1. Hipotezy badawcze 11

2.2. Cele pracy 11

Rozdział 3. Metody i zasoby **13**

3.1. Biologiczne bazy danych dotyczące sekwencji biologicznych 13

3.2. Porównywanie sekwencji i tworzenie drzew filogenetycznych 14

3.3. Symulowanie sieci filogenetycznych 16

3.4. Porównywanie drzew filogenetycznych 16

3.5. Specyfika sekwencji nukleotydowych oraz prawa biologiczne 17

3.6. Metody opisu i reprezentacji sekwencji nukleotydowych i białkowych 18

3.6.1. Entropia i logo 18

3.6.2. K-mery i analiza częstości 20

3.6.3. Złożoność Lempel-Ziv 20

3.6.4. Krzywa Z 21

3.6.5. Reprezentacja gry chaosu 23

3.6.6. Genomacierz 28

3.6.7. Spacer DNA 31

3.6.8. PseAAC 34

3.7. Dyskretna Transformata Fouriera (DFT) 35

3.8. Wykładnik Hursta 38

3.9. Ilościowa analiza rekurencji (RQA) 41

3.10. Momenty bezwładności 41

3.11. Indeks podobieństwa strukturalnego (SSIM) 42

Rozdział 4. Wyniki i dyskusja **44**

4.1. Motywy sekwencyjne w miRNA mogą być identyfikowane na podstawie analizy CGR 45

4.2. CGR pozwala scharakreryzować cechy sekwencji genów niezbędnych u bakterii, które mogłyby być wykorzystane jako deskryptory przy ich klasyfikacji 49

4.3. CGR może być stosowana w molekularnej analizie filogenetycznej do konstrukcji drzew filogenetycznych 63

4.3.1. Analiza sekwencji referencyjnych oraz genu neuraminidazy wirusów grypy 63

3:4350176605

4.3.2. Rekonstrukcja relacji ewolucyjnych wśród wybranych genomów referencyjnych dostępnych przez AFproject: wskazanie ograniczeń 75

4.4. Możliwe są rozszerzenia CGR dla DNA, RNA oraz białek, zachowujące najważniejsze własności oryginalnej reprezentacji, ale wykazujące szerszy zakres potencjalnych zastosowań 77

4.4.1. Przypadek DNA 77

4.4.2. Przypadek RNA 83

4.4.3. Przypadek białek 88

4.5. Efektywność kodowania sekwencji i rekonstrukcji relacji ewolucyjnych w oparciu o CGR może być optymalizowana 93

Rozdział 5. Podsumowanie 106

Dodatki 108

Literatura 108

Spis rysunków 117

Spis tabel 123

4:3382344970

Słowa kluczowe

analiza sekwencji; mikroRNA; gen niezbędny; porównywanie sekwencji; metody bez dopasowania; reprezentacja gry chaosu; drzewo filogenetyczne;

Wykaz najważniejszych skrótów

miRNA (mikroRNA) - mała, jednoniciowa cząsteczka niekodującego RNA o długości najczęściej między 21 a 23 nukleotydów, regulująca ekspresję innych genów

AF – metody bez dopasowania (ang. *Alignment-Free Methods*)

CGR – reprezentacja gry chaosu (ang. *Chaos Game Representation*)

UPGMA – metoda średnich połączeń (ang. *Unweighted Pair Group Method With Arithmetic Mean)*

DFT - dyskretna transformata Fouriera (ang. *Discrete Fourier Transform*)

SSIM – indeks podobieństwa strukturalnego (ang. *Structural Similarity Index Measure*)

5:3112620785

Streszczenie

Analiza sekwencji nukleotydowych oraz aminokwasowych ma szerokie zastosowanie w naukach o życiu. Zwyczajowo metody analizy takich sekwencji obejmują albo algorytmy konstrukcji dopasowań albo techniki porównywania sekwencji bez dopasowania. W poniższej pracy przedstawiono różne aspekty analizy sekwencji oparciu o reprezentację gry chaosu (ang. *Chaos Game Representation*, CGR). Wskazano na jej adekwatność jako modelu opisu sekwencji biologicznych oraz potencjalnie liczne zastosowania. CGR wykorzystano tutaj m.in. do oceny składu i identyfikacji motywów występujących w ludzkich mikroRNA na podstawie danych z miRBase oraz miRTarBase. Analiza CGR dla takich sekwencji pozwoliła zidentyfikować odmienne wzorce charakteryzujące sekwencje mikroRNA w zależności od siły oddziaływań z transkryptem. Wyniki analiz porównano z podzbiorem mikroRNA potwierdzonych eksperymentalnie czy charakteryzujących się wysokim stopniem podobieństwa między gatunkami. W innej części rozprawy, CGR oraz metodę spaceru DNA wykorzystano do analizy tzw. genów niezbędnych u bakterii zebranych w DEG. Wyniki tych badań umożliwiły zaproponowanie potencjalnie użytecznych deskryptorów przy budowie przyszłych modeli klasyfikacji genów bakteryjnych (geny niezbędne vs. geny opcjonalne). Kolejne części rozprawy dotyczą oceny użyteczności CGR, wykładnika Hursta, indeksu podobieństwa strukturalnego (SSIM) oraz dyskretnej transformaty Fouriera (DFT) przy porównywaniu sekwencji genów, genomów czy białek o różnym stopniu podobieństwa. Na podstawie wyznaczonych odległości Jaccarda-Robinsona-Fouldsa można stwierdzić, że taki dobór technik analitycznych może prowadzić do odtworzenia wiarygodnych relacji ewolucyjnych między badanymi sekwencjami. W rozprawie przedyskutowano zakres ich stosowalności oraz ograniczenia. Rozprawa zawiera również opis rozszerzenia CGR, oryginalnie zaproponowane wyłącznie do sekwencji DNA, na przypadek sekwencji aminokwasowych oraz RNA wraz z formalizmem matematycznym osadzonym w konkretnym, biologicznym kontekście.

6:5142457127

Abstract

The analysis of nucleotide and amino acid sequences has broad applications in life sciences. Typically, methods for analyzing such sequences involve either alignment-based algorithms or alignment-free sequence comparison techniques. This paper presents various aspects of sequence analysis based on the Chaos Game Representation (CGR). Its suitability as a model for describing biological sequences and its numerous potential applications are highlighted. CGR was applied here, among other uses, to evaluate the composition and identify motifs present in human microRNAs based on data from miRBase and miRTarBase. The CGR analysis of such sequences enabled the identification of distinct patterns characterizing microRNA sequences depending on their binding strength with transcripts. The results of these analyses were compared with a subset of experimentally validated microRNAs or those exhibiting high interspecies similarity. In another part of the study, CGR and the DNA walk method were used to analyze so-called essential genes in bacteria collected in the DEG database. The results of this research allowed the proposal of potentially useful descriptors for building future models for classifying bacterial genes (essential vs. non-essential genes). Subsequent sections of the dissertation address the evaluation of the utility of CGR, the Hurst exponent, the Structural Similarity Index (SSIM), and the Discrete Fourier Transform (DFT) in comparing gene, genome, and protein sequences of varying similarity. Based on the calculated Jaccard-Robinson-Foulds distances, it can be concluded that such a selection of analytical techniques may lead to the reconstruction of reliable evolutionary relationships among the studied sequences. The applicability and limitations of these methods are discussed in the dissertation. The dissertation also describes the extension of CGR, originally proposed solely for DNA sequences, to the case of amino acid and RNA sequences, along with a mathematical formalism embedded in a specific biological context.

7:9494025471

Rozdział 1. Wprowadzenie

Analiza sekwencji nukleotydowych oraz aminokwasowych stanowi jeden z kluczowych etapów w badaniach biologicznych [1,2]. W szczególności, sekwencje nukleotydowe kodują informacje genetyczne, które są podstawą funkcjonowania organizmów żywych. Dzięki takim analizom możliwa jest identyfikacja nowych genów, predykcja struktury białek, badanie zmienności genetycznej w populacjach czy ewolucyjnych powiązań między różnymi gatunkami [3,4,5,6]. Ma ona istotne znaczenie w badaniach nad adaptacją, a także w ochronie różnorodności biologicznej [7,8]. W medycynie, analiza sekwencji DNA umożliwia diagnozowanie chorób genetycznych, identyfikację mutacji oraz dostosowywanie terapii do pacjentów, co jest podstawą medycyny spersonalizowanej [9,10]. Dzięki rozwojowi technologii sekwencjonowania, które znacząco obniżyły koszty i zwiększyły szybkość uzyskiwania danych, analiza sekwencji biologicznych stała się jednym z najbardziej dynamicznie rozwijających się obszarów współczesnej nauki [11].

W tego typu badaniach, na początkowych etapach pojawia się zagadnienie porównania sekwencji. Klasyczne metody opierają się na ich dopasowaniu, które polega na wyrównywaniu dwóch lub więcej sekwencji w taki sposób, aby maksymalizować ich podobieństwo [12]. Metody bez dopasowania (ang. *Alignment-Free Methods,* AF) analizują globalne lub lokalne cechy sekwencji, takie jak częstości występowania merów (czyli podsekwencji o zadanej długości) lub wyznaczane są bardziej zaawansowane wskaźniki, jak złożoność Lempel-Ziv [13,14,15].

W 1990 roku Jeffrey zaproponował tzw. reprezentację gry chaosu (ang. *Chaos Game Representation,* CGR) do opisu sekwencji genetycznych [16]. Później, wprowadzono też jej rozszerzenia na sekwencje białkowe [17,18]. Metoda ta polega na iteracyjnym umieszczaniu punktów na płaszczyźnie w oparciu o prostą transformację i kolejne reszty występujące w sekwencji. CGR oferuje wiele zalet jako metoda analizy sekwencji. Przede wszystkim umożliwia jednoznaczne i klarowne przedstawienie informacji o całej sekwencji, w tym zależnościach dalekiego zasięgu. Może pozwolić to na identyfikację charakterystycznych wzorów czy motywów niezależnie od ich lokalizacji. W praktyce, CGR znalazła zastosowanie w analizie genomów, badaniach nad metagenomiką oraz w identyfikacji gatunków na podstawie sekwencji DNA [19,20]. Jej graficzna natura sprawia, że jest szczególnie użyteczna w eksploracyjnej analizie danych oraz wizualizacji wyników analizy sekwencji, ułatwiając interpretację danych [21].

W przypadku identyfikacji wzorców, CGR wraz z innym deskryptorami zostały wykorzystane do badania genów niezbędnych u bakterii. Te stanowią najczęściej grupę konserwatywnych sekwencji i są związane z podstawowymi funkcjami czy procesami w zachodzącymi komórce, jak produkcja energii czy przepływ informacji genetycznej [22]. Eksperymentalne metody ich odkrywania obejmują wprowadzanie mutacji do określonych genów, a następnie weryfikacji czy te okazały się letalne czy losową mutagenenzę za pomocą transpozonów [23,24]. Badanie genów niezbędnych pozwala z jednej strony lepiej zrozumieć podstawy życia, a z drugiej ma wiele praktycznych zastosowań, jak możliwość projektowania efektywniejszych antybiotyków [25]. Techniki eksperymentalne są jednak często kosztowne i czasochłonne, stąd pojawia się potrzeba poszukiwania charakterystycznych cech tych sekwencji, a następnie budowa modeli komputerowych. Pokazano m.in., że wartość współczynnika Hursta jest średnio mniejsza w przypadku genów niezbędnych w porównaniu z pozostałymi [26]. Wskazuje to na słabsze korelacje dalekiego zasięgu w tej grupie sekwencji. Dodatkowo, rozkład wykładnika Hursta wśród puli genów niezbędnych wydaje się w większości przypadków wykazywać cechy rozkładu

8:2127751151

normalnego [27]. CGR była także wielokrotnie wykorzystywana do kodowania sekwencji biologicznych, co stanowi istotny krok przy zastosowaniu metod nauczania maszynowego do komputerowej analizy sekwencji czy budowania modeli [28,29]. W [30], reprezentację gry chaosu zastosowano do zakodowania sekwencji genetycznych aby następnie przewidywać odporność bakterii na antybiotyki. Zbudowano modele do prognozy odporności na cyprofloksacynę, cefotaksym, ceftazydym oraz gentamycynę. Dodatkowo, zidentyfikowano mutacje związane z odpornością na poszczególne antybiotyki. Jeszcze inne analizy obejmują badania nad rakiem płuc czy haplotypem [31,32]. W 2024, w trakcie finalizacji pisania tej pracy, CGR z sukcesem został zastosowany do przewidywania celów mikroRNA (miRNA), czyli krótkich cząsteczek RNA regulujących ekspresję genów [33].

W literaturze pojawiło się też wiele prac aplikujących reprezentację gry chaosu jako metodę porównywania sekwencji bez dopasowania. Uzyskane reprezentacje można traktować jako obrazy, a porównanie sekwencji polega w tym przypadku na porównaniu tych obrazów [34]. Inne podejścia traktują otrzymany ciąg punków jako szereg i stosują metody im dedykowane, jak wykład Hursta czy dyskretna transformata Fouriera (ang. *Discrete Fourier Transform*, DFT) [35,36]. Jednak, nowe metody najczęściej prezentowane są na wybranych przykładach. Dodatkowo, brak kompleksowego porównania ich ze sobą oraz testowania na większej liczbie zestawów sekwencji celem oceny ich skuteczności.

W tej pracy przedstawiono kilka analiz sekwencji nukleotydowych oraz aminokwasowych przeprowadzonych w oparciu o reprezentację gry chaosu i inne deskryptory. Najogólniej można je podzielić na dwie kategorie – rozpoznawanie motywów oraz porównywanie sekwencji. W pierwszym przypadku przeanalizowano sekwencje ludzkich mikroRNA, celem wytypowania charakterystycznych wzorców w nich obecnych [37]. Dane pobrano z baz danych miRBase [38,39] oraz miRTarBase [40]. Znajomość konserwatywnych motywów z jednej strony poszerza bazę wiedzy dotyczącą badań fundamentalnych, a z drugiej ma oczywiście szereg zastosowań praktycznych, chociażby przy projektowaniu nowych terapeutycznych miRNA [41]. Kolejnym zbiorem sekwencji podjętym do analizy przez autora tej pracy był zestaw genów niezbędnych u bakterii. Dane pochodzą z bazy danych DEG (ang. *Database of Essential Genes)*, a otrzymano je metodami eksperymentalnymi [42]. Z powodu wysokich kosztów oraz czasochłonności takich badań, do przewidywania nowych genów niezbędnych wykorzystywane są też modele komputerowe. Te wymagają odpowiednich deskryptorów, czyli użytecznych przy klasyfikacji cech o rozważanych obiektach. W tej pracy autor z jednej strony podjął próbkę charakterystyki genów niezbędnych, a z drugiej zaproponował kilka współczynników w oparciu o proste statystyki oraz reprezentację gry chaosu wskazujące na odmienny charakter puli genów niezbędnych od pozostałych. W kolejnym rozdziale, autor ocenił efektywność metod do rekonstrukcji relacji ewolucyjnych w oparciu o reprezentację gry chaosu stosując odległość Jaccarda-Robinsona-Fouldsa [43,44]. Zastosowano dodatkowe metody, takie jak dyskretna transformata Fouriera (ang. *Discrete Fourier Transform,* DFT), wykładnik Hursta czy indeks podobieństwa strukturalnego (ang. *Structural Similarity Index Measure,* SSIM). Podejścia te były już proponowane w literaturze [34,35,36]. Autor testował zakres ich stosowalności na różnych zestawach sekwencji. Dodatkowo, zostały wprowadzone pewne modyfikacje potencjalnie poprawiające efektywność. Obliczenia przeprowadzono m.in. dla genu neuraminidazy oraz genomów bakteryjnych. Istotną częścią tej pracy było także badanie własności reprezentacji gry chaosu przy kodowaniu sekwencji biologicznych, a także potencjalnych jej rozszerzeń i wynikających z tego konsekwencji biologicznych. Z jednej strony taka transformacja może być użyteczna przy budowie późniejszych

9:2583669380

modeli, a z drugiej, co zdaniem autora jest nawet ważniejsze, może to być bardziej adekwatny sposób opisu sekwencji biologicznych; alternatywny wobec standardowego zapisu za pomocą liter. Obliczenia wykonano przede wszystkim stosując języki programowania Python 3.8.5 oraz R 4.1.3. Znacząca część wyników tej pracy została już opublikowana i nierzadko stanowią one inspiracje do dalszych badań.

10:1143961871

Rozdział 2. Hipotezy badawcze i cele pracy

2.1. Hipotezy badawcze

Reprezentacja gry chaosu wraz z wybranymi deskryptorami stanowi adekwatny i wszechstronny model do opisu sekwencji nukleotydowych oraz aminokwasowych, umożliwiając analizę ich składu, motywów i zależności filogenetycznych.

Zastosowanie reprezentacji gry chaosu umożliwia identyfikację charakterystycznych wzorców w grupie sekwencji ludzkich mikroRNA oraz genów niezbędnych u bakterii.

Rozszerzona reprezentacja gry chaosu pozwala na precyzyjne odwzorowanie cech strukturalnych i funkcjonalnych sekwencji biologicznych oraz ich ewolucyjnych relacji.

2.2. Cele pracy

Celem rozprawy doktorskiej jest zbadanie zakresu stosowalności i ograniczeń CGR w odniesieniu do analizy sekwencji nukleotydowych oraz jej rozwinięcie na przypadek sekwencji aminokwasowych. Metoda ta była już wielokrotnie wykorzystywana do analizy sekwencji biologicznych, w tym ich porównywania. Autor z jednej strony systematyzuje pewne rozwiązania, takie jak proponowane w literaturze techniki oparte o CGR, a drugiej bada własności tej reprezentacji oraz proponuje potencjalnie efektywniejsze jej modyfikacje. W pracy, autor wykorzystuje CGR do scharakteryzowania motywów sekwencyjnych pośród ludzkich mikroRNA, cech sekwencji genów niezbędnych u bakterii oraz rekonstrukcji relacji ewolucyjnych genów, genomów oraz białek. Dodatkowo, autor bada własności kodowania sekwencji nukleotydowych oraz aminokwasowych za pomocą CGR oraz proponuje użyteczne deskryptory.

mikroRNA, czyli krótkie jednoniciowe cząsteczki RNA o długości około 22 nukleotydów, regulują ekspresję genów m.in. poprzez oddziaływanie z transkryptem [45]. To może prowadzić do jego degradacji. Cząsteczki te charakteryzuje duża różnorodność, jednak wskazano też na pewne ich wspólne cechy. Większość badań koncentruje się jednak na analizie szczególnych przypadków. W pracy, autor stosuje CGR do wizualizacji motywów sekwencyjnych obecnych w zbiorze ludzkich mikroRNA. Celem jest z jednej strony potwierdzenie zapewne znanych z literatury wzorców, a z drugiej wskazanie potencjalnie nowych. Na podstawie obliczeń energii oddziaływania mikroRNA/mRNA, cała pula miRNA jest podzielona na dwa podzbiory i zostaje przeprowadzona analiza częstości występowania nukleotydów oraz dinukloetydów. Badania mogą w przyszłości pozwolić na łatwiejszą identyfikację nowych miRNA poprzez m.in. wskazanie konserwatywnych motywów sekwencyjnych. Inne potencjalne zastosowania obejmują projektowanie nowych miRNA przy terapiach genowych czy badaniach funkcjonalnych genów.

Geny niezbędne odpowiedzialne są za przetrwanie i reprodukcję organizmu w określonych warunkach [22]. Badania eksperymentalne są zazwyczaj czasochłonne i kosztowne. Stąd, następuje wzrost zainteresowania przewidywania tego typu genów za pomocą metod komputerowych. Klasyczna analiza obejmuje m.in. wykrywanie homologii. Z kolei modele klasyfikacji typowo wymagają pewnych deskryptorów do opisu danych. W literaturze pojawiło się kilka prac dotyczących zastosowania reprezentacji gry chaosu i opartych na nich miar, takich jak wykładnika Hursta, do rozróżnienia puli genów niezbędnych od pozostałych [26,27]. Cechy te mogą w przyszłości posłużyć do budowy jeszcze skuteczniejszych modeli klasyfikacji, a w konsekwencji pełniejszego zrozumienia podstaw życia czy przyczynić się do rozwoju biologii

11:7877530921

syntetycznej. W tej pracy, autor przedstawia kilkuetapową analizę tych genów, gdzie celem jest z jednej strony charakterystyka genów niezbędnych, a z drugiej poszukiwanie kolejnych użytecznych deskryptorów. Analiza obejmuje badanie częstości i rozkładu występowania reszt, a także próbę uchwycenia zależności dalekiego zasięgu. W tym celu zostaje wykorzystany tzw. spacer DNA połączony z reprezentacją gry chaosu. Na ich podstawie są wyznaczane odpowiednie deskryptory.

Jednym z najważniejszych celów tej pracy jest ocena skuteczności porównywania sekwencji biologicznych w oparciu o reprezentację gry chaosu. Badaniu poddane są zarówno sekwencje genów, genomów jak i białek. Dodatkowo, autor generuje kilka zestawów sekwencji referencyjnych o różnym stopniu podobieństwa, ale znanych relacjach ewolucyjnych. Poszukiwanie szybkich i efektywnych metod porównywania sekwencji stanowi nadal jedno z głównych wyzwań bioinformatyki. W literaturze przedstawiono wiele metod połączonych z reprezentacją gry chaosu. Wskazano m.in. na indeks podobieństwa strukturalnego, dyskretną transformatę Fouriera czy wykładnik Hursta [34,35,36]. Najczęściej jednak prace opierają wnioski na kilku przykładach. Stąd, pojawia się potrzeba analiz większej liczby zestawów danych aby w sposób wiarygodny ocenić ich potencjalną praktyczną przydatność. Autor wprowadza też dodatkowe modyfikacje, jak np. wyznaczanie nie jednego, a czterech wykładników Hursta aby lepiej odzwierciedlić charakter badanych sekwencji. Do porównywania wygenerowanych drzew filogenetycznych z referencją zostaje zastosowana odległość Jaccarda-Robinsona-Fouldsa [43,44].

Reprezentacja gry chaosu często wykorzystywana jest również jako metoda osadzania sekwencji biologicznych w modelach uczenia maszynowego. W tej pracy, badaniu zostają poddane własności między innymi pamięci czy stabilności takiej reprezentacji. Autor sprawdza wpływ mutacji takich jak substytucje, insercje czy delecje na ostateczną reprezentację. Dodatkowo, autor proponuje miarę do opisu tych zmian.

Znaczna część pracy jest także poświęcona rozszerzeniom klasycznej reprezentacji gry chaosu. Obejmują one zarówno sekwencje nukleotydowe jak i białkowe. Przedstawiona zostaje ogólniejsza wersja tej reprezentacji. Celem jest uzyskanie potencjalnie bardziej użytecznej metody kodowania sekwencji z zachowaniem najważniejszych własności oryginalnej reprezentacji, jak jednoznaczności. Przetestowane zostają przykłady rekonstrukcji relacji ewolucyjnych na przykładzie genu neuraminidazy u ptasich wirusów grypy oraz genu SPARC (ang. *Secreted Protein Acidic and Cysteine Rich*) u organizmów wyższych . W przypadku sekwencji RNA dodana jest możliwość uwzględnienia informacji o jej strukturze drugorzędowej (jeżeli taka jest znana). Z jednej strony, może to posłużyć do wizualizacji rozmaitych struktur dzielonych między różnymi sekwencjami RNA. Z drugiej, takie podejście może stanowić swego rodzaju walidację do struktur generowanych przez różne algorytmy. W przypadku sekwencji białkowych sprawdzone zostają różne parametryzacje aminokwasów. Najprostsze podejście polega na redukcji alfabetu dwudziestu znaków do czterech i zastosowaniu reprezentacji gry chaosu podobnie jak dla sekwencji nukleotydowych. Sprawdzone zostają różne własności fizykochemiczne aminokwasów. Przedstawiona jest też metoda dynamicznego tworzenia parametryzacji aminokwasów dla wybranej grupy sekwencji. Zbadany zostaje wpływ tych modyfikacji na generowanie drzew filogenetycznych dla białek SPARC oraz ND5.

12:6281967222

Rozdział 3. Metody i zasoby

3.1. Biologiczne bazy danych dotyczące sekwencji biologicznych

Bazy danych sekwencji biologicznych są podstawowymi zasobami w bioinformatyce, gromadzącymi informacje o sekwencjach DNA, RNA i białek [12, 46]. Wyróżnia się wiele rodzajów baz danych w zależności od stopnia przetworzenia zawartych informacji.

Bazy danych pierwszorzędowe zawierają surowe sekwencje uzyskane bezpośrednio z eksperymentów sekwencjonowania [12]. Przykładami takich baz są GenBank, prowadzony przez NCBI (*National Center for Biotechnology Information*), który zawiera sekwencje nukleotydowe wraz z odpowiednimi metadanymi, oraz ENA (*European Nucleotide Archive*) prowadzony przez EMBL-EBI, będący europejskim odpowiednikiem dla GenBank [47,48].

Bazy danych drugorzędowe przetwarzają i interpretują surowe sekwencje, dodając informacje takie jak struktury białek, domeny funkcjonalne i motywy [12]. Przykładem jest UniProt, który oferuje bogate adnotacje dotyczące sekwencji białkowych, ich funkcji oraz interakcji [49]. Innym przykładem jest Pfam, baza danych zawierająca informacje o rodzinach białek i ich domenach [50].

Wreszcie, istnieją bazy danych trzeciorzędowe, które integrują dane z różnych źródeł, oferując bardziej zaawansowane analizy i narzędzia [12]. Przykładem jest Ensembl, który łączy dane genomowe z różnych gatunków, umożliwiając porównania ewolucyjne i funkcjonalne [51].

Bazy danych sekwencji biologicznych, takie jak GenBank, ENA, UniProt i Ensembl umożliwiają przechowywanie, analizę i interpretację ogromnych ilości danych genetycznych i proteomicznych. Dzięki nim możliwe jest prowadzenie zaawansowanych badań w genomice, proteomice, filogenetyce i wielu innych dziedzinach biologii.

Prócz baz danych szerokiego zastosowania występują także bazy specjalistyczne, skupiające np. sekwencje biologiczne określonego typu czy poświęcone poszczególnym jednostkom chorobowym (np. bazy danych dotyczące mutacji). Przy analizie zastosowań CGR, autor badał m.in. relacje wśród tzw. genów niezbędnych u bakterii. Dane pochodziły z DEG 15 (*Database of Essential Genes*, strona internetowa: http://origin.tubic.org/deg/), która jest specjalistycznym zasobem gromadzącym informacje na temat genów niezbędnych dla przetrwania różnych organizmów [42]. Geny te są kluczowe dla podstawowych procesów życiowych, a ich usunięcie lub mutacja prowadzi do śmierci komórki lub organizmu [52-56]. Baza danych DEG zawiera informacje o genach niezbędnych dla różnych organizmów, w tym bakterii, archeonów i eukariontów. Dane te są zbierane z literatury naukowej oraz eksperymentów, takich jak knockout genów, badania z wykorzystaniem interferencji RNA i CRISPR-Cas9. DEG jest nieocenionym narzędziem w badaniach medycznych, szczególnie w identyfikacji nowych celów dla terapii przeciwnowotworowych i antybiotyków. W biotechnologii wiedza o genach niezbędnych jest kluczowa dla inżynierii metabolicznej i syntezy nowych szlaków metabolicznych. Analiza genów niezbędnych dostarcza również wglądu w ewolucję genów i ich funkcji. Dzięki narzędziom dostępnym online, użytkownicy mogą przeszukiwać i analizować dane DEG według różnych kryteriów, co wspiera rozwój nowych terapii oraz technologii.

Przy badaniu mikroRNA szczególnie istotne są bazy danych miRBase oraz miRTarBase. miRBase 22 stanowi repozytorium tych sekwencji oraz ich adnotacji, dostarczając standardowego

13:8619214097

nazewnictwa dla mikroRNA i szczegółowych informacji o ich lokalizacji oraz strukturze [38,39]. Część rekordów to te o najwyższym poziomie wiarygodności, m.in. potwierdzone eksperymentalnie (ang. *confident*). Z kolei miRTarBase 8.0 zawiera informacje o interakciach mikroRNA z docelowymi transkryptami [40]. W szczególności koncentruje się na eksperymentalnie potwierdzonych interakcjach miRNA-mRNA, co czyni ją jedną z najbardziej wiarygodnych i użytecznych baz do badania regulacji genów przez mikroRNA. Jest zintegrowana z innymi zasobami, takimi jak miRBase, NCBI Gene, czy Ensembl.

Prócz baz danych szczególną uwagę należy zwrócić na różnego typu narzędzia online do analizy sekwencji biologicznych jak np. *emboss* [57]. Inne umożliwiają weryfikację nowych metod i porównanie ich z obecnymi rozwiązaniami. AFproject.org (strona internetowa: https://afproject.org/) to platforma internetowa stworzona do analizy i porównania metod porównywania sekwencji biologicznych bez dopasowania [58]. Na stronie znajdują się:

- **Zestawy danych referencyjnych** podzielone na kategorie, takie jak sekwencje regulatorowe, klasyfikacja białek, analiza filogenezy genomowej czy transfer genów.
- **Narzędzia** pozwalające porównać skuteczność różnych metod bez dopasowania na podstawie predefiniowanych miar.
- **Możliwość testowania własnych metod** – użytkownicy mogą przesyłać wyniki swojej analizy, by porównać je z innymi dostępnymi metodami.

Platforma ma na celu standaryzację oceny jakości narzędzi i wspieranie naukowców w ulepszaniu ich algorytmów. Wyniki analiz mogą być przechowywane jako prywatne lub udostępniane publicznie. W tej pracy wykorzystano tę platformę do testowania metod porównywania sekwencji opartych o reprezentację gry chaosu.

3.2. Porównywanie sekwencji i tworzenie drzew filogenetycznych

Dopasowanie sekwencji jest pierwszym i często kluczowym krokiem w wielu analizach bioinformatycznych. Umożliwia porównanie sekwencji DNA, RNA lub białek w celu zidentyfikowania podobieństw i różnic między nimi. Proces ten polega na przyrównaniu sekwencji nukleotydów lub aminokwasów w taki sposób, aby jak najlepiej odzwierciedlać ewolucyjne i funkcjonalne relacje między nimi. Istnieją dwa główne typy dopasowania sekwencji: dopasowanie globalne i dopasowanie lokalne [12]. Dopasowanie globalne (np. algorytm Needleman-Wunsch) porównuje sekwencje na całej ich długości, natomiast dopasowanie lokalne (np. algorytm Smith-Waterman) skupia się na znalezieniu najbardziej podobnych fragmentów w sekwencjach.

Metody bez dopasowania (ang. *Alignment-Free Methods,* AF) polegają na analizie globalnych lub lokalnych cech sekwencji, takich jak częstości występowania merów (czyli podsekwencji o zadanej długości) lub wyznaczane są bardziej zaawansowane wskaźniki, jak złożoność Lempel-Ziv [13,14,15]. Mają one kilka istotnych zalet w porównaniu z metodami opartymi na dopasowaniu sekwencji. Przede wszystkim są one zazwyczaj szybsze i bardziej skalowalne, co czyni je atrakcyjnymi w analizie bardzo dużych zbiorów danych, takich jak całe genomy. Ponadto, metody bez dopasowania są mniej podatne na problemy związane z lokalnymi zmianami w sekwencjach, takimi jak insercje, delecje czy inne rodzaje mutacji, które mogą zakłócać dokładne dopasowanie. Dzięki temu, metody te mogą być bardziej uniwersalne i efektywne w porównywaniu sekwencji

14:9518934549

o niskim stopniu podobieństwa lub różnej długości. Metody bez dopasowania są także bardziej elastyczne w przypadku analizy sekwencji o niejednorodnej strukturze, gdzie klasyczne metody dopasowania mogą nie działać efektywnie z powodu wysokiej zmienności i złożoności danych. Ponadto, metody bez dopasowania są często mniej wymagające pod względem zasobów obliczeniowych, co czyni je bardziej dostępnymi w zastosowaniach, gdzie zasoby są ograniczone.

W kontekście filogenetyki, porównanie sekwencji jest często pierwszym krokiem w tworzeniu drzew filogenetycznych, które obrazują ewolucyjne relacje między organizmami lub genami. Po przyrównaniu sekwencji, uzyskane dane są analizowane za pomocą różnych metod filogenetycznych, takich jak metody oparte na maksymalnej parsymonii, maksymalnej wiarygodności czy metodach bayesowskich [12].

Metoda maksymalnej parsymonii (ang. *Maximum Parsimony*) polega na poszukiwaniu drzewa filogenetycznego, które minimalizuje liczbę zmian ewolucyjnych [59]. Jest to podejście zakładające, że najprostsze wyjaśnienie jest najbardziej prawdopodobne.

Metody oparte na maksymalnej wiarygodności (ang. *Maximum Likelihood*) oceniają różne drzewa filogenetyczne i wybierają to, które najlepiej wyjaśnia obserwowane dane sekwencji [60]. Metoda ta uwzględnia różne modele ewolucyjne, co pozwala na bardziej precyzyjne oszacowanie relacji filogenetycznych.

Metody bayesowskie, realizowane przez takie programy jak MrBayes, wykorzystują podejście probabilistyczne, integrując informacje z danych sekwencji z wcześniejszymi informacjami na temat ewolucji [61]. Metody te generują drzewa filogenetyczne oraz oszacowania prawdopodobieństwa dla każdej gałęzi drzewa, co pozwala na ocenę pewności wyników.

Inną grupę stanowią metody oparte na odległościach, które wykorzystują informacje o podobieństwie lub różnicy między sekwencjami [12]. Podstawą tych metod jest obliczanie odległości między każdą parą sekwencji w zbiorze danych, na podstawie których budowane są drzewa filogenetyczne. Po obliczeniu macierzy odległości, drzewo filogenetyczne jest konstruowane za pomocą algorytmów hierarchicznych [62]. Przykładami są tutaj metoda średnich połączeń (ang. *Unweighted Pair Group Method With Arithmetic Mean,* UPGMA*)* czy najbliższego sąsiada (ang. *Neighbor-Joining,* NJ) [63,64]. Metoda UPGMA zakłada równomierne tempo ewolucji w obrębie badanych linii rozwojowych i buduje drzewo filogenetyczne, grupując sekwencje na podstawie ich podobieństwa i średnich odległości między nimi [12]. Metoda NJ jest bardziej elastyczna, pozwalając na uwzględnienie nieliniowych zmian ewolucyjnych i zmiennego tempa ewolucji w obrębie analizowanych danych [12]. Algorytm ten iteracyjnie łączy najbliższe pary sekwencji, aby zbudować drzewo, minimalizując całkowitą długość gałęzi. Rzadziej wykorzystywane są też inne podejścia jak metoda Warda mająca na celu minimalizację wariancji wewnątrz grup [65]. Jest ona jednak stosunkowo odporna na nieliniowe zależności w danych, co oznacza, że może lepiej radzić sobie z nietypowymi kształtami klastrów lub zmiennymi grupami. Metody oparte na odległościach są stosunkowo proste w zastosowaniu i wydajne obliczeniowo, co czyni je popularnymi w analizie dużych zestawów danych. Pomimo pewnych ograniczeń, takich jak założenie o stałym tempie ewolucji w przypadku UPGMA, są one nadal używane w badaniach filogenetycznych, zwłaszcza w przypadku analiz porównawczych między dużymi grupami organizmów.

Porównywnie sekwencji czy tworzenie drzew filogenetycznych są kluczowymi technikami w badaniach ewolucyjnych, pozwalającymi na zrozumienie w jaki sposób organizmy i geny

15:8196033311

ewoluowały oraz jak są ze sobą spokrewnione. Dzięki tym narzędziom naukowcy mogą odkrywać mechanizmy ewolucyjne, identyfikować geny (lub ich regiony) konserwatywne i unikalne, a także rekonstruować historie ewolucyjne całych grup organizmów.

3.3. Symulowanie sieci filogenetycznych

Aby ocenić skuteczność nowych metod rekonstrukcji drzew filogenetycznych, badacze często korzystają ze sztucznie generowanych sekwencji genetycznych [66,67]. W tym procesie najpierw tworzy się hipotetyczne drzewo filogenetyczne, które reprezentuje znane relacje ewolucyjne. Następnie na podstawie tego drzewa symuluje się ewolucję sekwencji, uwzględniając różne procesy ewolucyjne, takie jak mutacje. Dzięki temu powstają sztuczne sekwencje genetyczne, których historia ewolucyjna jest dokładnie znana.

Używając tych sztucznie wygenerowanych sekwencji referencyjnych, można następnie zastosować nowe metody rekonstrukcji drzew filogenetycznych i porównać otrzymane wyniki z oryginalnym, znanym drzewem. Tego typu walidacja pozwala na obiektywną ocenę dokładności i skuteczności nowych metod, umożliwiając identyfikację ich mocnych i słabych stron w różnych scenariuszach ewolucyjnych.

3.4. Porównywanie drzew filogenetycznych

Porównywanie drzew filogenetycznych jest istotnym krokiem w ewolucyjnej analizie danych, pozwalającym na ocenę zgodności różnych metod rekonstrukcji drzewa oraz na weryfikację hipotez dotyczących historii ewolucyjnej organizmów. Ogólnie, porównywanie drzew filogenetycznych można przeprowadzać za pomocą różnych technik [68]. Jedną z najczęściej używanych miar jest odległość Robinsona-Fouldsa (RF) [69]. Jest to metoda, która mierzy liczbę operacji, takich jak cięcie i sklejanie gałęzi, potrzebnych do przekształcenia jednego drzewa w drugie. Miara RF porównuje struktury dwóch drzew, ignorując długości gałęzi, co czyni ją przydatną do oceny zgodności topologii drzew. Bardziej precyzyjnie

$$RF = |S_1 \backslash S_2| + |S_2 \backslash S_1|, \tag{1}$$

gdzie S_1, S_2 to zbiory podziałów dla obydwu drzew, a $|\ |$ oznacza moc zbioru.

Odległość Jaccarda-Robinsona-Fouldsa (JRF) jest rozszerzeniem miary Robinsona-Fouldsa [43,44]. W JRF obliczamy liczbę wspólnych podziałów między dwoma drzewami, dzieląc ją przez łączną liczbę klastrów w obu drzewach (miara Jaccarda, J)

$$J = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}. \tag{2}$$

Następnie, odległość JRF wyznaczana jest jako

$$JRF = 1 - J = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}. \tag{3}$$

Dzięki temu JRF może lepiej oddać stopień pokrewieństwa między drzewami, uwzględniając zarówno wspólne, jak i różne elementy ich struktur. Rozpisując prawą stronę

16:5182110500

$$\frac{|S_1 \cup S_2|}{|S_1 \cup S_2|} - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = \frac{|S_1 \cup S_2| - |S_1 \cap S_2|}{|S_1 \cup S_2|} = \frac{|S_1 \backslash S_2| + |S_2 \backslash S_1|}{|S_1 \cup S_2|} = \frac{RF}{|S_1 \cup S_2|}. \quad (4)$$

.

Widzimy, że JRF jest w pewnym sensie znormalizowaną wersję RF przyjmującą wartości z zakresu [0,1], gdzie 0 oznacza identyczność, 1 maksymalne zróżnicowanie. Istnieje też bardziej ogólna definicja odległości JRF

$$JRF = \frac{RF^k}{N}, \quad (5)$$

gdzie k to pewna stała, a N czynnik normujący. Wtedy odległości nie muszą już znajdować się w przedziale [0,1]. W przypadku klasycznej wersji mamy $k = 1, N = |S_1 \cup S_2|$. Niektóre oprogramowania jak np. metoda *JaccardRobinsonFoulds z* **R** (biblioteka *TreeDist*) z której korzystano w tej pracy domyślnie zwracają wartości nieunormowane. Aby porównać uzyskane odległości można je porównać z tymi dla drzew losowych. Dzięki temu mamy dodatkowo informację o rozkładzie odległości. Inne, prostsze i częściej używane podejście polega na podzieleniu odległości przez maksymalne możliwe wartości (suma liczby podziałów).

Porównywanie drzew filogenetycznych za pomocą tych miar pozwala na ocenę, jak dobrze różne metody rekonstrukcji drzewa oddają rzeczywiste relacje ewolucyjne między organizmami. Dzięki temu możliwe jest zidentyfikowanie metod, które najlepiej radzą sobie w różnych scenariuszach ewolucyjnych, oraz weryfikacja hipotez dotyczących historii ewolucyjnej badanych organizmów.

3.5. Specyfika sekwencji nukleotydowych oraz prawa biologiczne

W obrębie sekwencji biologicznych odkrytych zostało wiele zależności. Nie są to oczywiście prawa w ścisłym tego słowa znaczeniu, jednak są na tyle charakterystyczne, że warto zwrócić na nie uwagę.

Na początku trzeba podkreślić, że sekwencje biologicznie nie są losowe, co oznacza w tym przypadku, że zarówno częstości występowania poszczególnych nukleotydów jak i ich rozmieszczenie nie jest jednostajne. Już prosta statystyka wyznaczająca procent GC w genomach pokazuje, że róże organizmy mają odmienne ilości tych dwóch nukleotydów [70,71]. Drobniejsze różnice pojawiają się też oczywiście w obrębie gatunków. W przypadku genów prokariotycznych w regionach promotorowych nade często obserwuje się charakterystyczną sekwencję TATAAT [72,73]. Innym przykładem jest występowanie wysp GC w obrębie regionów promotorowych u ssaków [74,75]. Kolejnym przykładem są liczne powtórzenia w genomie czy charakterystyczne sekwencje do których przyłączają się czynniki transkrypcyjne i inne białka [76].

Jeżeli chodzi o generalne prawidłowości, to oczywiście nie sposób nie wspomnieć o kodzie genetycznym który oprócz kilku istotnych wyjątków jest dzielony pomiędzy orgazmami [77,78]. Istnieje wiele interesujących prób wyjaśnienia obecnej postaci kodu genetycznego. Ciekawe podejście zaproponował Petoukhov, który w oparciu o teorię macierzy i analizę sygnałów podzielił kodony na dwie grupy [79]. Innym podejściem jest zastosowanie tzw. liczb p-adycznych [80]. Okazuje się, że kodony które są mniej odległe w sensie metryki opartej na liczbach 2- i 5-adycznych kodują te same aminokwasy.

Warto też zaznaczyć, że u organizmów często występuje też pewna preferencja kodonów (ang. *codon bias*) [81]. Dotyczy to szczególnie genów ulegającej wysokiej ekspresji. Jednym

17:1039367270

z parametrów opisujących to zjawisko się wskaźnik CAI (ang. *Codon Adaptation Index*) [82]. Wacław Szybalski zauważył, że u wirusów występuje średnio więcej puryn niż pirymidyn [83]. Ma to zapobiegać tworzeniu struktur dwuniciowych pomiędzy fragmentami RNA. To z jednej strony spowalnia replikację RNA, a z drugiej może wpływać na formowanie się białek.

Kolejne prawidłowości stanowią tzw. reguły Chargaffa [84]. Dotyczą one przede wszystkim organizmów dwuniciowych. Pierwsza z nich stwierdza, że u organizmów na obu niciach występuje ta sama ilość adenin co tymin, a guanin co cytozyn. Uzasadnienie tego prawa jest proste i wynika z obecności podwójnej helisy w jądrze oraz zasady komplementarności. Mniej trywialna jest druga reguła Chargaffa stwierdzająca, że podobna prawidłowość z przybliżeniem występuje także na każdej z nici. Przez długi czas poszukiwano biologicznego uzasadnienia tego faktu odwołując się między innymi do transpozonów. W 2021 roku ukazała się praca, w której podjęto się tego tematu na gruncie fizyki statystycznej [85]. W oparciu o kilka naturalnych założeń dotyczących symetrii udało się dowieść tej własności. Co więcej, okazuje się, że można tę regułę uogólnić stwierdzając, że na jednej nici ilość danej sekwencji i odwrotnie do niej komplementarnej jest w przybliżeniu jednakowa. Dla przykładu, ilość ATGG i CCAT jest podobna. Warto tutaj podkreślić, że ta sama zależność nie występuje dla zwykłej komplementarności. W przypadku genomu *E. coli* stosunki A/T oraz G/C wynoszą odpowiednio 0.997 oraz 1.023. W przypadku dinukleotydów średni stosunek pomiędzy daną sekwencją a do niej odwrotnie komplementarną wynosi 1.007 z odchyleniem 0.030. Nie wliczano przy tym przypadków, gdzie sekwencja sama dla siebie jest odwrotnie komplementarna jak np. dla GC. Podobne obliczenia można dokonać dla dłuższych sekwencji. Wyniki zestawiono w tabeli. Obserwujemy istotne odstępstwo w przypadku zwykłej komplementarności. Ze wzrostem długości rozważanych sekwencji bardzo szybko rośnie też odchylenie standardowe wyników. Z kolei dla sekwencji odwrotnie komplementarnych stosunek bliski jedności występuje nawet dla dłuższych fragmentów. W tym przypadku nie obserwujemy też gwałtownego wzrostu odchylenia standardowego.

Tabela 1. Porównanie częstości występowania podsekwencji i do nich komplementarnych oraz odwrotnie komplementarnych dla genomu E. coli (NCBI ID: HG738867.1). Podano średnie wartości stosunków częstości oraz odchylenia.

	dinukleotydy	trinukleotydy	tetranukleotydy
komplementarność	1.033 (0.206)	1.127 (0.585)	1.342 (1.928)
odwrotna komplementarność	1.007 (0.030)	1.011 (0.033)	1.011 (0.044)

3.6. Metody opisu i reprezentacji sekwencji nukleotydowych i białkowych

3.6.1. Entropia i logo

Zliczenie wystąpień pojedynczych reszt stanowi podstawową statystykę na temat sekwencji. Mimo prostoty, już na podstawie tego parametru można dokonać pewnej analizy. Przykładem jest choćby zawartość GC, czyli suma częstości guaniny i cytozyny, która może służyć do rozpoznania miejsc występowania genów. W ogólności, jeżeli rozważamy alfabet n literowy, to dla danej sekwencji można policzyć częstości f_i odpowiadające procentowej zawartości i-tej reszty, gdzie $i = 1, \dots, n$. Otrzymujemy zatem wektor wartości

18:4882854890

$$f = [f_1, f_2, \dots, f_n]. \tag{6}$$

W przypadku sekwencji nukleotydowych $n = 4$, białkowych $n = 20$. Parametrem, który wykorzystuje te wielkości jest entropia Shannona [86]

$$H = -\sum_{i=1}^{n} f_i \, log_2 f_i. \tag{7}$$

Określa ona na ile skład danej sekwencji odbiega od rozkładu jednostajnego. W przypadku sekwencji nukleotydowych i jednorodnych wartości częstości, tzn. $f_i = \frac{1}{4}$, wartość entropii wynosi

$$H = -\sum_{i=1}^{4} \frac{1}{4} log_2 \frac{1}{4} = -4 * \frac{1}{4} * (-2) = 2 = log_2 4. \tag{8}$$

Analogicznie, w przypadku sekwencji białkowych $H = log_2 20$. Są to wartości maksymalne. Z drugiej strony, jeżeli w sekwencji występowałaby tylko jedna reszta, to $H = 0$ (zakładając $0 \log_2 0 = 0$). Mając do dyspozycji N sekwencji o tej samej długości (ewentualnie po dopasowaniu) można wyznaczyć relatywną ilość informacji przypadającą na kolejną pozycję j zgodnie z

$$R_j = log_2 4 - H_j \tag{9}$$

lub

$$R_j = log_2 20 - H_j \tag{10}$$

gdzie

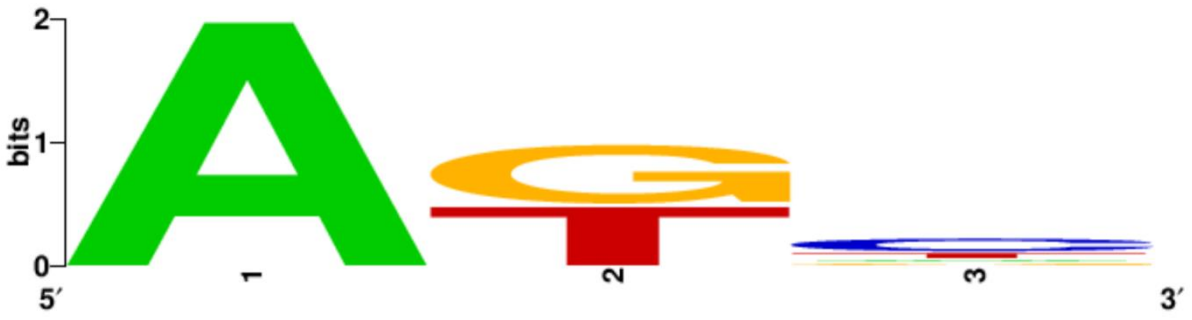
$$H_j = -\sum_{i=1}^{n} f_{ij} \, log_2 f_{ij}, \tag{11}$$

gdzie f_{ij} jest częstością i-tej reszty w j-tej kolumnie. Takie wyznaczanie entropii dla kolejnych pozycji zwane jest logiem i często prezentowane graficznie. Wysokość i -tej litery w j-tej kolumnie dana jest przez

$$f_{ij} R_j. \tag{12}$$

Z uwagi na pewne przyczyny techniczne i praktyczne, programy najczęściej modyfikują lekko częstości uwzględniając przy tym zliczenia pozorne. Dodatkowo, relatywna ilość informacji również uwzględnia liczbę porównywanych sekwencji. Dla dużej liczby sekwencji wyniki są zbieżne z powyższą formułą. Przykładowe logo wygenerowane poprzez serwis WebLogo [87,88] zaprezentowano poniżej:

19:2023815288



Rysunek 1.Logo wygenerowano dla 200 sekwencji nukleotydowych o długości 3. Wszystkie sekwencje posiadały A na pierwszej pozycji. Na drugiej pozycji 100 sekwencji posiadało G, a pozostałe T. Na trzeciej pozycji C wystąpiło w 100 sekwencjach, T w 50, A w 25 oraz G w 25.

W przypadku gdy na danej pozycji w sekwencji nukleotydowej występuje tylko jedna z liter, obserwujemy ją na wysokości 2. Z kolei, gdy występują tylko dwie litery w tej samej ilości, łączna wysokość stanowi 1, a wysokości tych liter stanowią po połowie tej wartości. W przypadku pozycji, gdzie występują różne reszty z różnymi częstościami obserwujemy istotnie niższą łączną wysokość – w przypadku rozkładu jednorodnego byłoby to 0.

3.6.2. K-mery i analiza częstości

Chociaż analiza pojedynczych reszt dostarcza istotnych informacji dotyczące sekwencji, następuje tutaj jednak znacząca redukcja informacji. W przypadku prostej analizy częstościowej tracimy informacje na temat położenia samych reszt w obrębie sekwencji. Problem ten częściowo został rozwiązany przez logo, jednak i w tym przypadku kolejne pozycje traktowane są niezależnie. Aby uwzględnić informację na temat szerszego kontekstu stosuje się analizę tzw. k-merów [89]. Sekwencja dzielona jest na mniejsze części – tzw. mery, czyli nakładające się podsekwencje o zadanej długości. Dla przykładu biorąc $k = 3$, dla sekwencji $ATGCAC$ jej 3-mery to ATG, TGC, GCA oraz CAC. Analiza takich fragmentów może być przeprowadzona za pomocą podobnych technik jak tych wprowadzonych w poprzednim podrozdziale. W przypadku sekwencji zawierających reszty z mniejszego alfabetu stosowana jest większa wartość parametru k. W szczególności, dla sekwencji nukleotydowych preferowana wartość k jest istotnie większa niż dla sekwencji białkowych. Związane jest to z ilością potencjalnych k-merów. Dla przykładu, w przypadku alfabetu 4 literowego i $k = 5$ istnieje $4^5 = 1024$ unikalnych podsekwencji. Dla alfabetu 20 literowego i tej samej wartości k mamy 20^5, czyli ponad 3 miliony takich sekwencji. Poprzez zliczenie wystąpień wszystkich unikalnych k-merów można między innymi określić stopień złożoności sekwencji. Technika k-merów z powodzeniem stosowana jest do analizy danych z sekwencjonowania czy metod porównywania sekwencji bez dopasowania [90,91].

3.6.3. Złożoność Lempel-Ziv

Złożoność Lempel-Ziv stanowi metodę porównania sekwencji w oparciu o cechy lingwistyczne [15]. W podstawowej wersji wyznacza się tutaj liczbę słów wchodzących w skład pierwszej $C(x)$ i drugiej sekwencji $C(y)$, a także ich konkatenacji $C(xy)$. Słowo określa tutaj pewną możliwie najkrótszą podsekwencję. Przykładowo, dla sekwencji $x = ATAGAGC$ słowami są A, T, AG oraz

20:4523063539

C. Rozpoczynamy od lewej strony, jako pierwsze słowo bierzemy literę A, drugie słowo to T. Następnie występuje A, która już się pojawiła; wtedy wydłużamy nasz ciąg otrzymując AG. Kolejno, ponownie pojawia się AG, ale za nim następuje litera C, której do tej pory jeszcze nie było (zatem słowem będzie C a nie AGC).

Odległość liczona jest jako

$$d_{xy} = \frac{C(xy) - \min\big(C(x), (Cy)\big)}{\max\big(C(x), C(y)\big)}. \tag{13}$$

Przy takiej definicji wielkość ta w ogólnym przypadku nie jest symetryczna, bo $C(xy) \neq C(yx)$. Aby temu zapobiec ostateczną odległość można zdefiniować jako

$$d = \frac{d_{xy} + d_{yx}}{2}. \tag{14}$$

3.6.4. Krzywa Z

Poprzednie metody reprezentacji sekwencji opierały się pewnych miarach składu sekwencji, np. częstościach. Takie podejście, mimo, że z sukcesem stosowane, traci oczywiście zasadniczą informację o ułożeniu poszczególnych części w sekwencji. Dla przykładu zwykła statystyka zawartości GC dla całego genomu, mimo że może z powodzeniem służyć charakteryzacji gatunków, to oczywiście niewiele mówi o lokalnym znaczeniu sekwencji. Okazuje się, że w przypadku bakterii dla miejsc rozpoczęcia i zakończenia replikacji następuje istotna zmiana składu G i C [92]. Miarą umożliwiającą zlokalizowanie takich fragmentów jest parametr asymetrii w składzie G/C (ang. *GC skew*) definiowany jako

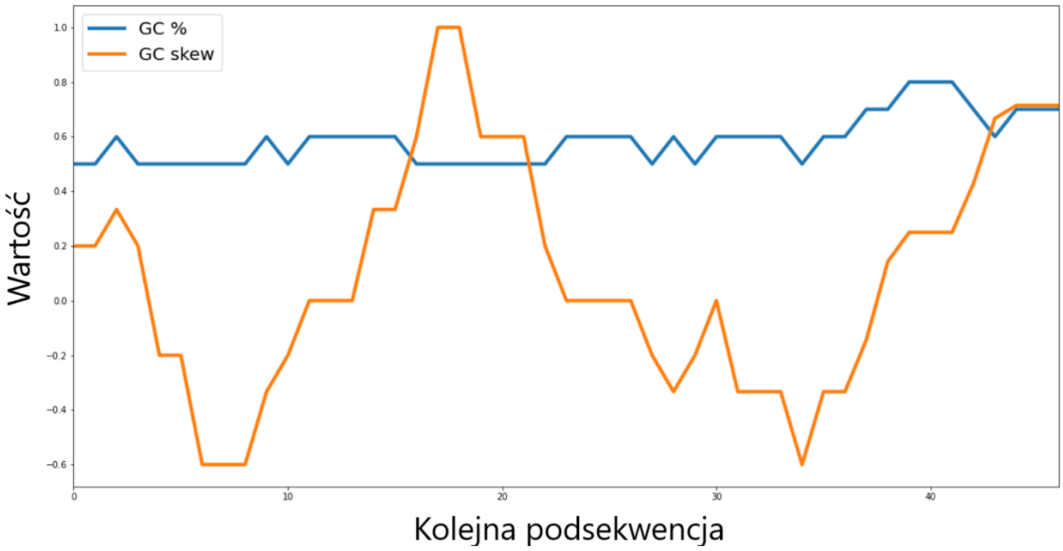
$$G_{skew} = \frac{G - C}{G + C}, \tag{15}$$

a liczony dla kolejnych nakładających się tzw. okien, czyli podsekwencji o zadanej długości. Dla przykładu, dla sekwencji

5’ - …GGATCGAAGCTAAGGGCT…-3’

i długości okna 10 rozważane są sekwencje GGATCGAAGC, GATCGAAGCT, ATCGAAGCTA, TCGAAGCTAA, CGAAGCTAAG, GAAGCTAAGG, AAGCTAAGGG, AGCTAAGGGC oraz GCTAAGGGCT. Następnie liczona jest miara G_{skew} dla każdej z nich. W efekcie otrzymujemy profil wartości dla kolejnych sekwencji. Oczywiście następuje tutaj też pewna utrata informacji, między innymi dlatego, że nie zajmujemy się zawartością A/T. Jednak zasadniczym atrybutem jest informacja o zachowaniu sekwencji na kolejnych pozycjach (informacja o zmianie zachowania) a nie tylko globalny parametr. Na poniższym rysunku przedstawiono przykładowy taki profil.

21:9447467587



Rysunek 2. Zawartość GC (GC %) oraz parametr asymetrii G/C (GC skew) dla kolejnych podsekwencji ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT. Przyjęto długość okna 10.

Zawartość GC informuje o średnim poziomie GC dla kolejnych podsekwencji. W przedstawionym przykładzie utrzymuje się na relatywnie podobnym poziomie. W przypadku asymetrii G/C widzimy miejsca bardziej i mniej bogate w G w stosunku do C (dodatnie lub ujemne wartości). Jeżeli ilość G jest porównywalna z C, to wartości skoncentrowane są w pobliżu zera.

Kolejnym opisem sekwencji, uwzględniającym tym razem wszystkie nukleotydy w jej obrębie jest tzw. krzywa Z [93]. Transformacja ta została wprowadzona w 1994, a dana jest za pomocą trzech zależności, wyliczając każdorazowo różnicę zliczeń pomiędzy danymi dwoma nukleotydami a pozostałymi

$$x_n = (A_n + G_n) - (C_n + T_n), \tag{16}$$

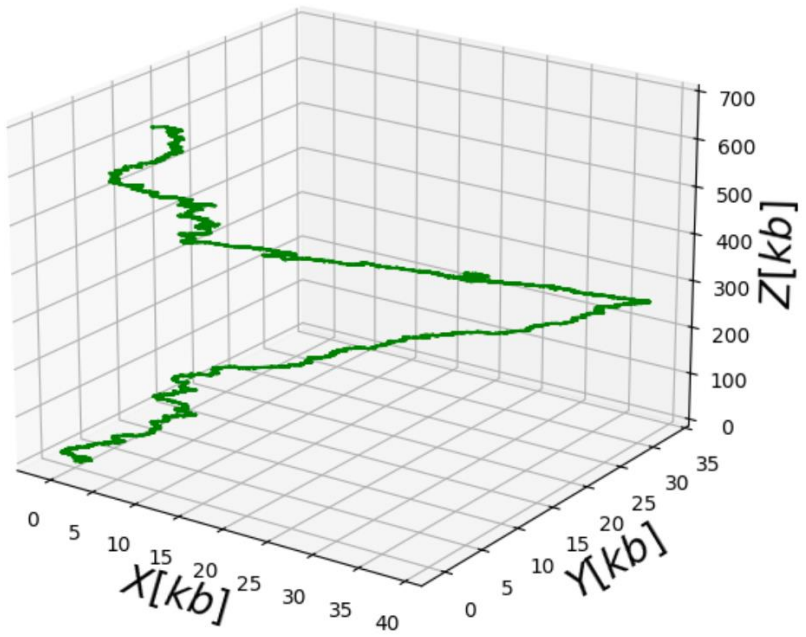
$$y_n = (A_n + C_n) - (G_n + T_n), \tag{17}$$

$$z_n = (A_n + T_n) - (C_n + G_n), \tag{18}$$

gdzie $n = 1, 2, \dots, N$ (N jest długością sekwencji), a A_n, C_n, G_n $oraz$ T_n reprezentują liczbę występowania odpowiednich nukleotydów dla kolejnych podsekwencji długości n zaczynając każdorazowo od pierwszej pozycji. Zauważmy, że taka transformacja wyczerpuje wszystkie możliwości kombinacji zestawienia ze sobą kolejnych par nukleotydów. Z definicji, otrzymujemy trójki liczb, a sama krzywa jest krzywą trójwymiarową. Uwzględnia ona dalekozasięgowe zależności oraz informuje o ewentualnym zmianie składu. Zastosowanie krzywej Z zostało bogato opracowane w literaturze, gdzie wskazano jej użyteczność m.in. w analizie genomu. Podobnie jak wprowadzony już parametr G_{skew} pozwala na rozpoznanie miejsc rozpoczęcia replikacji a także typowanie genów, identyfikację izochor czy realizowanie genomiki porównawczej [94]. Przykładem może być rozpoznanie oriC pośród Archeonów jako ostrego piku ze względu na bogatość tego regionu w AT [95]. Umożliwia również rozpoznanie intronów w obrębie genów. Inne jej zastosowania obejmują między innymi poszukiwanie relacji ewolucyjnych, w tym tworzenie drzew filogenetycznych [96]. W przypadku krótkich sekwencji, już sama analiza wizualna krzywej pozwala na rozpoznanie interesujących miejsc. Wraz z rosnącą długością stosowane są także automatyczne metody rozpoznawania i detekcji zachowania

22:1025840563

krzywej Z. Poniżej przedstawiono przykładową krzywą Z dla genomu *Methanosarcina mazei* (NCBI ID: NC_003901.1). Organizm ten należy do domeny Archeonów.



Rysunek 3. Krzywa Z reprezentująca genom Methanosarcina mazei (NCBI ID: NC_003901.1). Na osiach OX, OY oraz OZ przedstawione są kolejne pozycje w genomie w skali tysiąca par zasad (kb).

Obserwujemy miejsca bardziej jednorodne z określonym trendem jak i takie, gdzie krzywa widocznie zmienia kierunek i zwrot. Takie miejsca mogą wskazywać na występowanie wspomnianych wcześniej składowych genomu.

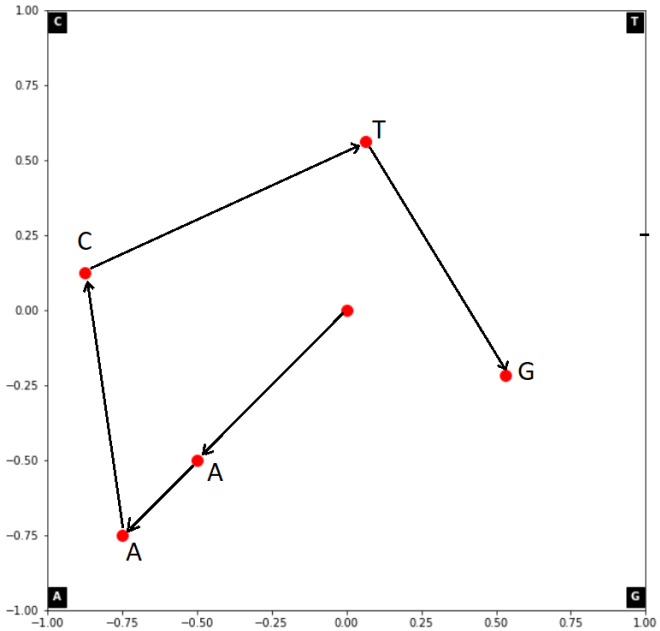
3.6.5. Reprezentacja gry chaosu

Reprezentacja gry chaosu stanowi główną metodę wykorzystywaną w tej pracy [16]. W dalszej części pracy będę też używał skrótu CGR (ang. *Chaos Game Representation*). Aby przekształcić sekwencję nukleotydową na numeryczną, należy w tym przypadku zastosować następującą, prostą transformację rekurencyjną

$$x_n = \frac{x_{n-1} + N}{2}, \tag{19}$$

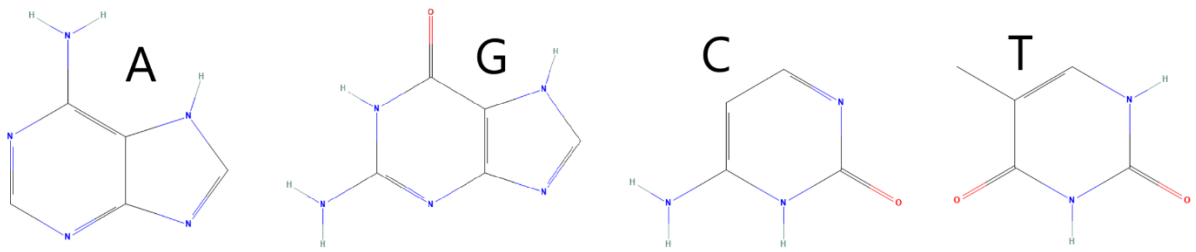
gdzie $x_0 = (0,0)$ jest punktem początkowym, a N jest aktualnym nukleotydem w sekwencji na pozycji n. Nukleotydy reprezentowane są jako wierzchołki kwadratu na płaszczyźnie. Dane są przez $A(-1,-1), C(-1,1), T(1,1)$ oraz $G(1,-1)$. Poniżej znajduje się przykładowa reprezentacja, stworzona dla sekwencji AACTG.

23:6816476967



Rysunek 4. Reprezentacja gry chaosu dla sekwencji AACTG. Startujemy z punktu $(0,0)$, a każdy kolejny kładziemy w połowie odległości pomiędzy poprzednim punktem a aktualnym nukleotydem (symbolizowany przez wierzchołek kwadratu).

Zaczynając więc z punktu $(0,0)$ idziemy w kierunku odpowiedniego wierzchołka kwadratu związanego z pierwszym nukleotydem i zatrzymujemy się w połowie drogi. Powtarzamy ten proces każdorazowo idąc od punktu postawionego w poprzednim kroku. Czasami przyjmuje się nieco inne wartości współrzędnych dla nukleotydów (wierzchołków kwadratu) lub podaje się je w innej kolejności. Przykładowo, środek znajduje się w $(0.5,0.5)$ a kolejne nukleotydy znajdują się w wierzchołkach kwadratu jednostkowego o wierzchołkach $(0,0), (0,1), (1,0), (1,1)$. Zauważmy, że występuje tutaj pewna asymetria. Litery A i T oraz G i C występują naprzeciw siebie. Pozostałe kombinacje obejmują sąsiednie nukleotydy, dla przykładu A i C lub T i G. Istnieje trzy istotnie różne możliwości umiejscowienia nukleotydów w wierzchołkach kwadratu. Dodatkowo, można nadać im interpretację.



Rysunek 5. Zasady azotowe występujące w DNA. Struktury pochodzą z bazy danych PubChem.

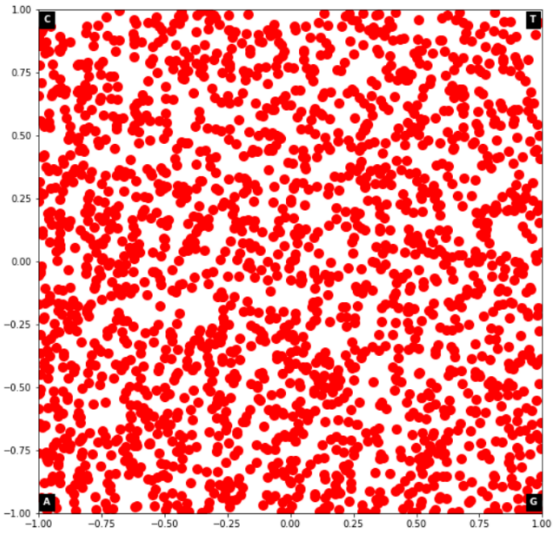
24:1137779297

Tabela 2. Własności fizykochemiczne oraz cechy zasad azotowych występujących w DNA wyznaczają naturalne ich podziały. Te stanowią potencjalne parametryzacje CGR – ułożenie nukleotydów w rogach kwadratu.

Kryterium podziału	Nukleotydy naprzeciw siebie
Puryny/Pirymidyny	A/G oraz C/T
Liczba wiązań wodorowych	A/T oraz G/C
Amino-mutujące (amino) i keto	A/C oraz G/T

W pierwszym przypadku rozważamy podział na puryny posiadające 2 pierścienie (A i G) oraz pirymidyny z 1 pierścieniem (C i T). W drugim, nukleotydy naprzeciw to te które tworzą tę samą liczbę wiązań wodorowych – dwa wiązania dla A i T oraz trzy wiązania dla C i G. Ostatnia grupa dotyczy podatności na mutacje. Okazuje się, że adenina i cytozyna w obecności kwasu azotowego (III) HNO_2 ulega deaminacji, co powoduje parowanie tych zasad azotowych w niewłaściwy sposób [97]. Związane jest to z obecnością grupy aminowej $-NH_2$. Adenina po deaminacji przekształca się w hipoksantynę (paruje się z C zamiast T), cytozyna w uracyl (który paruje się z A zamiast G). Możliwa (jednak występuje istotnie rzadziej) jest również deaminacja guaniny; w tym przypadku przekształcając ją w ksantynę, która ma ograniczone możliwości parowania.

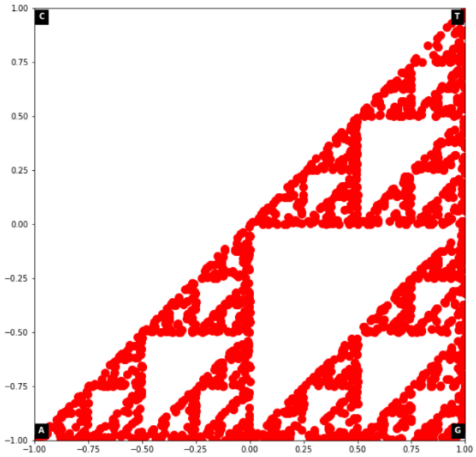
Wartości uzyskane po zastosowaniu CGR mogą być analizowane w postaci samego obrazu jak i ciągu liczb. w drugim przypadku, możemy traktować je jako ciąg liczb zespolonych, jako że dla każdej pozycji otrzymujemy dwie wartości (x oraz y). Już na podstawie graficznej reprezentacji można odkryć występujące w sekwencji nukleotydowej wzorce. Spójrzmy na kilka przykładów. Dla sekwencji losowej obserwujemy jednorodny rozkład punków.



Rysunek 6. CGR dla losowej sekwencji nukleotydowej o długości 2000 nt. Obserwujemy równomierny rozkład punktów.

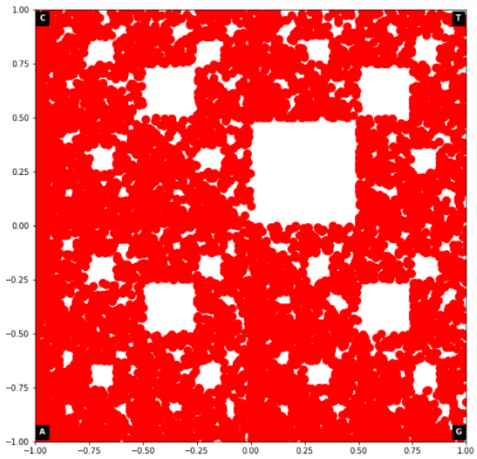
W przypadku nieobecności jednego z nukleotydów w sekwencji pojawia się charakterystyczny wzór fraktalny.

25:6118883539

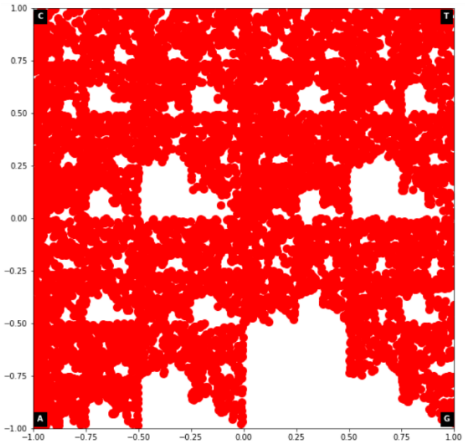


Rysunek 7. CGR dla losowej sekwencji nukleotydowej o długości 2000 nt. która nie zawiera C.

Różnie przedstawiają się również reprezentacje w przypadku nieobecności dinukleotydów z przekątnej (np. AT) oraz tych sąsiednich (np. AG).



Rysunek 8. CGR dla losowej sekwencji nukleotydowej o długości 10000 nt. która nie zawiera AT.



Rysunek 9. CGR dla losowej sekwencji nukleotydowej o długości 10000 nt. która nie zawiera AG.

W każdym z przypadków obserwujemy charakterystyczną nieobecność punktów w pewnych polach w obrębie kwadratu. Formanie tworzą one wzór fraktalny.

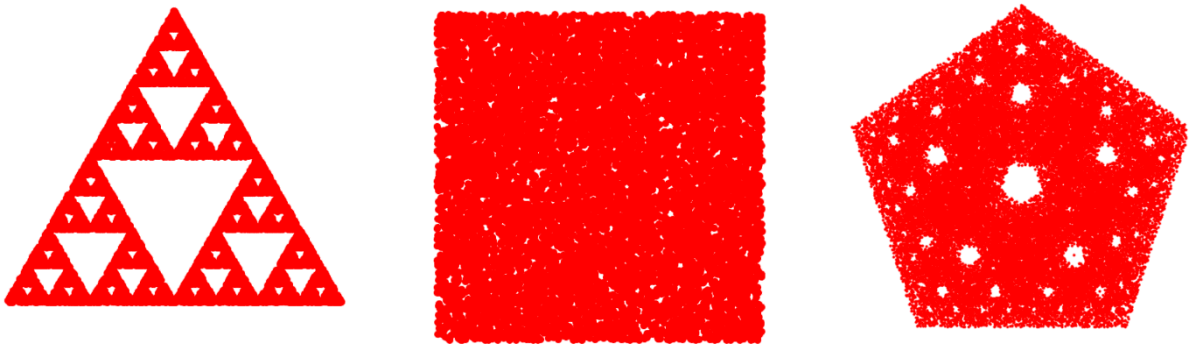
26:8079606459

Analogiczna reprezentacja może być wygenerowana w przypadku sekwencji RNA z tym, że tymina (T) jest zastąpiona poprzez uracyl (U). W dalszej części pracy autor przedstawia też zmodyfikowaną wersję CGR dla sekwencji RNA z uwzględnieniem jej struktury drugorzędowej. W literaturze proponowane były również reprezentacje dla sekwencji białkowych. Jedno z rozwiązań proponuje podział 20 aminokwasów na cztery grupy ze względu na ich własności fizykochemiczne [17]. Takie podejście stanowi proste uogólnienie CGR dla sekwencji nukleotydowych. Główną wadą takiego rozwiązania jest utrata informacji o występujących resztach bowiem następuje tutaj ich redukcja do czterech znaków. Innym, również naturalnym, uogólnieniem CGR na białka jest zastosowanie dwudziestościanu foremnego zamiast kwadratu [18]. W tym przypadku współrzędne aminokwasów dane są przez

$$x_n = r\sin(\frac{2\pi i}{n} + \theta), \tag{20}$$

$$y_n = r\cos\left(\frac{2\pi i}{n} + \theta\right), \tag{21}$$

gdzie n jest liczbą wierzchołków (20) a θ jest kątem orientacji. Okazuje się jednak, że takie podejście prowadzi do obrazów, w których następuje nakładanie się punktów czy też występują naturalnie wolne przestrzenie. Taka sytuacja występuje m.in. już w przypadku trójkąta równobocznego czy pięciokąta foremnego.



Rysunek 10. Reprezentacje gry chaosu dla losowej sekwencji z alfabetu 3, 4 oraz 5 literowego. Tylko w przypadku 4 liter losowa sekwencja wypełnia jednorodnie wnętrze wielokąta.

Żeby przezwyciężyć tę kwestie stosuje się różnego typu czynniki skalujące. Lochel w 2020 [18] zaproponował następującą formułę na taki czynnik (ang. *scaling factor*)

$$sf = 1 - \frac{\sin(\frac{\pi}{n})}{\sin(\frac{\pi}{n}) + \sin(\frac{\pi}{n} + \frac{2\pi m}{n})}, \tag{22}$$

$$m = \left[\frac{n}{4}\right]. \tag{23}$$

Kolejne zagadnienie dotyczy przyporządkowania kolejnych aminokwasów do wierzchołków dwudziestościanu foremnego. Istnieje bowiem $\frac{(n-1)!}{2}$ możliwości, co w przypadku $n = 20$ prowadzi do ogromnej liczby. Autor zaproponował w dalszej części inne podejście. Z jednej strony zachowujące 20 wierzchołków, a z drugiej uwzględniające fizykochemiczne właściwości aminokwasów. Niweluje to m.in. problem arbitralności przy przypisaniu wierzchołków.

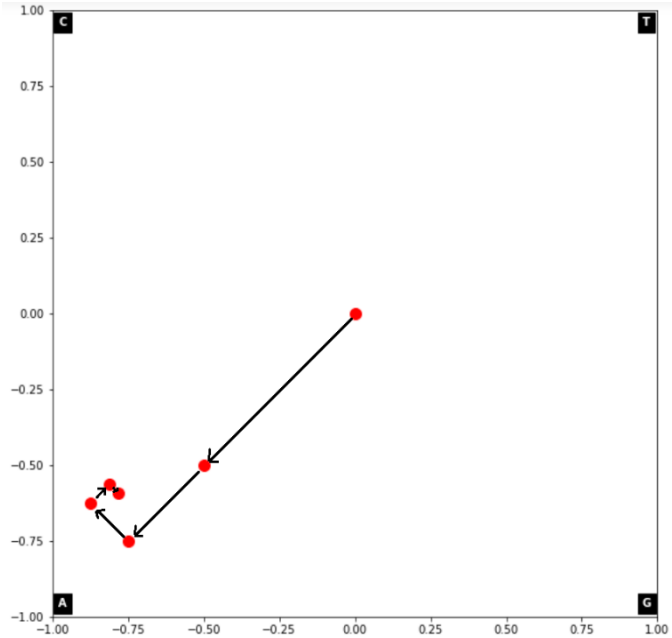
27:1233753712

3.6.6. Genomacierz

Genomacierz, podobnie jak reprezentacja gry chaosu stanowi dwuwymiarową reprezentację sekwencji [79]. Tutaj także nukleotydy umieszczane są w wierzchołkach kwadratu jednak sama transformacja jest nieco odmienna

$$x_n = x_{n-1} + \frac{\vec{N}}{2^{n-1}}, \tag{24}$$

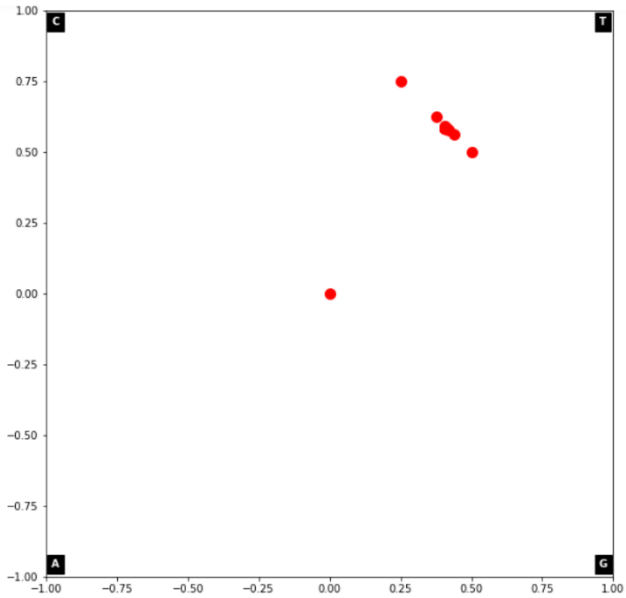
gdzie $\vec{N}$ wskazuje na aktualny nukleotyd na n-tej pozycji, $\vec{A} = [-0.5; -0.5], \vec{C} = [-0.5; 0.5], \vec{T} = [0.5; 0.5], \vec{G} = [0.5; -0.5]$ a $x_0 = (0,0)$.



Rysunek 11. Genomacierz wygenerowana dla sekwencji AACTG.

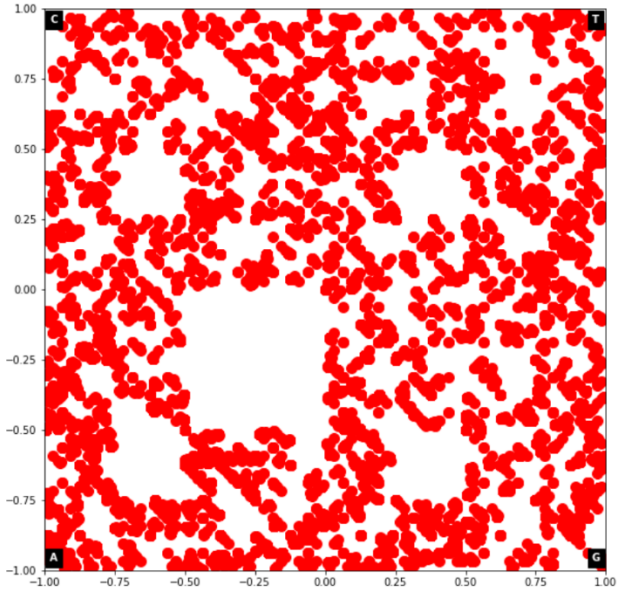
W przypadku CGR, dla losowej sekwencji zapełniona była cała powierzchnia kwadratu (rozkład jednostajny). W tym przypadku pierwsza litera występująca w sekwencji determinuje pierwszą ćwiartkę kwadratu dla pozostałych punktów, druga kolejną podćwiartkę itd. W efekcie również punkty znajdują się coraz bliżej siebie – taka macierz jest rzadka dla jednej sekwencji.

28:7317592464



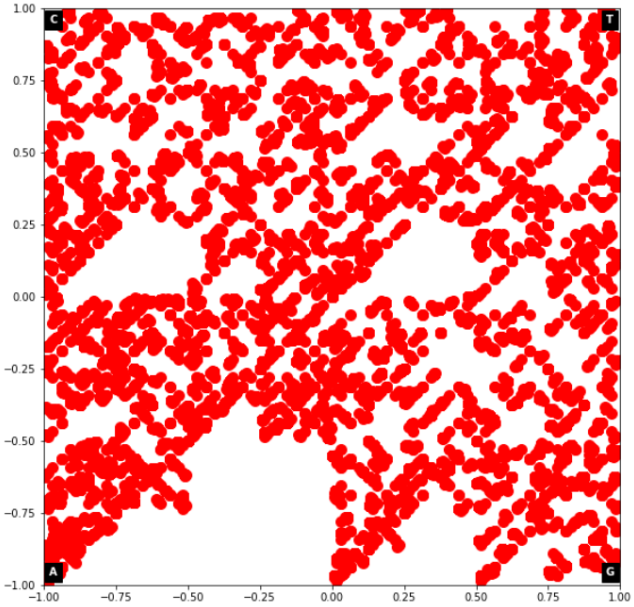
Rysunek 12. Genomacierz wygenerowana dla losowej sekwencji o długości 2000 nt. Kolejne punkty są na tyle blisko, że różnice przestają być widoczne.

Z tego powodu genomacierz można z powodzeniem zastosować do jednoczesnego zaprezentowania zbioru sekwencji. Z uwagi na wrażliwość kolejnych pozycji w sekwencji, za pomocą takiej reprezentacji również można prześledzić występujące wzorce. Podobnie jak w przypadku CGR nieobecność pewnych podsekwencji odzwierciedla się w niezapełnionych polach.



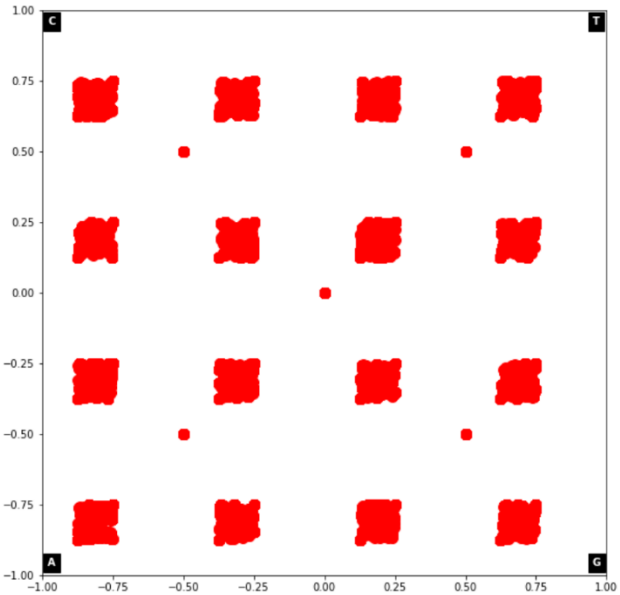
Rysunek 13. Genomacierz przedstawiająca 1000 losowych sekwencji o długości 100 nt. każda, gdzie nie występuje podsekwencja AT.

29:8054982321



Rysunek 14. Genomacierz przedstawiająca 1000 losowych sekwencji o długości 100 nt. każda, gdzie nie występuje podsekwencja AG.

Możemy też prześledzić jaki wpływ na reprezentację genomacierzy ma występowanie danej podsekwencji na danej pozycji, co nie było takie proste w przypadku CGR.



Rysunek 15. Genomacierz przedstawiająca 1000 losowych sekwencji o długości 100 nt. każda, gdzie AT występuje na trzeciej pozycji.

Oczywiście rzeczywiste przykłady sekwencji biologicznych są niejednorodne, a tym samym obserwuje się bardziej złożone wzorce. Jednak znajomość relacji obrazów zarówno z genomacierzy jak i reprezentacji gry chaosu do sekwencji wzorcowych, pozwala na ich łatwiejszą interpretację.

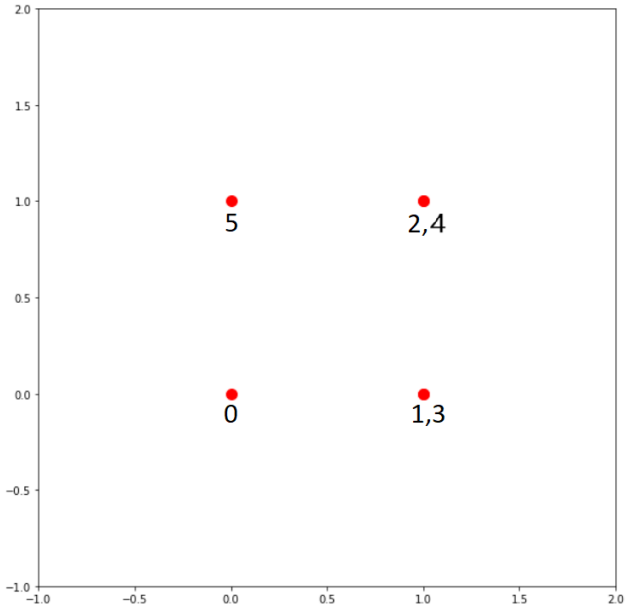
30:4362947533

3.6.7. Spacer DNA

Spacer DNA stanowi jeszcze jeden przykład reprezentacji sekwencji nukleotydowej na płaszczyźnie. Rozważany był m.in. przy badaniu zmian genetycznych występujących w nowotworze płuc [32]. Podobnie jak w reprezentacji gry chaosu, zaczynamy od początku układu współrzędnych $x_0 = (0,0)$. Jednak w tym przypadku w zależności od kolejnego nukleotydu występującego w sekwencji poruszamy się w górę (A), dół (T), prawo (G), lewo (C) o wektor jednostkowy. Oczywiście ponownie możliwe jest inne usytuowanie nukleotydów. W ten sposób otrzymujemy ciąg punktów, których wartości nie są ograniczone. Formalnie stosujemy transformację

$$x_n = x_{n-1} + \overrightarrow{N}. \tag{25}$$

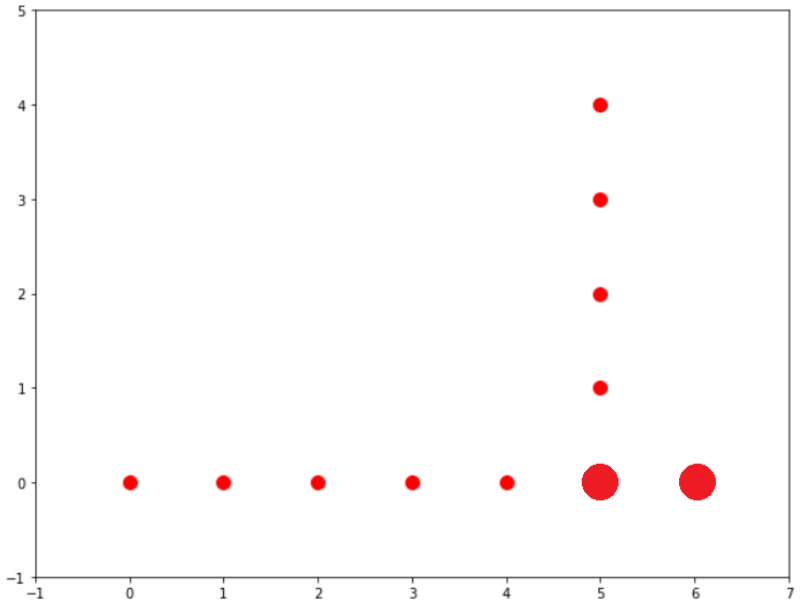
Gdzie $x_0 = (0,0)$ a N jest $A = [0,1], T[0,-1], G[1,0]$ lub $C[-1,0]$. Zauważmy, że w takiej reprezentacji mogą pojawić się punkty zajmowane wielokrotnie przez kolejne reszty. Przykładem może być sekwencja GATAC, gdzie zaczynamy od kierunku w prawo, następnie w górę, dół i ponownie górę oraz lewo. W przypadku reprezentacji graficznej można zastosować odpowiednio większe punkty dla miejsc zajmowanych wielokrotnie. Oczywiście w przypadku analizy ciągu punktów, te są zapisywane dla każdej pozycji. Czasem redukuje się pełny ciąg do wartości unikalnych, a tym przypisuje się odpowiednie wagi (np. zliczenia). Poniżej przedstawiono graficznie sposób konstrukcji tej reprezentacji dla przykładowej sekwencji.



Rysunek 16. Spacer DNA dla sekwencji GATAC. Kolejnymi liczbami naturalnymi zaznaczono punkty odpowiadające kolejnym resztom w sekwencji (pozycja). Zauważmy, że dwa z położeń zajmowane są dwukrotnie.

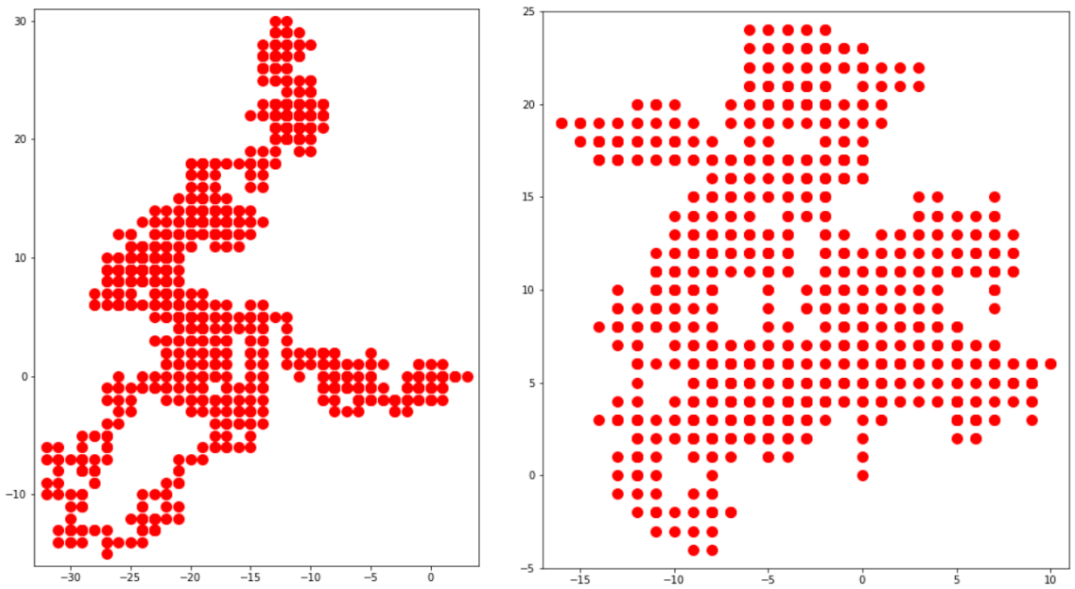
W tym przypadku również mamy obszerną informację dotyczącą sekwencji, w tym jej dalekozasięgowych zależności.

31:4124256468



Rysunek 17. Spacer DNA dla sekwencji GGGGGGCGCGCAAAA. Dla przejrzystości rysunku nie zaznaczono tym razem numeracji kolejnych punktów. Punkty zajmowane wielokrotnie zwiększono.

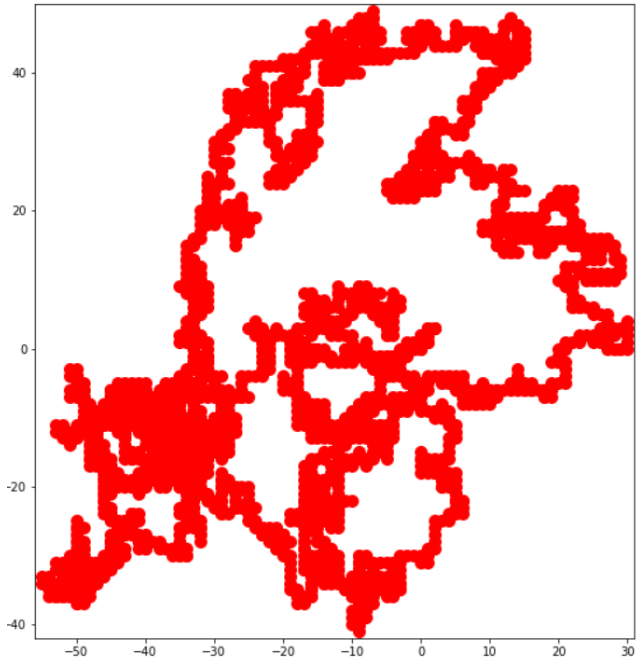
Na powyższym rysunku widać ciąg punktów w prawo wskazujących na powtórzenia G oraz ciąg punktów do góry z powodu wystąpień A. Dwa zwiększone punkty wskazują na lokalne wystąpienie kilku powtórzeń GC. Warto tutaj zaznaczyć istotnie różne zachowanie sekwencji losowych od sekwencji biologicznych.



Rysunek 18. Spacer DNA dla dwóch losowych sekwencji nukleotydowych o długości 1000 każda.

Jak łatwo zauważyć, wartości są w stosunkowo małym przedziale w porównaniu z długością sekwencji. Dodatkowo punkty stanowią gęsto upakowaną reprezentację. Dla porównania, dla sekwencji ludzkiego genu SPARC (długość 3448) obserwujemy następujący rozkład punktów

32:6892425558



Rysunek 19. Spacer DNA dla mRNA genu SPARC (NCBI ID: NM_001309443.2).

W tym przypadku widać nieco więcej nieregularności aniżeli poprzedni losowy spacer. Obserwujemy m.in. pewne lokalne utrzymujące się motywy występujące w sekwencji (np. powtórzenia). Związane jest to oczywiście z niejednostajnym rozkładem nukleotydów w obrębie tego genu.

Spacer DNA może być rozważany w wyżej wymiarowych przestrzeniach. Przykładem może być przestrzeń czterowymiarowa, gdzie kolejne kierunki osi wskazują na dany nukleotyd. Takie podejście eliminuje problem nakładania się punktów, ponieważ nie ma tutaj możliwości powrotu – ścieżkę w przeciwnym kierunku do aktualnego. Wprowadzoną reprezentację dla DNA można też w sposób naturalny uogólnić dla aminokwasów [98]. W tym przypadku rozważana jest przestrzeń 20 wymiarowa, a aminokwasy określają kolejne kierunki. Należy jednak pamiętać, że wraz ze wzrostem wymiaru zyskujemy znaczną ilość dodatkowego miejsca, a same punkty są średnio dalej oddalone od siebie. Pojawia się też problem z wizualizacją wyników, jednak ten można częściowo rozwiązać poprzez prezentację profili, czyli rzutów na przestrzenie niżej wymiarowej, najczęściej 2D. W przypadku sekwencji aminokwasowych byłyby to ustalone dwa aminokwasy.

Na koniec warto jeszcze wspomnieć, że do stworzenia spaceru DNA można wykorzystać też inne reprezentacje. Jednym z rozwiązań jest wykorzystanie współrzędnych z reprezentacji gry chaosu. Zamiast traktować je jako niezależne można policzyć skumulowaną sumę poprzednich punktów. Takie rozwiązanie częściowo rozwiązuje zagadnienie częstego nakładania się punktów w tradycyjnym spacerze DNA (szczególnie gdy analiza opiera się o reprezentację graficzną). To podejście zostanie w dalszej części zastosowane przez autora do analizy genów niezbędnych u bakterii.

33:7243955433

3.6.8. PseAAC

Poprzednie reprezentacje były dedykowane głównie sekwencjom nukleotydowym, jednak ich modyfikacje można było też zastosować dla sekwencji białkowych. W tym podrozdziale wprowadzimy jeszcze inny sposób opisu sekwencji, tym razem dedykowany tym drugim. Metoda PseAAC (ang. *Pseudo Amino Acid Composition*) uwzględnia zarówno korelacje występujące w sekwencjach jak i fizykochemiczne właściwości aminokwasów. Wprowadzona została w 2001 roku przez Kuo-Chen Chou [99]. Jej zastosowania obejmują między innymi ułatwienie w rozpoznaniu lokalizacji komórkowej białek czy przewidywanie typu białek błonowych [100,101,102]. Podobnie jak w metodzie częstościowej (AAC) zliczane są wystąpienia aminokwasów w sekwencji. Ponadto uwzględniane są pewne korelacje w obrębie sekwencji występujące w zadanym oknie. Aby wyznaczyć reprezentację PseAAC ustalamy parametr λ określający rząd tych korelacji (mniejszy od długości sekwencji L), a następnie wyznaczamy następujący ciąg liczbowy

$$p_1, p_2, \dots, p_{20}, p_{20+1}, \dots p_{20+\lambda} \tag{26}$$

zgodnie z następującymi formułami:

- dla $i = 1, \dots, 20$

$$p_i = \frac{f_i}{\sum_{k=1}^{20} f_k + w \sum_{k=1}^{\lambda} \tau_k}, \tag{27}$$

- dla $i = 20 + 1, \dots, 20 + \lambda$

$$p_i = \frac{w\tau_{i-20}}{\sum_{k=1}^{20} f_k + w \sum_{k=1}^{\lambda} \tau_k}, \tag{28}$$

gdzie w jest również ustalonym parametrem (wagą) a τ_k są parametrami opisującymi występujące w sekwencji korelacji. Obliczane są jako

$$\tau_k = \frac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+k}, \tag{29}$$

gdzie

$$J_{i,i+k} = \frac{1}{\Gamma} \sum_{q=1}^{\Gamma} [\Phi_q(X_{i+k}) - \Phi_q(X_i)]^2 \tag{30}$$

gdzie $\Phi_q(X_i)$ jest wartością q-tej funkcji, określającą wartość aminokwasu X_i, $q = 1, \dots, \Gamma$. W praktycznych przypadkach często rozważanymi właściwościami są hydrofobowość, hydrofilowość oraz masa łańcucha bocznego ($\Gamma = 3$). Oczywiście w ogólnym przypadku można uwzględnić też inne cechy takie jak punkt izoelektryczny (pI) czy stała dysocjacji (pK). Upraszczając nieco interpretację, wartości $p_1, p_2, \dots, p_{20}$ związane są przede wszystkim z częstościami występowania każdego z aminokwasów (odpowiednio znormalizowane). Z kolei kolejne człony $p_{20+1}, \dots p_{20+\lambda}$ charakteryzują dalekozasięgowe korelacje występujące pomiędzy

34:3125423736

resztami w sekwencji. Te z kolei opierają się na funkcji $J_{i,i+k}$ wyznaczającej średni kwadrat różnic pomiędzy własnościami aminokwasów występującymi w sekwencji w odległości k. Aby lepiej zrozumieć te korelacje przedstawiono poniżej następujący schemat pokazujący zakres $J_{i,i+k}$. Dla sekwencji

$$X_1 X_2 X_3 \dots X_L$$

$\tau_1$zbiera informację dotyczącą zależności pierwszego rzędu, czyli pomiędzy $X_1\ i\ X_2, X_2\ i\ X_3\ \dots\ X_{L-1}\ i\ X_L$. Podobnie τ_2 odnosi się do drugiego rzędu, czyli bada zależność pomiędzy $X_1\ i\ X_3, X_2\ i\ X_4\ \dots\ X_{L-2}\ i\ X_L$. Maksymalny indeks τ_k określony jest za pomocą λ. To uzasadnia konieczność wyboru parametru lambda mniejszego niż długość sekwencji. W szczególności dla $\lambda = L - 1$ badamy oddziaływanie pomiędzy X_1 oraz X_L.

Dla przykładu, dla białka SPARC (NCBI ID: CAG33080.1) wybierając właściwości: hydrofobowość oraz hydrofilowość oraz paramenty $\lambda = 3$ oraz $w = 0.05$ otrzymujemy wyniki zaprezentowane w Tabeli 3.

Tabela 3. Reprezentacja PseAAC dla ludzkiego białka SPARC. Wyniki wygenerowano za pomocą narzędzia dostępnego online (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/).

4.560	3.800	5.827	8.614	3.547	3.547	3.040	3.040	5.067	7.347
1.267	3.547	4.560	2.533	2.533	1.773	4.307	5.067	1.013	1.773
7.909	7.883	7.446							

Pierwsze 20 wartości związane są z kolejnymi częstościami, a trzy ostatnie liczby związane są z korelacjami 1., 2. oraz 3. rzędu mierzonymi zadanymi własnościami fizykochemicznymi.

Omówione zostaną teraz techniki analizy szeregów (DFT oraz wykładnik Hursta) oraz obrazów (w przypadku metody SSIM) w połączeniu z metodą CGR zostały one zastosowane do porównywania sekwencji biologicznych.

3.7. Dyskretna Transformata Fouriera (DFT)

Mając dany ciąg wartości liczbowych $X = \{X_0, X_2, .. X_{N-1}\}$ (w ogólności zespolonych) można przekształcić go na $Y = \{Y_0, Y_2, .. Y_{N-1}\}$ zgodnie z formułą

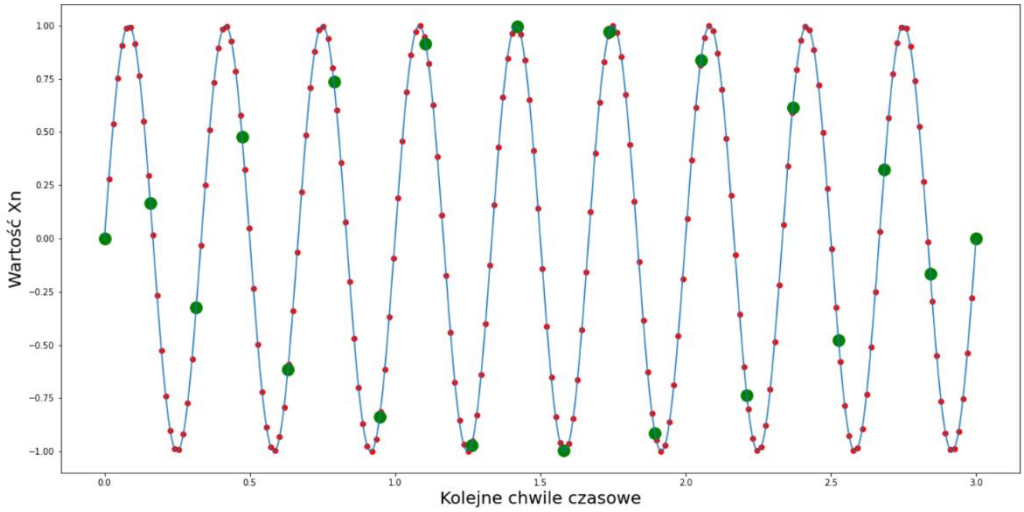
$$Y_k = \sum_{n=0}^{N-1} X_n e^{-\frac{2\pi i k n}{N}}, \tag{31}$$

gdzie i jest jednostką urojoną, a $k = 0,1 \dots, N-1$. Zauważmy, że do wyznaczenia każdej wartości Y_k wykorzystywane są wszystkie wartości ciągu X. Ponadto ciąg Y jest ciągiem liczb zespolonych. Taka transformacja danych jest bardzo często używana przy analizie sygnałów różnego typu do uwidocznienia pewnych cech tych sygnałów [103]. Dane przechodzą z domeny czasu na domenę częstotliwości. Aby uzyskać wartości rzeczywiste można policzyć tzw. widmo (ang. *power spectrum*) zgodnie z

$$PS_k = |Y_k|^2. \tag{32}$$

35:8114545516

W przypadku prostych sygnałów będących kombinacjami funkcji trygonometrycznych, na widmie można dostrzec charakterystyczne piki. Należy tutaj zwrócić uwagę na zagadnienie próbkowania, a dokładniej kryterium Nyquista które stwierdza w jakich odstępach czasu należy zbierać sygnały, aby potrafić z nich następnie wyznaczyć charakterystyczne częstości.

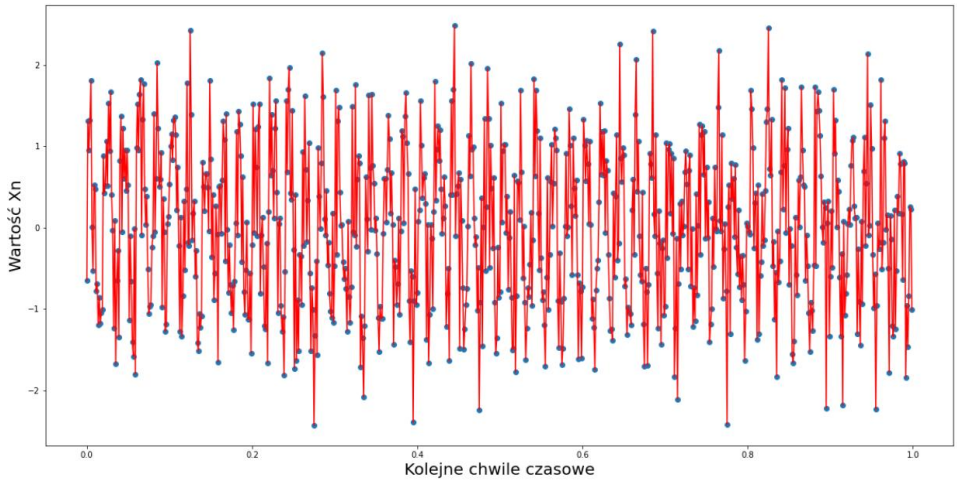


Rysunek 20. Wygenerowano dwa ciągi liczbowe według tej samej formuły – 200 wartości (czerwone punkty) oraz 20 wartości (zielone punkty). Zbyt małe próbkowanie może spowodować brak możliwości rekonstrukcji oryginalnego ciągu czy jego własności.

Rozważmy następujący przykład. Wygenerowano $N = 800$ wartości wg zależności

$$f(t) = \sin(50 * 2\pi t) + \frac{1}{2}\sin(250 * 2\pi t) + \varepsilon_t, \tag{33}$$

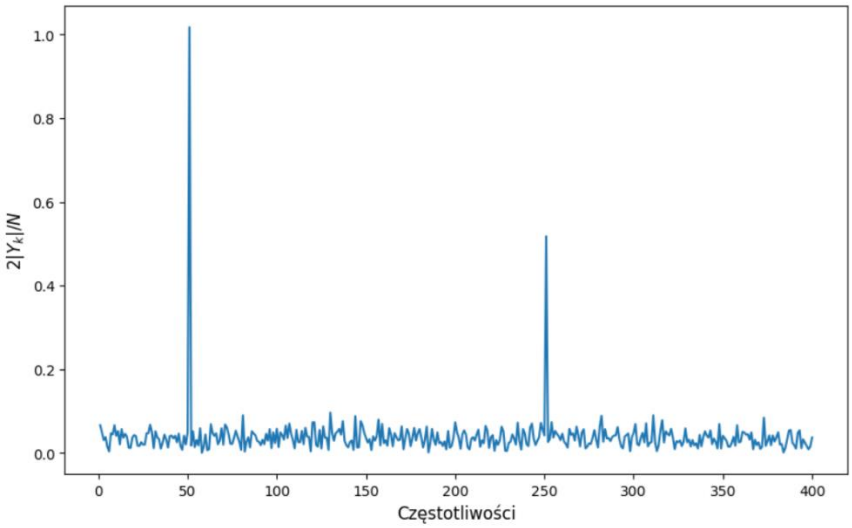
z przedziału czasu od 0 do 1 z krokiem T=1/800, gdzie ε_t jest pewną liczbą losową z rozkładu jednostajnego w zakresie od -1 do 1 (dla każdej chwili czasowej inna). Narysujmy wykres kolejnych obserwacji



Rysunek 21. Wartości ciągu x_t określonego zależnością $sin(50 * 2\pi t) + \frac{1}{2}sin(250 * 2\pi t) + \varepsilon_t$, gdzie rozważano czas z przedziału [0,1].

36:8301846534

Odpowiadający mu wykres znormalizowanych modułów z wartości ciągu Y_k prezentuje charakterystyczne częstości. Obserwujemy dwa piki przy 50 i 250. Wykres narysowano tylko dla połowy wartości Y_k z powodu symetrii.



Rysunek 22. Wykres znormalizowanych modułów z wartości ciągu Y_k. Pojawiają się dwie charakterystyczne częstości przy 50 oraz 250. Wykres przycięto w połowie z powodu symetrii.

Takie piki występują także w sekwencjach biologicznych. Okazuje się, że np. dla oksydazy cytochromowej puchacza zwyczajnego (NCBI ID: GU571285) pojawia się charakterystyczny pik w okolicy 1/3 długości. Wyniki tego typu łączy się z informacją dotyczącą trójkowego kodu genetycznego [104]. Aby porównać ze sobą sygnały różnej długości można użyć ich widm. Te jednak powinny być tej samej długości. Yin oraz Yau zaproponowali transformację, dzięki której można wydłużyć krótszą sekwencję do dłuższej [105]. Jeżeli przez n oraz m oznaczamy długości dwóch sekwencji oraz $m > n$ to ciąg T_n przekształcamy do ciągu T_m wg wzoru

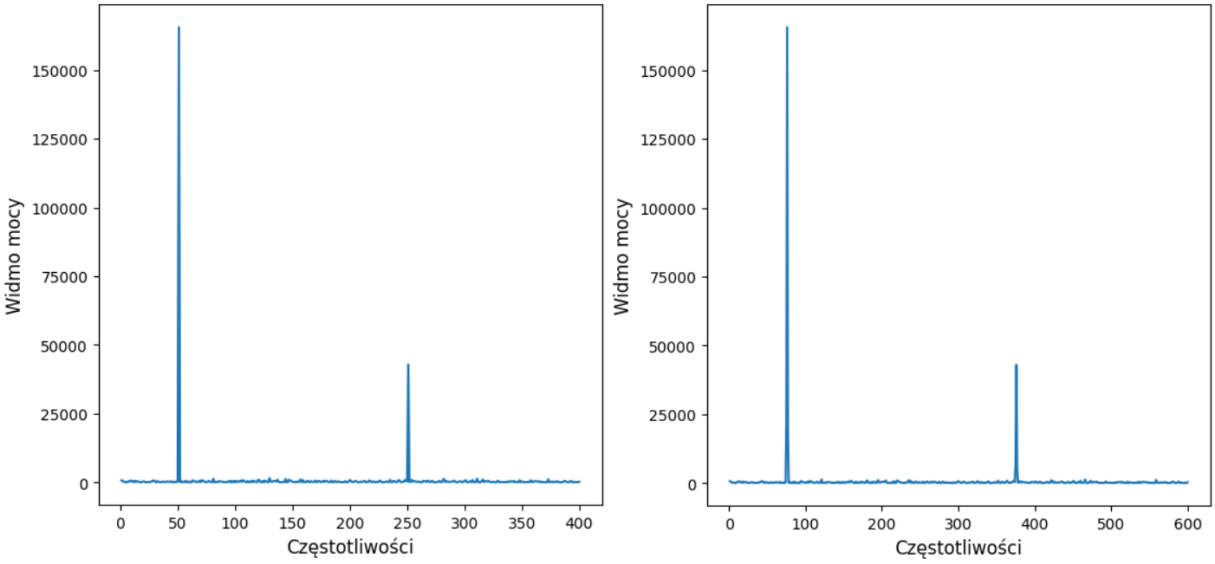
$$T_m(k) = T_n(Q), \tag{34}$$

jeżeli $Q = \frac{kn}{m} \in Z$, w przeciwnym wypadku

$$T_m(k) = T_n(R) + (Q - R)\big(T_n(R+1) - T_n(R)\big), \tag{35}$$

dla Q zdefiniowanego jak wyżej oraz R będącego zaokrągleniem Q w dół do najbliższej liczby całkowitej.

37:9655487140



Rysunek 23. Widma mocy o długości 800 (rysunek z lewej, przycięto w połowie) wydłużono do 1200 (rysunek z prawej, przycięto w połowie). Na pierwszym rysunku pik znajduje się na pozycji 50, czyli w 50/400 = 0.125 części; podobnie na drugim wykresie na 75 pozycji, czyli 75/600= 0.125 części.

3.8. Wykładnik Hursta

Podobnie jak poprzednio, wykładnik (lub współczynnik) Hursta służy analizie ciągu wartości $X = \{X_1, X_2, .. X_N\}$. Wprowadzony został w XX wieku przy analizę poziomu wód w Nilu [106]. W tym przypadku wartości powinny przyjmować wartości rzeczywiste. Wyznaczana jest jedna wartość z przedziału [0,1] charakteryzująca zachowanie wejściowego ciągu. Postępujemy wg następującej procedury

1. Wyznaczamy średnie cząstkowe

$$< X >_n = \sum_{i=1}^{n} X_i \tag{36}$$

2. Wyznaczamy różnice cząstkowe

$$X_j(n) = \sum_{i \le j} (X_i - < X >_n) \tag{37}$$

3. Wyznaczamy różnice pomiędzy maksymalną i minimalną wartości różnic cząstkowych

$$R(n) = max_{j=1,2,\dots,n} X_j(n) - min_{j=1,2,\dots,n} X_j(n) \tag{38}$$

4. Wyznaczamy odchylenie standardowe

$$S(n) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - < X >_n)^2} \tag{39}$$

38:4642952788

dla $n = 1,2, \dots, N$.

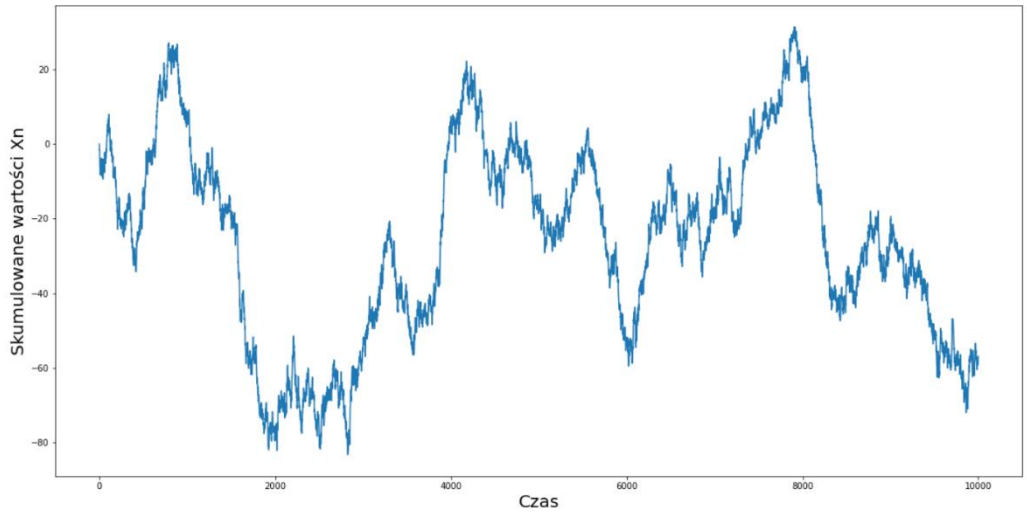
Okazuje się, że przy dostatecznie dużej liczbie obserwacji zachodzi przybliżona zależność

$$\frac{R(n)}{S(n)} \approx Cn^H \tag{40}$$

dla pewnych stałych C oraz H. Tę drugą nazywamy właśnie wykładnikiem Hursta. Aby go efektywnie wyznaczyć, najprościej zlogarytmować powyższą zależność otrzymując liniową formułę pomiędzy $\log\left(\frac{R(n)}{S(n)}\right)$ oraz $\log(n)$

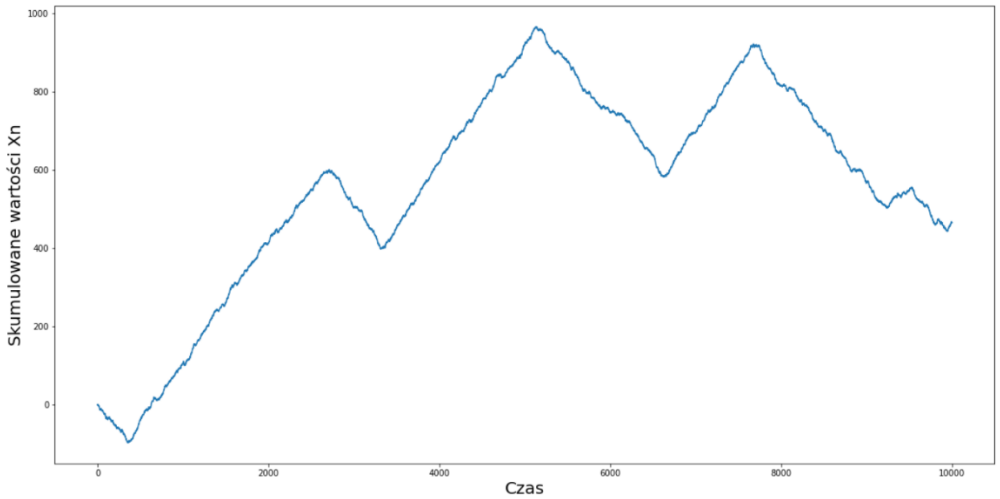
$$\log\left(\frac{R(n)}{S(n)}\right) = H\log(n) + C_1, \tag{41}$$

gdzie $C_1 = \log(C)$. Wartości tego współczynnika bliskie 0.5 świadczą o wysokim stopniu losowości i nieprzewidywalności kolejnych wyrazów ciągu. Jeżeli współczynnik Hursta przyjmuje wartość poniżej 0.5 obserwowane jest zjawisko odwracania trendu i powracania do średniej (antypersystencja). Obserwujemy tutaj dużą zmienność. Przykładowo, jeżeli wartości malały, to prawdopodobnie tendencja ta zostanie odwrócona w przyszłości. Z kolei, jeżeli wartości współczynnika są powyżej 0.5 wskazuje, że istnieją pewne dłuższe tendencje i wzmocnienie trendu oraz tzw. efekt pamięci (persystencja). W praktyce często wyznaczane są skumulowane wartości ciągu X. Spójrzmy na kilka przykładów ciągów wraz z odpowiednimi współczynnikami Hursta.

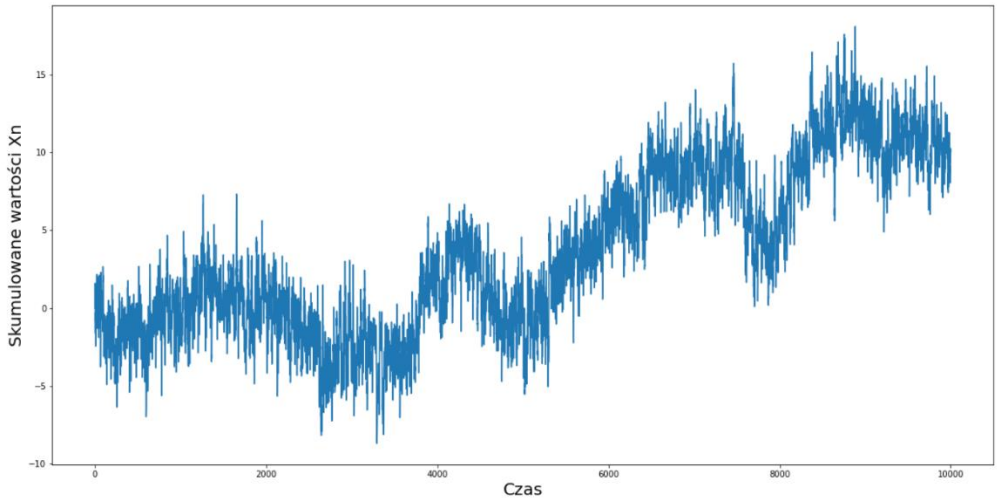


Rysunek 24. Przykład szeregu losowego, dla którego współczynnik Hursta wynosi 0.47. Następują tutaj losowe zmiany w kolejnych chwilach czasowych.

39:8726769472



Rysunek 25. Przykład szeregu losowego dla którego współczynnik Hursta wynosi 0.82 (tzw. szereg persystentny). Pewne trendy utrzymują się przez określony okres.



Rysunek 26. Przykład szeregu losowego dla którego współczynnik Hursta wynosi 0.26 (tzw. szereg antypersystentny). Obserwujemy tutaj nagłe zmiany i odwracanie trendu.

W porównaniu z dyskretną transformatą Fouriera, w tej metodzie analizy szeregów czasowych, otrzymujemy ostatecznie jedną liczbę. Porównanie pomiędzy różnymi szeregami polega na porównaniu ich współczynników Hursta. Redukuje to problem różnej długości rozważanych szeregów. Oczywiście sprowadzenie zachowania pełnego zestawu wartości do jednej końcowej wartości niesie ze sobą kilka problemów. Pojawiają się pytania dotyczące czułości tego współczynnika na zmianę wartości szeregu oraz na ile różne od siebie szeregi mogą osiągać ostatecznie podobną jego wartość.

Zarówno współczynnik Hursta jak i dyskretna transformata Fouriera rozważane są typowo dla szeregów czasowych. Te same metody można dostosować także do analizy szeregów przestrzennych, tzn. ciągi wartości nie obserwowane w czasie a kolejnych położeniach jak np. w przypadku sekwencji nukleotydowych. Takie podejście będzie stosowane w dalszej części.

40:7673425102

3.9. Ilościowa analiza rekurencji (RQA)

Ilościowa analiza rekurencji (RQA) stanowi kolejną metodę analizy szeregów czasowych [107]. Z powodzeniem wielokrotnie była stosowana do badania do systemów nieliniowych i chaotycznych [108,109]. Szczególną zaletą tej metody jest możliwość badania krótkich, niestacjonarnych danych. Bazuje na macierzy rekurencji (ang. *Recurrence Plot,* RP). Wartości tej macierzy definiujemy przez $R_{ij} = H\big(\epsilon - |x_i - x_j|\big)$, gdzie x_i, x_j są punktami w przestrzeni fazowej w chwilach i, j odpowiednio, ϵ to parametry określający maksymalną odległość między punktami, | | to metryka, a H to funkcja Heaviside'a.

$$H(x) = \begin{cases} 0, x < 0 \\ 1, x \geq 0 \end{cases}. \tag{42}$$

W praktyce przestrzeń fazowa często rekonstruowana jest na podstawie wartości szeregu czasowego u. Dokładniej, $x_i = [u(i), u(i+\tau), \dots, u\big(i + \tau(m-1)\big)]$, gdzie m to wymiar tej przestrzeni (tzw. wymiar zanurzenia, ang. *embedding dimension*), a τ opóźnienie czasowe; obydwa parametry zadawane z góry. Na jej podstawie obliczane są pewne miary określające własności tej macierzy takie jak: częstość rekurencji (RR), długość diagonalnych linii, długość pionowych linii, średnia długość diagonalnych linii itd.

Dla dwóch szeregów u oraz w generujemy zestaw punktów x_i oraz y_i odpowiednio i wyznaczamy macierz rekurencji krzyżowej (ang. *Cross Recurrence Plot*)

$$CR_{ij} = H\big(\epsilon - \big|x_i - y_j\big|\big) \tag{43}$$

i w ten sposób możemy je ze sobą porównać używając analogicznych miar [110].

Zgodnie z sugestią jednego z recenzentów przy pisaniu pracy [111], sprawdzono użyteczność tej metody w połączniu z reprezentacją gry chaosu do tworzenia drzew filogenetycznych m.in. w przypadku genu neuraminidazy u wirusów grypy. Wyniki jednak nie potwierdziły użyteczności takiej kombinacji w porównaniu do omawianej później dyskretnej transformaty Fouriera. W przyszłych badaniach być może warto jednak będzie wrócić jeszcze do tego podejścia; należałoby wtedy dostosować tę technikę do danych oraz ich specyfiki (jak np. naturalnie ograniczony przedział wartości i duża zmienność).

3.10. Momenty bezwładności

Mając do dyspozycji spacer DNA, będący reprezentacją sekwencji nukleotydowej, można określić jego geometryczne właściwości takie jak kształt czy rozkład punktów. Interesujące podejście, zaczerpnięte z klasycznej mechaniki, zaproponowali Państwo prof. Piotr i Dorota Wąż [112]. Idea opiera się na traktowaniu kolejnych punktów jako punkty materialne i wyznaczeniu odpowiednich momentów bezwładności. Zostało pokazane, że takie podejście może być z powodzeniem stosowane w analizie sekwencji, a przykładowe prace dotyczą analizy ewolucji wirusa Zika czy metod bez dopasowania [112,113]. Te ostatnie mogą również dotyczyć sekwencji białkowych, gdzie te reprezentowane są w przestrzeni 20 wymiarowej. Na początku punkty przesuwamy w ten sposób, aby środek ciężkości znajdował się w początku układu współrzędnych, czyli punkcie (0,0). W tym celu wystarczy wyjściowy ciąg punktów $Z_i = (X_i, Y_i)_{i=1,\dots,n}$ przekształcić na $Z_i' = (X'_i, Y'_i)_{i=1,\dots,n}$ według

41:4383203199

$$Z_i' = Z_i - \bar{Z}_i \tag{44}$$

gdzie $\bar{Z}_i$ jest średnią wartością po wszystkich punktach. Następnie, wyznaczany jest tensor momentów bezwładności

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \tag{45}$$

gdzie $I_{11} = \sum_i (Y_i')^2$, $I_{22} = \sum_i (X_i')^2$, $I_{21} = I_{12} = -\sum_i Y_i' X_i'$. Wektory własne I_1, I_2 macierzy I, zwane głównymi momentami bezwładności, odzwierciedlają geometryczny rozkład punktów. Wspomniani autorzy wprowadzili też pewne dodatkowe deskryptory oparte o I_1, I_2, w tym $r_1 = \sqrt{\frac{I_1}{n}}$ oraz $r_1 = \sqrt{\frac{I_2}{n}}$, czyli znormalizowane główne momenty bezwładności. W dalszej części pracy będą one dla uproszczenia nazywane momentami bezwładności. Dodatkowo, punktom można przypisać pewne arbitralne masy. Domyślnie zakładamy, że każdy punkt ma masę 1.

3.11. Indeks podobieństwa strukturalnego (SSIM)

Reprezentacje gry chaosu dla dwóch sekwencji można też porównywać traktując je jako obrazy. Jedną z klasycznych metod porównywania ze sobą obrazów jest indeks podobieństwa strukturalnego (SSIM) [34]. Załóżmy, że mamy dwa obrazy X oraz Y niech x_i oraz y_i oznaczają wartości kolejnych pikseli. Definiujemy

- średnie wartości pikseli

$$\mu_x = \frac{1}{N} \sum_i x_i \,, \tag{46}$$

$$\mu_y = \frac{1}{N} \sum_i y_i \,, \tag{47}$$

- odchylenia standardowe

$$\sigma_x = \sqrt{\frac{1}{N-1} \sum_i (x_i - \mu_x)^2} \,, \tag{48}$$

$$\sigma_y = \sqrt{\frac{1}{N-1} \sum_i \left(y_i - \mu_y\right)^2} \tag{49}$$

oraz kowariancję

$$\sigma_{xy} = \frac{1}{N-1} \sum_i (x_i - \mu_x)\left(y_i - \mu_y\right). \tag{50}$$

42:3341980979

Następnie na ich podstawie definiujemy

- luminację

$$I(X,Y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}, \tag{51}$$

- kontrast

$$c(X,Y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \tag{52}$$

- parametr struktury

$$s(X,Y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}, \tag{53}$$

gdzie c_1, c_2, c_3 to pewne stałe stabilizujące przebieg obliczeń. Następnie wyznaczamy

$$SSIM(X,Y) = I(X,Y)^\alpha c(X,Y)^\beta s(X,Y)^\gamma, \tag{54}$$

gdzie α, β, γ to stałe uwzględniające różny udział zdefiniowanych wcześniej komponentów (wagi dla jasności, kontrastu i parametru struktury). Domyślnie przyjmuje się wartości 1 dla tych parametrów. Jeżeli chcemy teraz wyznacz odległość między dwoma obrazami, liczymy

$$d = 1 - SSIM(X,Y). \tag{55}$$

Miara ta jest symetryczna, jednak w ogólności nie jest przechodnia, więc nie spełnia formalnie definicji metryki.

43:1137267711

Rozdział 4. Wyniki i dyskusja

W tym rozdziale przedstawiono problemy i zagadnienia jakie podjął autor w tej rozprawie, w związku z zastosowaniem reprezentacji gry chaosu do analizy sekwencji nukleotydowych oraz aminokwasowych. Często, na początkowych etapach, przeprowadzano dodatkowo standardowe analizy statystyczne rozważanych sekwencji. Poniżej znajduje się lista opublikowanych już prac, w których autor przedstawił znaczną część tych wyników.

K. Sarapata, A. Kania*, ***Revealing miRNAs patterns by employing matrix representations and energy analysis**, J Mol Graph Model.,* 132, 2024.

A. Kania*, ***Harnessing the information theory and chaos game representation for pattern searching among essential and non-essential genes in Bacteria**, J Theor Biol.,* 531, 110917, 2021.

A. Kania*, K. Sarapata, ***The robustness of the chaos representation to mutations and its application in free-alignment method**s*, *Genomics*, 113(3), 1428-1437, 2021.

A. Kania*, K. Sarapata, ***Multifarious aspects of the chaos game representation and its applications in biological sequence analysis,** Comput Biol Med.,* 151(PtA), 2022.

*- autor korespondencyjny

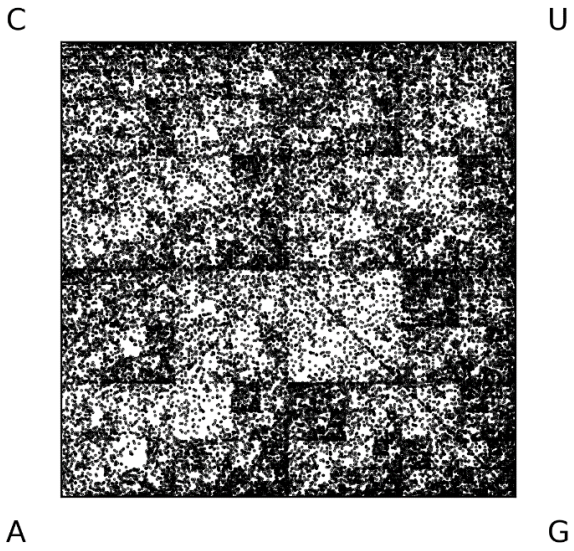
Pierwsza praca dotyczy analizy ludzkich miRNA, gdzie reprezentacja gry chaosu pozwoliła przedstawić, scharakteryzować i porównać motywy występujące w tej grupie sekwencji (rozwinięcie w Rozdział 4.1). W drugiej pracy, autor analizuje pulę genów niezbędnych u bakterii (rozwinięcie w Rozdział 4.2). Dwie ostatnie prace dotyczą metod porównywania sekwencji nukleotydowych oraz aminokwasowych w oparciu o reprezentację gry chaosu (rozwinięcie w Rozdziałach 4.4 oraz 4.5). Przedstawione są też tam pewne rozszerzenia klasycznej reprezentacji i ich biologiczne konsekwencje.

Rysunki pochodzące z prac opatrzone są odnośnikiem do odpowiedniej pozycji. Zgodnie z polityką *Elsevier* (https://www.elsevier.com/about/policies-and-standards/copyright), w celach niekomercyjnych, w szczególności na potrzeby rozprawy doktorskiej, autor pracy ma prawo zamieszczać zarówno wybrane rysunki, tabele jak i treści (wraz z tłumaczeniem).

44:6882806859

4.1. Motywy sekwencyjne w miRNA mogą być identyfikowane na podstawie analizy CGR

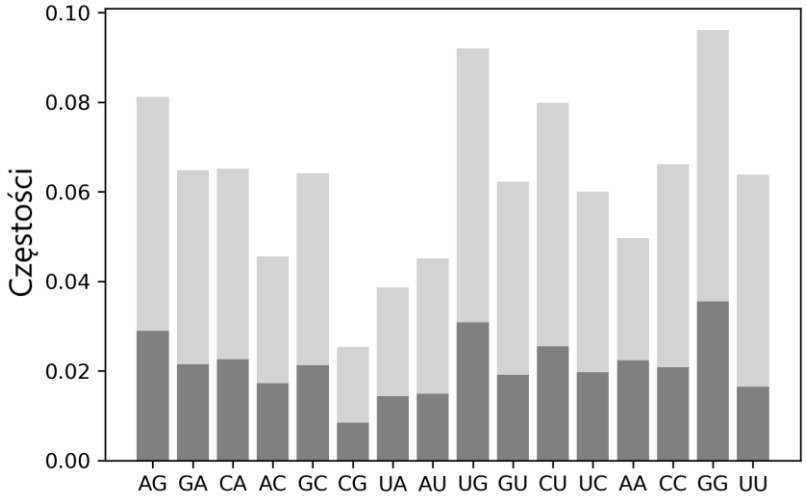
Klasycznym zastosowaniem reprezentacji gry chaosu jest przedstawienie składu sekwencji (częstości występowania kolejnych k-merów) w postaci graficznej. W pracy z 2024, wraz z K. Sarapata badaliśmy ludzkie mikroRNA (miRNA), czyli krótkie jednoniciowe cząsteczki RNA regulujące ekspresję genów [37]. Dane pochodziły z baz danych miRBase 22 [38,39] oraz miRTarBase 8.0 [40]. Przeprowadziliśmy analizę energetyczną oddziaływania miRNA/mRNA za pomocą narzędzia *miRanda 3.3a* [114] oraz analizowaliśmy skład nukleotydowy miRNA. Poniżej przedstawiono część wyników odnoszących się do zastosowania CGR w tym badaniu. Wygenerowano CGR dla wszystkich rozważanych ludzkich miRNA. Taki obraz pokazuje charakterystyczne motywy dzielone między sekwencjami, zlokalizowane potencjalnie na różnych pozycjach.



Rysunek 27. CGR dla wszystkich sekwencji ludzkich miRNA. Taki obraz pozwala na globalną ocenę składu tych cząsteczek. Rysunek opublikowano w J Mol Graph Model. w 2024.

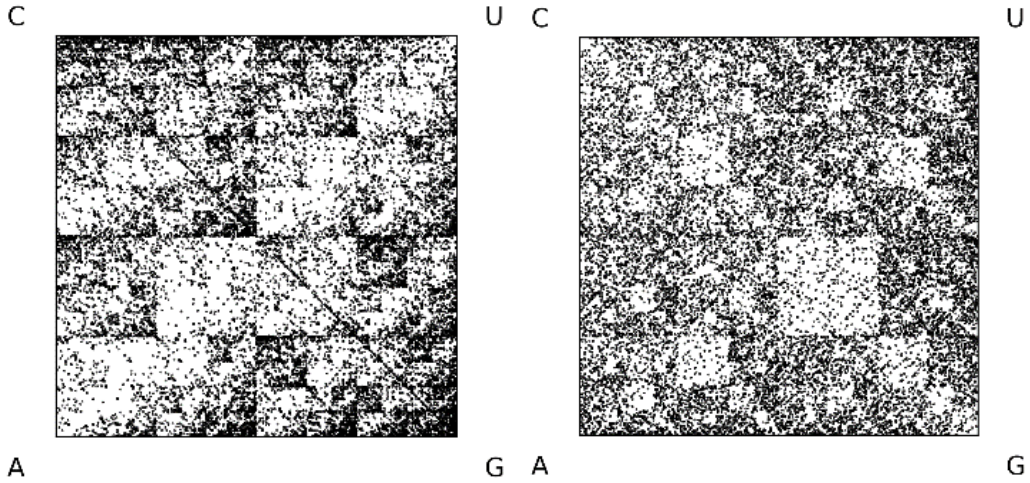
Pojawia się charakterystyczny wzór fraktalny związany z występowaniem w mniejszej ilości określonych k-merów. Jaśniejsze pola w kształcie kwadratów wskazują na istotnie mniej częste występowanie dinukleotydu CG, ale też częściowo AU. Skupiając uwagę na dinukleotydach wygenerowano wykres ich częstości osobno dla pierwszych 7 pozycji (tzw. region *seed*) oraz pozostałych.

45:1039753805



Rysunek 28. Częstość występowania dinukleotydów pośród ludzkich miRNA. Wyróżniono przy tym podział miRNA na tzw. region seed (pierwsze 7 nt., oznaczono ciemniejszym kolorem) oraz pozostałą część (oznaczono jaśniejszym kolorem). Rysunek opublikowano w J Mol Graph Model. w 2024.

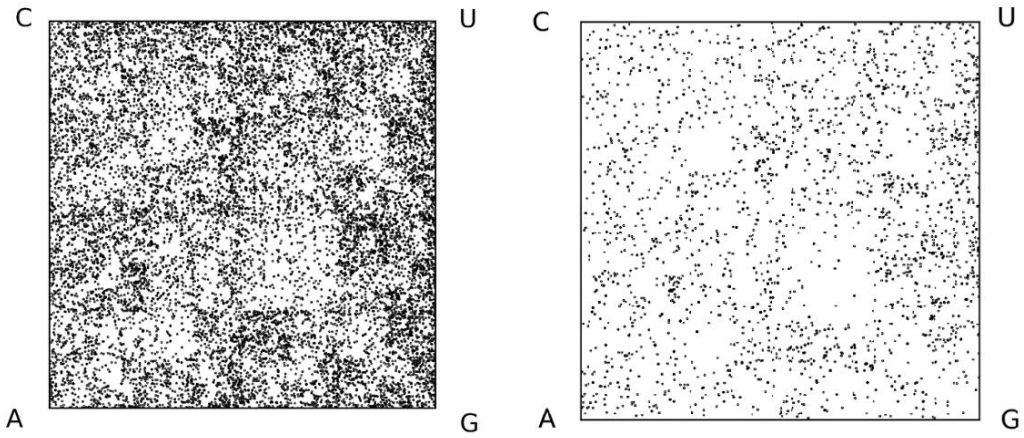
Najczęściej występującymi dinukleotydami są UG oraz GG; potwierdza się również obserwacja, że najmniej częstym jest CG. Dotyczy to obydwu części podziału miRNA. Następnie, przeprowadzono obliczenia energii oddziaływania miRNA z transkryptem. Na tej podstawie podzielono zbiór wszystkich miRNA na nisko- oraz wysokoenergetyczne (ze względu na siłę tego oddziaływania, jako wartość krytyczną wybrano medianę energii).



Rysunek 29. CGR dla wszystkich sekwencji ludzkich miRNA. Te, po analizie energetycznej, podzielono ten zbiór na dwa podzbiory – niskoenergetyczne (po lewej) oraz wysokoenergetyczne (po prawej). Rysunek opublikowano w J Mol Graph Model. w 2024.

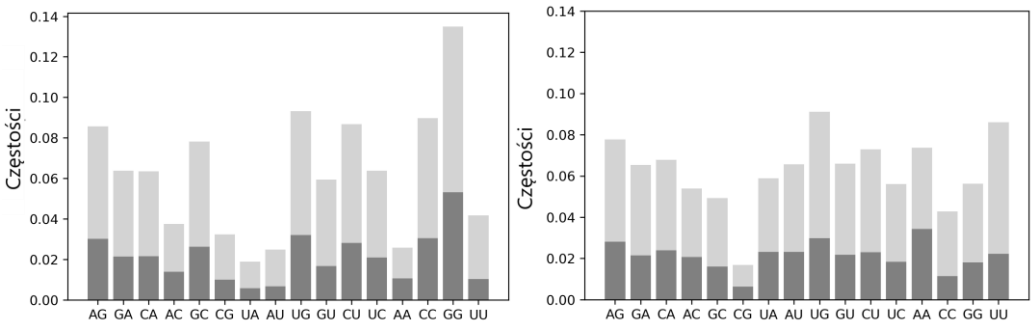
Ponownie obserwujemy charakterystyczne wzorce fraktalne, odmienne dla każdej z grup. W przypadku pierwszej występuje przede wszystkim niedobór dinukleotydów AU oraz UA. Dla drugiej jest to CG. Z baz danych miRBase oraz miRTarBase wybrano dodatkowo podzbiory wszystkich miRNA o największej wiarygodności, w których m.in. istnieją eksperymentalne dowody ich istnienia oraz wysokiej konserwatywności (tzw. rekordy *confident* i *conserved*). Zwizualizowano je za pomocą CGR.

46:3240105992



Rysunek 30. CGR dla podzbioru wszystkich miRNA o największej wiarygodności (ang. confident) oraz konserwatywnych (ang. conserved). Tego typu sekwencje są dostępne poprzez miRBase oraz miRTarBase. Rysunek opublikowano w J Mol Graph Model. w 2024.

W obydwu przypadkach obserwujemy analogiczne wzorce jak w grupie wysokoenergetycznej (porównaj z Rysunek 29). Może to pośrednio wskazywać na istotną rolę innych czynników w procesie degradacji, w tym kompleksu RISC (ang. *RNA-induced silencing complex*). Ponownie wygenerowano wykres częstości, aby w sposób kompletny porównać te wartości dla kolejnych dinukleotydów.

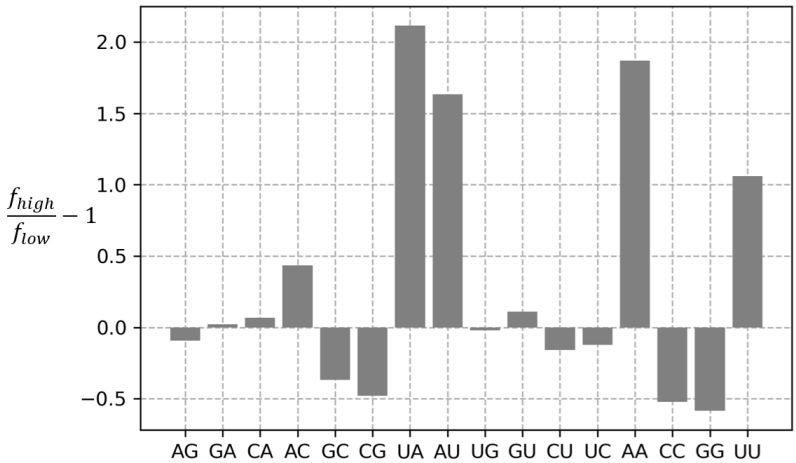


Rysunek 31. Częstość występowania dinukleotydów pośród ludzkich miRNA. Te, po analizie energetycznej podzielono na dwa podzbiory – niskoenergetyczne (po lewej) oraz wysokoenergetyczne (po prawej). Wyróżniono przy tym podział miRNA na tzw. region seed (pierwsze 7 nt., oznaczono ciemniejszym kolorem) oraz pozostałą część (oznaczono jaśniejszym kolorem). Rysunek opublikowano w J Mol Graph Model. w 2024.

Grupy wykazują istotnie różne wartości częstości dla wielu dinukleotydów. Aby łatwiej porównać ze sobą te wyniki, wygenerowano dodatkowo wykres stosunku częstości dinukleotydów w grupie wysokoenergetycznej (f_{high}) do niskoenergetycznej (f_{low}). Dodatkowo, od wyznaczonego stosunku odjęto 1, tak aby 0 było poziomem odniesienia

$$\frac{f_{high}}{f_{low}} - 1 \tag{56}$$

47:1044150419



Rysunek 32. Stosunek częstości występowania dinukleotydów pośród miRNA wysokoenergetycznych oraz niskoenergetycznych. Dodatkowo, od wyznaczonego stosunku odjęto 1. Rysunek opublikowano w J Mol Graph Model. w 2024.

Obserwujemy, że dinukleotydy GC, CG, CC i GG występują szczególnie obficie w grupie o niskiej energii. Wynika to zapewne z faktu, że te nukleotydy należą do zasad silnych (z trzema wiązaniami wodorowymi). Z kolei dinukleotydy UA, AU, AA i UU, które reprezentują zasady słabe, występują częściej w grupie o wysokiej energii. Różnice te są również bardziej wyraźne. Co ciekawe, oba zestawy dinukleotydów obejmują komplementarne podsekwencje: GC/CG, CC/GG, UA/AU oraz AA/UU. Regiony bogate w AU, GU czy CA w regionie 3'UTR odgrywają ważną rolę w kontroli translacji [115,116,117]. Znanym jest również fakt, że cząsteczki miRNA preferencyjnie wiążą się częściej do regionów UTR niż CDS (ang. *Coding Sequences*) prawdopodobnie z powodu większej dostępności jak i obecności charakterystycznych struktur drugorzędowych. Dinukleotydy AU, GU i CA oraz ich komplementarne odpowiedniki występują częściej w zestawie miRNA o wysokiej energii. Dotyczy to zwłaszcza AU. W części wiążącej miRNA w odpowiadającym miejscu docelowym powinny pojawić się podobne wzorce, ze względu na komplementarność. Obecność elementów bogatych w AU (tzw. AREs) w regionach UTR często prowadzi do degradacji mRNA przez białka, takie jak TTP (tristetraprolina) lub AUF1 [118,119]. Wskazano też na możliwość interakcji TTP z mikroRNA poprzez kompleks RISC [120]. Oddziaływanie między miRNA a białkami wiążącymi RNA była jeszcze później badana, szczególnie w kontekście motytwów AU [121]. Pokazano, że elementy AREs mogą być rozpoznawane na wiele różnych sposobów. W [122] autorzy podkreślali rolę miRNA bogatych w GU w neurogenezie; wskazali m.in. procesy związane z miRNA bogatymi w GU. W 2024 roku, w trakcie finalizowania tej pracy, ukazała się publikacja, w której w oparciu o reprezentację gry chaosu przewidywane są miejsca wiązania miRNA [33]. Autorzy jako jedną z zalet wyboru tej reprezentacji wskazali możliwość uwzględnienia funkcjonalnych motywów w różnych częściach sekwencji. W połączeniu z konwolucyjnymi sieciami neuronowymi zbudowano model o dokładności ok. 80%.

Należy zaznaczyć, że reprezentacja gry chaosu dostarcza pełnej informacji o analizowanej sekwencji (lub grupy sekwencji). W szczególności, nie musimy ograniczać się tutaj do badania jedynie częstości występowania dinukleotydów. Powyższe rozważania dotyczyły ludzkich miRNA. W przyszłości warto byłoby zastosować tę metodę do innych organizmów, przeprowadzić analogiczną analizę i ostatecznie porównać ze sobą wyniki – wizualnie lub np. za pomocą metod do analizy obrazów.

48:1155042167

4.2. CGR pozwala scharakreryzować cechy sekwencji genów niezbędnych u bakterii, które mogłyby być wykorzystane jako deskryptory przy ich klasyfikacji

Rozważano 48 bakterii i towarzyszące im zbiory genów niezbędnych (ang. *essential genes*) oraz pozostałych genów (ang. *non-essential genes)* [123]. O ile w pierwszym przypadku, nazwa gen niezbędny funkcjonuje w języku polskim i jest szeroko stosowana, tak w drugim, nie ma tutaj konsensu; autor nie chciał używać mylącej nazwy „gen zbędny", sugerującej, że jakaś sekwencja jest niepotrzebna dla danego organizmu. Zastosowano więc terminologię gen opcjonalny/uzupełniający. Dane pochodziły z bazy danych DEG 15 (*Database of Essential Genes*) [42]. Możemy tam znaleźć również informacje dotyczące pozostałych domen, czyli Eukariontów czy Archeonów, jednak większość publikacji naukowych bazuje głównie na sekwencjach bakteryjnych ze względu na ich wysoką wiarygodność. Przykładowo, niezbędne geny organizmów wyższych są często typowane na podstawie linii komórkowych co może nie odzwierciedlać sytuacji natywnej. Należy też zaznaczyć, że informacje w bazie danych są statyczne, w tym znaczeniu, że dla danej bakterii, gen jest jednoznacznie opisany jako niezbędny lub uzupełniający podczas gdy w środowisku naturalnym może to być kwestia bardziej płynna (np. czasem funkcja jednego genu może być zastąpiona przez udział innego lub grupę innych genów). Prócz sekwencji nukleotydowych znajdują się tam również sekwencje białkowe, jednak do analizy wybrano tylko te pierwsze.

Tabela 4. Rozważany zbiór bakterii. Oprócz nazwy organizmu podano również liczbę genów niezbędnych i opcjonalnych. W przypadku trzech bakterii zbiór genów opcjonalnych nie był dostępny. Tabelę opublikowano w J Theor Biol. w 2021.

Indeks	Organizm	Liczba genów niezbędnych	Liczba genów opcjonalnych
0	Acinetobacter baumannii ATCC 17978a	452	2899
1	Acinetobacter baumannii ATCC 17978b	154	2646
2	Acinetobacter baylyi ADP1	499	2594
3	Agrobacterium fabrum str C58	361	4793
4	Bacillus subtilis 168	271	3904
5	Bacillus thuringiensis BMB171	495	-
6	Bacteroides fragilis 638R	547	3743
7	Bacteroides thetaiotaomicron VPI 5482	325	4453
8	Brevundimonas subvibrioides ATCC 15264	412	2605
9	Burkholderia pseudomallei K96243	505	5222
10	Burkholderia thailandensis E264	406	5226
11	Campylobacter jejuni subsp jejuni 81 176	384	1269
12	Campylobacter jejuni subsp jejuni NCTC 11168a	228	1395
13	Campylobacter jejuni subsp jejuni NCTC 11168b	166	1406
14	Caulobacter crescentus	480	3224
15	Escherichia coli MG1655 I	609	2923
16	Escherichia coli MG1655 II	296	4077
17	Escherichia coli ST131 strain EC958	300	-
18	Francisella novicida U112	392	1329
19	Haemophilus influenzae Rd KW20	642	512
20	Helicobacter pylori 26695	323	1135

49:9575460059

21	Mycobacterium tuberculosis H37Rv	614	2552
22	Mycobacterium tuberculosis H37Rv II	771	3171
23	Mycobacterium tuberculosis H37Rv III	687	3070
24	Mycoplasma genitalium G37	381	94
25	Mycoplasma pulmonis UAB CTIP	310	322
26	Porphyromonas gingivalis ATCC 33277a	463	1627
27	Porphyromonas gingivalis ATCC 33277b	281	1808
28	Pseudomonas aeruginosa PAO1a	117	5454
29	Pseudomonas aeruginosa PAO1b	336	5179
30	Pseudomonas aeruginosa UCBPP PA14	335	960
31	Rhodopseudomonas palustris CGA009	544	4340
32	Salmonella enterica serovar Typhi	353	4005
33	Salmonella enterica serovar Typhi Ty2	358	3906
34	Salmonella enterica serovar Typhimurium SL1344	353	4035
35	Salmonella enterica subsp e. s. T. str. 14028S	105	5210
36	Salmonella typhimurium LT2	230	4228
37	Shewanella oneidensis MR 1	402	1103
38	Sphingomonas wittichii RW1	535	4315
39	Staphylococcus aureus N315	302	2281
40	Staphylococcus aureus NCTC 8325	351	2541
41	Streptococcus agalactiae A909	309	1590
42	Streptococcus pneumoniae	244	-
43	Streptococcus pyogenes MGAS5448	227	1337
44	Streptococcus pyogenes NZ131	241	1177
45	Streptococcus sanguinis	218	2052
46	Synechococcus elongatus PCC 7942	682	1740
47	Vibrio cholerae N16961	779	2943

Jak łatwo można zauważyć, liczba genów niezbędnych waha się w okolicach kilkuset, a dla opcjonalnych kilku tysięcy. Analizę tych sekwencji autor przeprowadził w trzech krokach. Po pierwsze, wyznaczono proste statystyki dotyczące częstości występowania nukleotydów używając m.in. entropii Shannona czy dywergencji Kullbacka-Leiblera [86]. Autor sprawdził dodatkowo m.in. czy drugie prawo Chargaffa jest spełnione w tych dwóch grupach genów. Następnie rozważano rozkład występowania nukleotydów w sekwencji. Ostatecznie, przyjrzano się dalekozasięgowym relacjom występującym w obrębie rozważanych sekwencji.

Analizę genów niezbędnych i opcjonalnych rozpoczęto od zbadania rozkładu długości sekwencji w obu grupach. To ważny krok, ponieważ dalsze etapy mogłyby być bezpośrednio naznaczone tymi potencjalnymi wzorcami. Okazało się jednak, że są one bardzo podobne. Analizując wszystkie geny niezbędne i opcjonalne ze wszystkich bakterii okazało się, że średnia długość w obu grupach to ok. 1027 i 962 odpowiednio. W przypadku kolejnych kwartyli (Q1, Q2 oraz Q3) było to 537/504, 888/828, 1304/1227. Widzimy więc brak istotnych tendencji w którejś z grup. Dodatkowo, wyznaczano współczynnik korelacji tych statystycznych miar (średnia, pierwszy, drugi i trzeci kwartyl) dla każdej z bakterii.

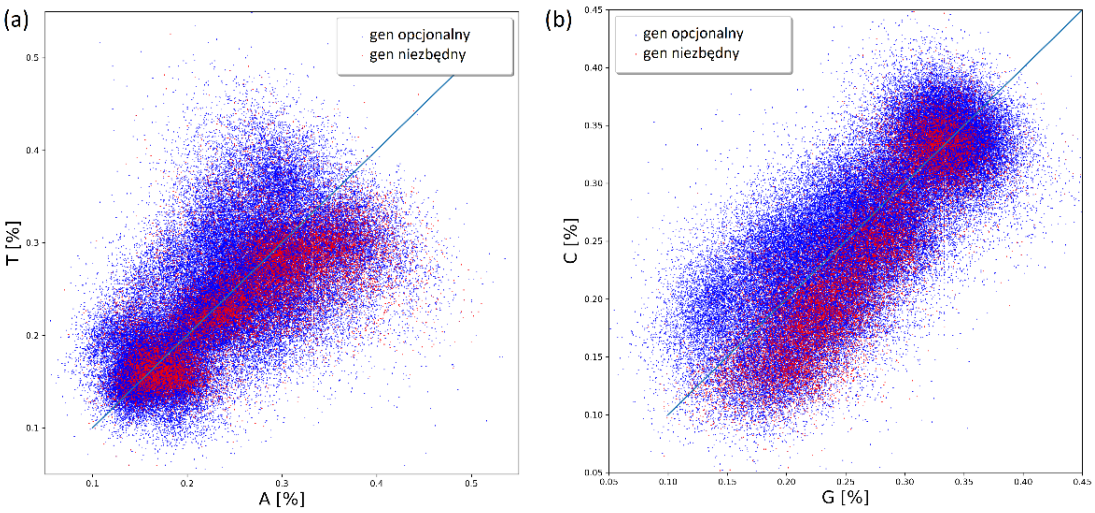
50:1084973931

Tabela 5. Współczynnik korelacji pomiędzy grupą genów niezbędnych i opcjonalnych dla kolejnych bakterii, gdzie rozważano kolejne kwartyle oraz średnią ich długość. Niskie współczynniki oraz wysokie wartości prawdopodobieństwa testowego (p-value) potwierdzają brak liniowej zależności pomiędzy tymi grupami genów. Tabelę opublikowano w J Theor Biol. w 2021.

Zmienne	wsp. korelacji	p-value
Q1	0.30	0.05
Q2	-0.14	0.37
Q3	0.02	0.92
średnia	-0.05	0.77

Niskie współczynniki korelacji oraz wysokie wartości prawdopodobieństwa testowego wskazują na brak liniowej zależności pomiędzy długością genów wśród rozważanych bakterii. Geny nie są zatem ani średnio dłuższe ani krótsze w żadnej z grup. To pożądany wynik ze wspomnianych wcześniej powodów.

Następnie wykreślono zależność pomiędzy G i C oraz A i T dla każdego genu wśród wszystkich rozważanych bakterii. Drugie prawo Chargaffa jest spełnione na poziomie całych genomów, jednak autor uznał interesującym sprawdzenie go na poziomie genów.



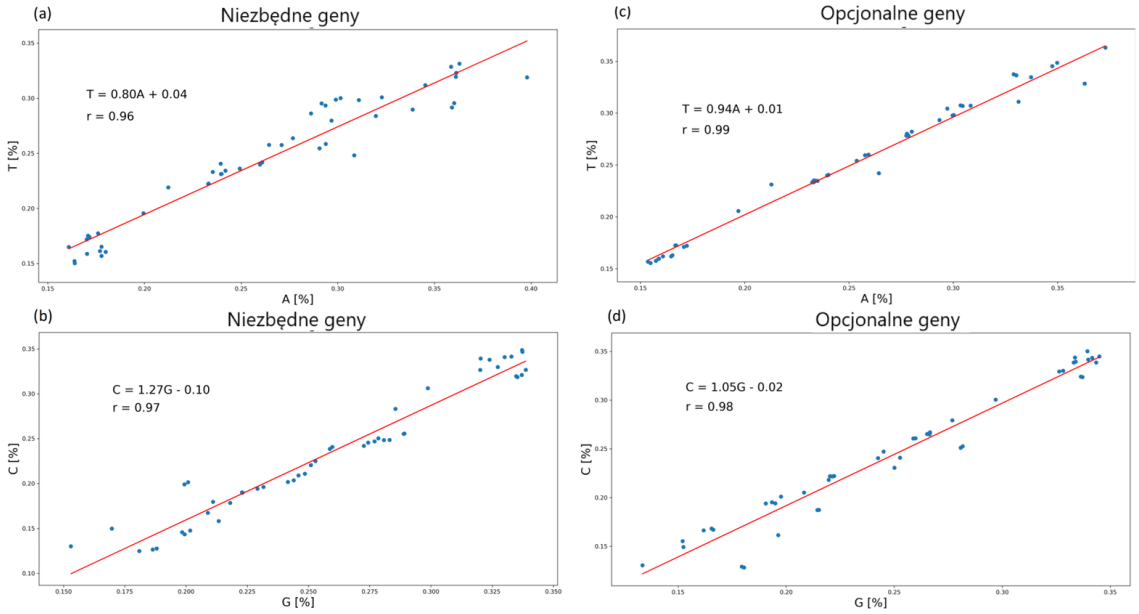
Rysunek 33. Zależność pomiędzy zawartością A i C oraz G i C dla genów niezbędnych i opcjonalnych. Jedna kropka symbolizuje jeden gen. Kolorem czerwonym oznaczone są geny niezbędne, a niebieskim geny opcjonalne. Rysunek opublikowano w J Theor Biol. w 2021.

Obserwujemy tutaj dość intrygującą zależność. Geny opcjonalne są równo rozłożone wokół prostej, co wskazuje na symetrię. W przypadku genów niezbędnych ta symetria jest złamana; kolejne geny znajdują się częściej pod prostą. Dokładniej, w przypadku A/T liczba genów pod i nad prostą to odpowiednio 49% oraz 51% dla puli genów opcjonalnych. Dla genów niezbędnych to już 65% oraz 35%. W przypadku G/C są to 53%/47% oraz 71%/29% odpowiednio. Wskazuje to na istotnie różny skład sekwencji pośród tych dwóch grup.

Następnie, dla każdej bakterii złączono w dwie ciągłe sekwencje jej geny niezbędne oraz opcjonalne tworząc swego rodzaju „pseudogenomy" (sekwencje złożone tylko z genów jednej z kategorii). Analiza ta stanowi rozszerzenie poprzedniej części, a jej celem było określenie czy

51:1065600890

obserwowane zależności uogólniają się dla każdej z bakterii lub czy jest to efekt tylko niewielkiej grupy bakterii. Podobnie jak poprzednio wykreślono zależności A/T oraz G/C.



Rysunek 34. Zależność pomiędzy A i C oraz G i C dla pseudogenomów złożonych tylko z genów niezbędnych lub opcjonalnych. Jedna kropka symbolizuje jedną bakterię. Rysunek opublikowano w J Theor Biol. w 2021.

W każdym przypadku współczynnik korelacji r wskazuje na istotną zależność liniową. Warto tutaj szczególnie przyjrzeć się współczynnikowi kierunkowemu. Dla genów opcjonalnych jego wartość jest bliska 1 (0.94 oraz 1.05). W przypadku genów niezbędnych obserwujemy istotne odstępstwa (0.80 oraz 1.27 odpowiednio). Jest to konsystente z poprzednimi wnioskami. Oszacowano również błędy wyestymowanych parametrów oraz prawdopodobieństwo testowe (p-value). Niektóre z rozważanych bakterii są bardziej spokrewnione. Istotnym wydaje się tutaj pytanie, czy nie występuje swego rodzaju nadreprezentacja jakiejś grupy bakterii, która istotnie wpływa na wyniki. Wszystkie 48 bakterie podzielono więc na mniejsze grupy (związane z pokrewieństwem – gatunkiem lub rodzajem), otrzymując zbiór 24 grup {(0, 1, 2), (3), (4, 5), (6, 7), (8), (9, 10), (11, 12, 13), (14), (15, 16, 17), (18), (19), (20), (21, 22, 23), (24, 25), (26, 27), (28, 29, 30), (31), (32, 33, 34, 35, 36), (37), (38), (39, 40), (41, 42, 43, 44, 45), (46), (47)}. Powtórzono poprzednie obliczenia wykonując 1000 symulacji, każdorazowo losując po 1 bakterii z każdej z grup (1000 kontroli).

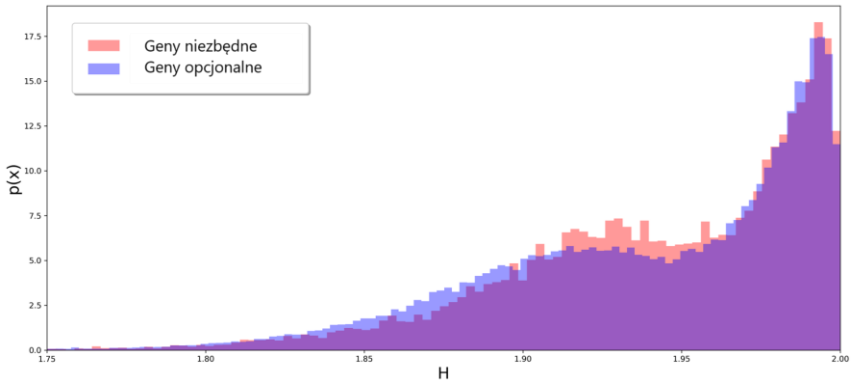
Tabela 6. Wartości oszacowanych współczynników a i b dla modeli regresji liniowej $y = ax + b$. Podano również prawdopodobieństwo testowe (p-value) oraz ocenę modelu (R^2). W nawiasach znajdują się błędy oszacowań. Gwiazdką oznaczono parametry oszacowane dla kontroli. Tabelę opublikowano w J Theor Biol. w 2021.

	r	a	p-value	b	p-value	R^2	r*	a*	b*
Niezbędne GC	0.97	1.27 (0.04)	e-16	-0.10 (0.010)	e-10	0.95	0.98	1.32 (0.03)	-0.10 (0.007)
Opcjonalne GC	0.98	1.05 (0.03)	e-16	-0.02 (0.009)	e-2	0.96	0.99	1.04 (0.02)	-0.01 (0.007)
Niezbędne AT	0.96	0.80 (0.03)	e-16	0.04 (0.009)	e-4	0.93	0.97	0.80 (0.02)	0.03 (0.005)
Opcjonalne AT	0.99	0.94 (0.02)	e-16	0.01 (0.005)	e-2	0.99	0.99	0.96 (0.02)	0.01 (0.004)

52:1198255236

Obserwujemy istotność statystyczną otrzymanych wyników, w szczególności dla parametru a. Wszystkie modele osiągnęły też wysoką wartość R^2. Wyniki dla kontroli są porównywalne z wynikami dla wszystkich bakterii i mieszczą się w ramach błędu statystycznego (te podano w nawiasach). Wskazuje to na możliwość generalizacji wskazanych wyżej zależności.

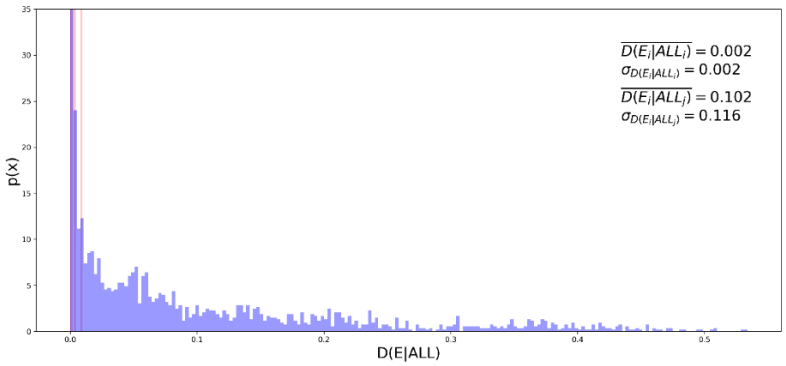
Uwzględniając wszystkie częstości nukleotydów równocześnie, wyznaczono entropię H dla każdego genu. W przypadku losowej sekwencji wynosiłaby ona 2.



Rysunek 35. Rozkład entropii wyznaczonej dla każdego genu. Wyróżniono geny niezbędne i opcjonalne. Wykres częściowo przycięto z lewej strony. Rysunek opublikowano w J Theor Biol. w 2021.

Obserwujemy przede wszystkim wysokie wartości entropii (bliskie 2). Rozkłady entropii w obu grupach są zbliżone, chociaż występują też pewne subpopulacje genów niezbędnych o wyższej entropii oraz genów opcjonalnych o niższej entropii. Z tego powodu test Kołomogorowa-Smirnowa wskazał na istotność statystyczną różnicy pomiędzy tymi grupami.

Inną istotną miarą z teorii informacji jest dywergencja Kullbacka-Leiblera. Celem jej zastosowania dla rozważanych sekwencji było stwierdzenie czy geny niezbędne stanowią swego rodzaju reprezentatywną grupę względem całego genomu dla każdej z bakterii (w sensie częstości występowania nukleotydów). Wyznaczono tę wartość porównując każdy pseudogenom genów niezbędnych z odpowiednim całym zbiorem genów dla kolejnej bakterii oraz z innymi genomami bakterii.



Rysunek 36. Dywergencja Kullbacka-Leiblera da pseudogenomów złożonych z genów niezbędnych (E) oraz całych genomów (ALL). Porównywano korespondujące pseudogenomy i genomy oraz te przypadające dla różnych organizmów. Wykres częściowo przycięto na osi OY. Rysunek opublikowano w J Theor Biol. w 2021.

53:4860337845

Rozkłady dywergencji wskazują na ich istotnie odmienny charakter. Porównanie pseudogenomu genów niezbędnych z genomem tej samej bakterii prowadzi do mniejszych wartości (średnio 0.002) niż w przypadku analogicznych porównań pomiędzy różnymi bakteriami (0.102). Wyniki te można podsumować stwierdzeniem, że z punktu widzenia teorii informacji, informacja zawarta tylko w zbiorze genów niezbędnych jest porównywalna z informacją zawartą w całym genomie dla każdej bakterii. Należy tutaj podkreślić, że analiza dotyczy częstości występowania nukleotydów i nie bierze ona pod uwagę samego rozkładu nukleotydów w obrębie tych grup. Taką analizę przeprowadzono w dalszej części.

Dla alfabetu 4 literowego, dla losowej sekwencji, prawdopodobieństwo wystąpienia każdej z liter na danej pozycji wynosi $p = \frac{1}{4}$. W tym przypadku, rozkład geometryczny opisuje rozkład przerwy pomiędzy wystąpieniem dwóch tych samych nukleotydów w sekwencji. Wtedy, wartość oczekiwana $EX = \frac{1}{p} = 4$ oraz odchylenie standardowe $DX = \sqrt{\frac{1-p}{p^2}} = 3.46$. Sekwencje biologiczne oczywiście nie są losowe i wykazują odmienny charakter. Wyznaczono te wartości dla każdego z nukleotydów (A, C, T oraz G) w sekwencjach genów niezbędnych i opcjonalnych.

Tabela 7. Wartości oczekiwane i odchylenia standardowe dla każdego z nukleotydów pośród genów niezbędnych i opcjonalnych. Tabelę opublikowano w J Theor Biol. w 2021.

Nukleotyd	Zbiór genów niezbędnych	Zbiór genów opcjonalnych
A	4.03/4.03	4.36/4.36
C	3.99/3.68	3.70/3.30
T	4.29/4.14	4.36/4.32
G	3.69/3.21	3.65/3.22

Obserwujemy istotnie różne od sekwencji losowych wartości. Dodatkowo, inne wzorce występują pośród genów niezbędnych oraz opcjonalnych. W drugim przypadku wartości oczekiwane i odchylenia standardowe dla tyminy oraz adeniny są zbliżone; podobnie dla guaniny oraz cytozyny. Koresponduje to bezpośrednio z wykresem G/C oraz A/T w poprzedniej części oraz rozszerza go. Okazuje się bowiem, że nie tylko częstości, ale i rozkład przerw pomiędzy kolejnymi wystąpieniami tego samego nukleotydu są zbliżone. Te zależności nie występują w przypadku zbioru genów niezbędnych. Dla przykładu, wartość oczekiwana dla A to 4, a dla T to 4.3, czyli średnio dłużej czekamy na wystąpienie tyminy niż adeniny.

54:4873299933

Przeprowadzano ponownie analogiczną analizę, jednak dla każdej bakterii osobno. Tym razem również uwzględniono wprowadzoną wcześniej kontrolę (wyniki oznaczone są gwiazdką).

Tabela 8. Zależności liniowe dla średnich i odchyleń standardowych rozkładu przerw dla kolejnych nukleotydów. Wartości te liczone były dla każdej bakterii – pomiędzy zbiorem genów niezbędnych i opcjonalnych. Gwiazdką oznaczono parametry oszacowane dla kontroli. Tabelę opublikowano w J Theor Biol. w 2021.

Nukleotyd	r	a	p-value	b	p-value	R^2	r*	a*	b*
$\bar{A}$	0.99	1.07 (0.02)	e-16	-0.19 (0.09)	e-2	0.98	0.99	1.07 (0.01)	-0.13 (0.05)
σ_A	0.99	1.04 (0.02)	e-16	-0.03 (0.09)	0.73	0.98	0.99	1.03 (0.01)	0.04 (0.04)
$\bar{C}$	0.98	0.90 (0.03)	e-16	0.28 (0.12)	e-2	0.96	0.98	0.88 (0.04)	0.37 (0.14)
σ_C	0.99	0.94 (0.02)	e-16	0.15 (0.09)	e-1	0.98	0.99	0.93 (0.03)	0.18 (0.09)
$\bar{T}$	0.99	1.04 (0.02)	e-16	-0.33 (0.10)	e-3	0.98	0.99	1.03 (0.01)	-0.30 (0.05)
σ_T	0.99	1.11 (0.03)	e-16	-0.49 (0.11)	e-5	0.97	0.99	1.10 (0.01)	-0.46 (0.04)
$\bar{G}$	0.98	1.26 (0.04)	e-16	-0.83 (0.16)	e-6	0.96	0.98	1.30 (0.04)	-0.95 (0.15)
σ_G	0.98	1.25 (0.04)	e-16	-0.52 (0.14)	e-4	0.96	0.99	1.28 (0.05)	-0.59 (0.13)

Obserwujemy bardzo charakterystyczne wzorce. Parametry oceny modelu są wysokie. Podobnie jak wcześniej, kontrolne rachunki wykluczyły efekt jakiejś grupy bakterii. Gdyby $a = 1$ oraz $b = 0$, oznaczałoby to, że geny niezbędne i opcjonalne nie różnią się w kontekście rozkładu przerw. Jednak okazuje się, że są one istotnie różne, co widoczne jest najbardziej w przypadku średniej wartości G. Tutaj wartość współczynnika kierunkowego wynosi 1.26 a wartość wyrazu wolnego -0.83.

Dotychczasowe analizy pokazały odrębny charakter sekwencji genów niezbędnych oraz opcjonalnych. Te drugie wydają się po uśrednieniu spełniać drugie prawo Chargaffa, co zostało pokazane na wykresach A/T ora G/C. Co więcej, okazało się, że nawet rozkład przerw jest podobny dla tych grup nukleotydów. Takie zależności zapewne są istotne biologicznie. Warto tutaj wspomnieć chociażby o regule Szybalskiego według której niektóre organizmy (w tym wirusy) posiadają mniej pirymidyn niż puryn [83]. W przypadku wirusów chodzi o zapobieżeniu formowaniu się dwuniciowej struktury RNA. Reguła ta dotyczy również niektórych bakterii. Jeżeli chodzi o regułę Chargaffa i jej termodynamiczne wyjaśnienie, może być ono niewystarczalne. O ile geny opcjonalne zdają się w prosty sposób poddawać regułom termodynamiki, o tyle pośród genów niezbędnych istotną rolę odgrywają zapewne względy biologiczne jak np. wpływ kompozycji sekwencji na formowanie się struktury drugorzędowej transkryptu. Jednym ze znanych fenomenów, jest tzw. zjawisko preferencji kodonów (ang. *codon bias*) mierzone np. za pomocą współczynnika CAI (ang. *Codon Adaptation Index*) [82]. Okazuje się bowiem, że pośród wysoko ekspresjonowanych genów występuje pewna preferencja kodonów nad innymi.

55:2669847097

Dotychczas analiza opierała się na częstości nukleotydów oraz rozkładzie przerw pomiędzy nimi. Pomimo tych uproszczeń otrzymano kilka ciekawych wyników. Naturalnie jednak tracimy tutaj część informacji związaną np. z korelacjami pomiędzy występowaniem różnych nukleotydów. Mimo, że przerwy dość dobrze obrazują charakter sekwencji z punktu widzenia danej litery alfabetu, to już zależności między nimi nie są tutaj widoczne. Z tego powodu postanowiono poddać analizie rozważane sekwencje bez istotnej redukcji informacji. W tym celu sekwencje zostały przetworzone na ciągi punktów za pomocą spaceru DNA. Ten z kolei oparto o reprezentację gry chaosu, gdzie kolejny punkt zostaje wygenerowany standardowo wg

$$N_{i+1} = \frac{N_i + N}{2}, \tag{57}$$

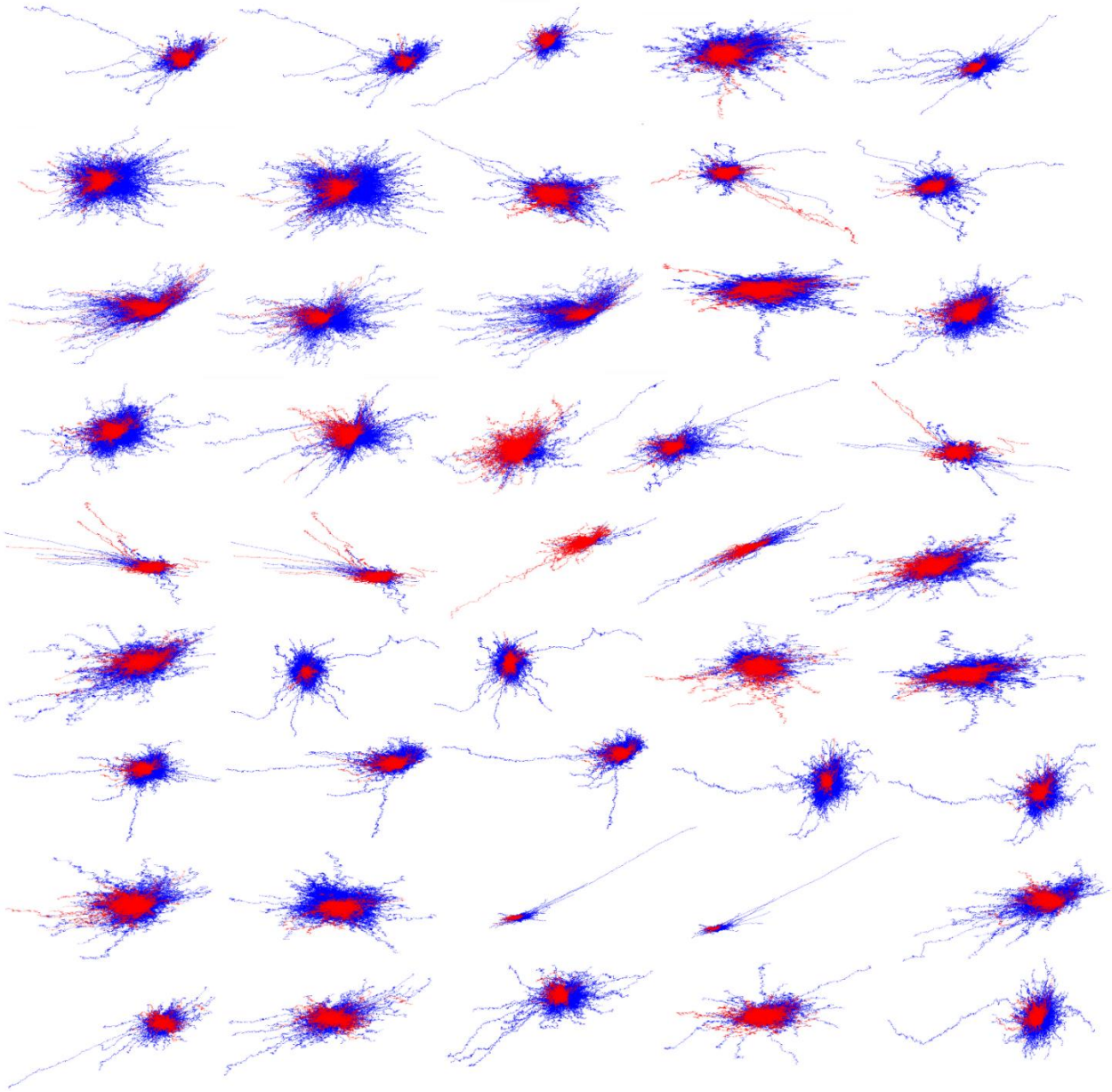
gdzie N związany jest z danym nukleotydem $N_A(-1,-1), N_G(-1,1), N_C(1,-1)$ oraz $N_T(1,1)$. Następnie z tego ciągu punktów powstał spacer DNA $\{Z_i\} = \{\sum_{j=1}^{i} N_j\}$. Zastosowano taką modyfikacje (zaprzężenie CGR ze spacerem DNA) aby bardziej zróżnicować rozkład punktów na płaszczyźnie. W klasycznym podejściu może się zdarzyć (i dość często się to dzieje), że punkty nachodzą na siebie. Poprzez wprowadzenie CGR ograniczamy ten efekt. Przed wyznaczeniem momentów bezwładności, środki ciężkości zostały jeszcze przeniesione do początku układu współrzędnych. Formalnie, stworzony został kolejny ciąg punktów

$$Z_i \mapsto Z_i - \bar{Z}, \tag{58}$$

gdzie $\bar{Z} = \sum_{j=1}^{n} Z_j$, gdzie n to liczba nukleotydów w sekwencji.

Na rysunku poniżej przedstawiono rozważane bakterie z wyszczególnieniem ich genów niezbędnych i opcjonalnych.

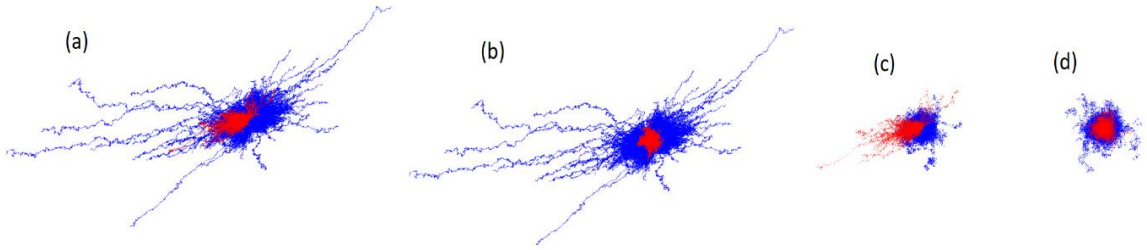
56:4469368303



Rysunek 37. Spacery DNA dla 45 bakterii, dla których dostępne były zarówno zbiory genów niezbędnych oraz opcjonalnych. Każda składowa związana jest z jednym genem. Geny niezbędne oznaczono kolorem czerwonym, opcjonalne niebieskim. Rysunek opublikowano w J Theor Biol. w 2021.

Różne gatunki charakteryzują odmienne reprezentacje. Część organizmów jest bliżej spokrewniona, co też łatwo tutaj zauważyć. Każda gałąź odpowiada jednemu genowi. Sekwencje zbliżone do losowych prowadzą do krótkich reprezentacji. Z kolei te o niskiej złożoności czy o wielokrotnych powtórzeniach jakiegoś nukleotydu odpowiadają długim, częściowo liniowym krzywym. Sekwencje genów niezbędnych wydają się bardziej skupione w centrum. Może to być jednak efekt liczebności genów. Jak pamiętamy, genów niezbędnych jest kilkaset podczas gdy genów opcjonalnych kilka tysięcy.

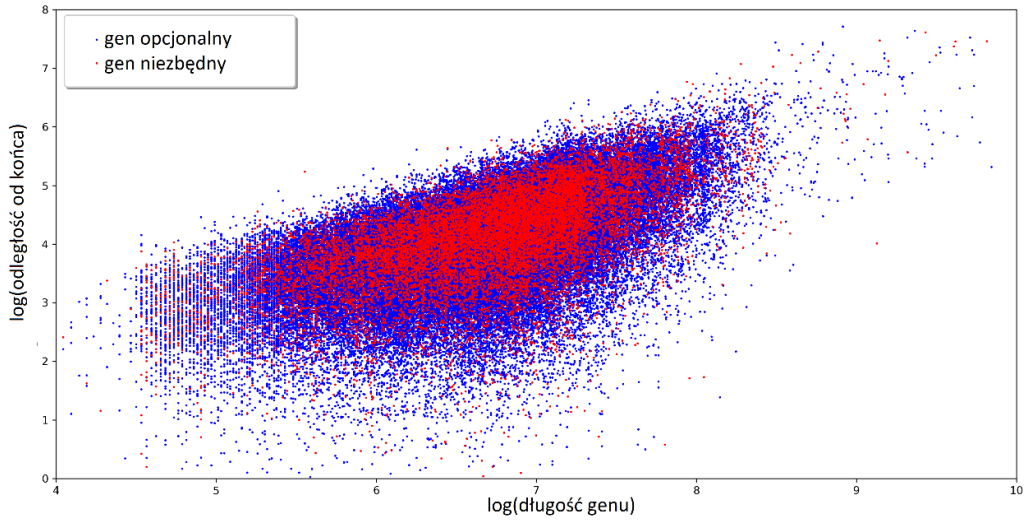
57:8629850066



Rysunek 38. Oryginalne sekwencje jednej z bakterii (a); geny niezbędne zastąpione losowymi o tym samym rozkładzie długości (b); geny opcjonalne zostały zastąpione sekwencjami losowymi o tym samym rozkładzie długości (c); zarówno sekwencje niezbędne jak i optymalne zostały zastąpione sekwencjami losowymi (d). Rysunek opublikowano w J Theor Biol. w 2021.

Zarówno zbiór genów niezbędnych jak i opcjonalnych istotnie różni się od sekwencji losowych wygenerowanych z tego samego rozkładu długości. W tym przypadku nie obserwujemy dłuższych fragmentów, a cała reprezentacja jest bardziej kompaktowa. Związane jest to ze wspomnianym wcześniej wyjaśnieniem. W dalszej części zostanie zbadane czy sekwencje niezbędne i opcjonalne istotnie różnią się i jak/czy te zależności utrzymują się dla kolejnych bakterii.

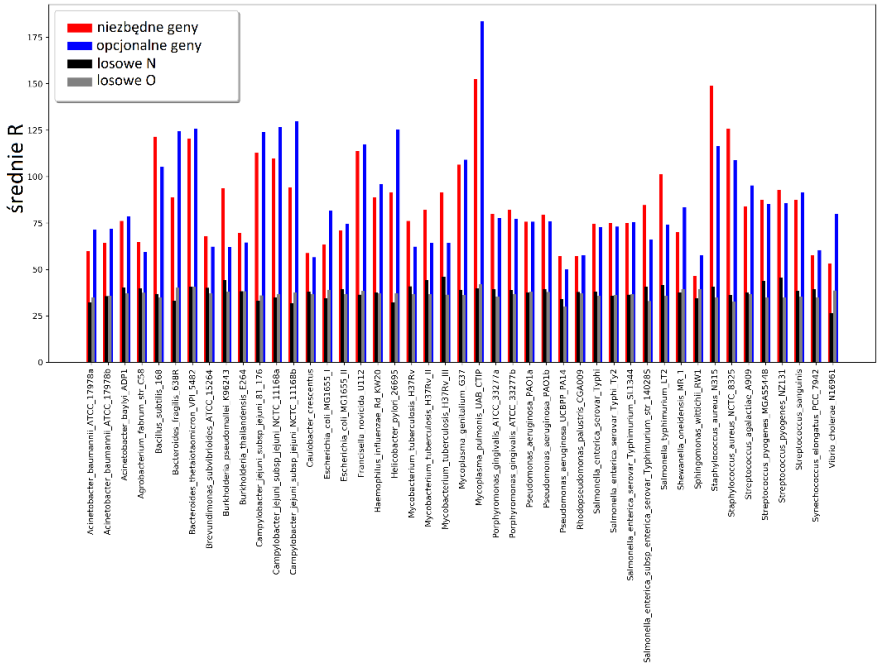
Dla każdego genu wyznaczono odległość geometryczną od początku do końca reprezentacji. Ta prosta miara charakteryzuje m.in. na ile sekwencja jest jednorodna oraz czy występują tam dłuższe powtórzenia.



Rysunek 39. Zależność pomiędzy długością genu o odległością do końca reprezentacji. Każdy punkt związany jest z jedną sekwencją. Generalnie, wraz ze wzrostem długości genu obserwujemy wzrost odległości od końca. Dane przedstawiono w skali logarytmicznej. Rysunek opublikowano w J Theor Biol. w 2021.

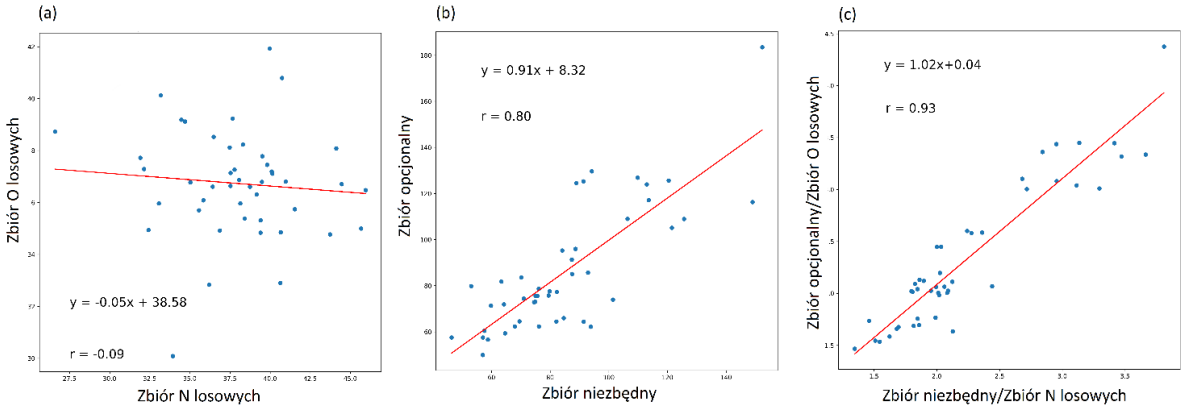
Oprócz oczekiwanej, pozytywnej zależności wzrostu tej miary z długością genu (oznaczenie R, chociaż nie jest to zależność silnie liniowa) nie obserwujemy tutaj istotnych różnic pomiędzy rozważanymi grupami. Zarówno niezbędne jak i opcjonalne geny wykazują podobny rozkład. Warto przyjrzeć się średniej wartości tego współczynnika dla kolejnych bakterii.

58:2412969620



Rysunek 40. Średni poziom R, czyli odległości do końca reprezentacji, pośród genów niezbędnych i opcjonalnych dla kolejnych bakterii. Dodatkowo, wygenerowano losowe sekwencje z tego samego rozkładu długości. Rysunek opublikowano w J Theor Biol. w 2021.

Można zauważyć, że wartość ta różni się pomiędzy bakteriami. W przypadku sekwencji losowych o tym samym rozkładzie długości genów które występują w kolejnych bakteriach ten efekt nie występuje. a zatem obserwowane różnice wśród bakterii nie są konsekwencją ilości sekwencji ani ich długości. Aby to jeszcze lepiej zilustrować wygenerowano kolejne wykresy.



Rysunek 41. Zależności pomiędzy średnim poziomem R, czyli odległości do końca reprezentacji, dla losowych sekwencji (a); oryginalnych sekwencji (b); stosunków oryginalnych sekwencji do losowych (c). Rysunek opublikowano w J Theor Biol. w 2021.

W przypadku sekwencji losowych nie obserwujemy istotnej zależności liniowej. Współczynnik kierunkowy prostej jest bliski zero, podobnie współczynnik korelacji. W przypadku sekwencji bakteryjnych obserwujemy umiarkowaną zależność liniową pomiędzy zbiorem genów niezbędnych oraz opcjonalnych. Tę zależność można jeszcze poprawić stosując poprawkę na rozkład długości, który sam w sobie nie wykazuje istotnego wpływu, ale przynosi wymierne korzyści jako czynnik normalizujący. Po podzieleniu średnich poziomów współczynnika R przez odpowiednią wartość dla zbioru losowego otrzymujemy silną zależność liniową.

59:7631724169

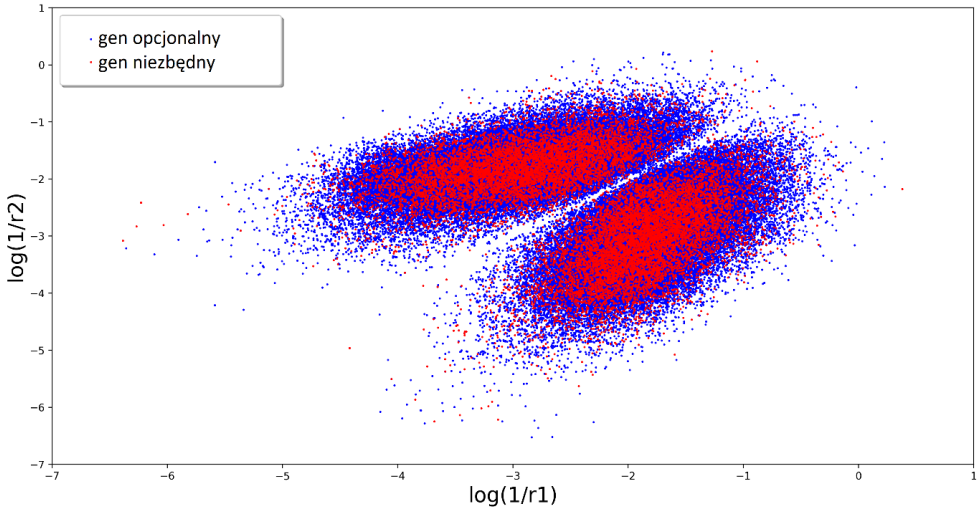
Tabela 9. Wartości współczynników modelu $y = ax + b$, gdzie badano odległość od końca reprezentacji między pulą genów opcjonalnych i niezbędnych. Dodatkowo umieszczono informację o prawdopodobieństwie testowym (p-value) i jakości modelu (R^2). Gwiazdką oznaczono parametry oszacowane dla kontroli. Tabelę opublikowano w J Theor Biol. w 2021.

	r	a	p-value	b	p-value	R^2	r^*	a^*	b^*
Losowe sekwencje	-0.09	-0.05 (0.08)	0.60	38.58 (3.15)	e-15	0.01	-	-	-
Niezbędne vs Opcjonalne	0.80	0.91 (0.10)	e-11	8.32 (9.23)	0.37	0.64	0.88	0.89 (0.09)	12.14 (7.08)
Znormalizowane N vs O	0.93	1.02 (0.06)	e-16	0.04 (0.14)	0.76	0.87	0.94	0.99 (0.04)	0.09 (0.07)

W przypadku sekwencji losowych jedynie wyraz wolny jest istotny statystycznie. Dla genów niezbędnych i opcjonalnych (zarówno przed jak i po normalizacji) współczynnik kierunkowy jest poprawnie wyestymowany. Okazuje się, że niwelując efekty długości obserwujemy prostą zależność pomiędzy niezbędnym a opcjonalnym zbiorem genów. Jest to wynik interesujący i jednocześnie wydawać by się mogło stający w sprzeczności do wyników analizy długości przerwy pomiędzy sekwencjami (tam pokazaliśmy, że zależności między tymi grupami genów istnieją, jednak są bardziej złożone). Należy jednak pamiętać, że w tamtej analizie rozważany był jedynie aspekt długości przerwy, a nukleotydy analizowane były niezależnie. Analiza ze współczynnikiem R daje pełniejszy obraz tych zależności (jest to uśrednienie wielu cech sekwencji).

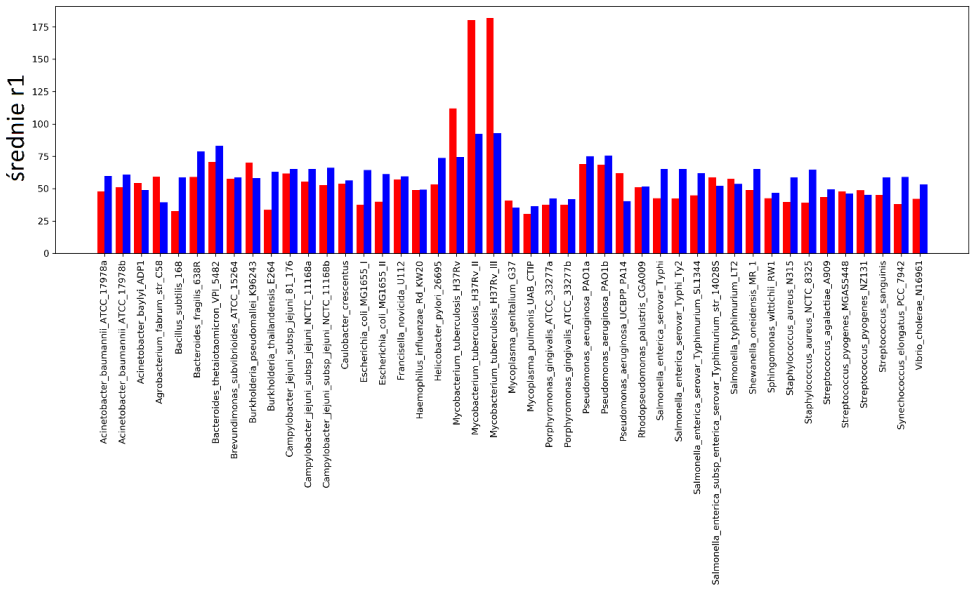
Przeprowadzono też analogiczną analizę z uwzględnieniem momentów bezwładności r_1 oraz r_2. Tym razem bezpośrednio rozważamy rozkład punktów na płaszczyźnie. Przy wyznaczaniu odległości od końca (R) brano pod uwagę tylko ostatni punkt, chociaż oczywiście uwzględniał on całą wcześniejszą historię. Technika ma swoją inspirację w fizyce. Punkty traktowane są jako masy jednostkowe, a cała sekwencja reprezentowana jako obiekt o danej masie (równej ilości reszt) oraz odpowiedniej geometrii (związanej z kolejnością występowania tych reszt). Poniższy wykres zawiera zależność logarytmów $\frac{1}{r_1}$ od $\frac{1}{r_2}$. Taka transformacja została wybrana ze względu na dogodny rozkład punktów do ich prezentacji, po jej zastosowaniu. Nie wpływa to oczywiście na relacje między punktami.

60:1043732676



Rysunek 42. Zależność pomiędzy r1 a r2. Zastosowano dodatkowe transformacje dla wygodniejszej reprezentacji punktów. Czerwonym kolorem oznaczono geny niezbędne, niebieskim opcjonalne. Rysunek opublikowano w J Theor Biol. w 2021.

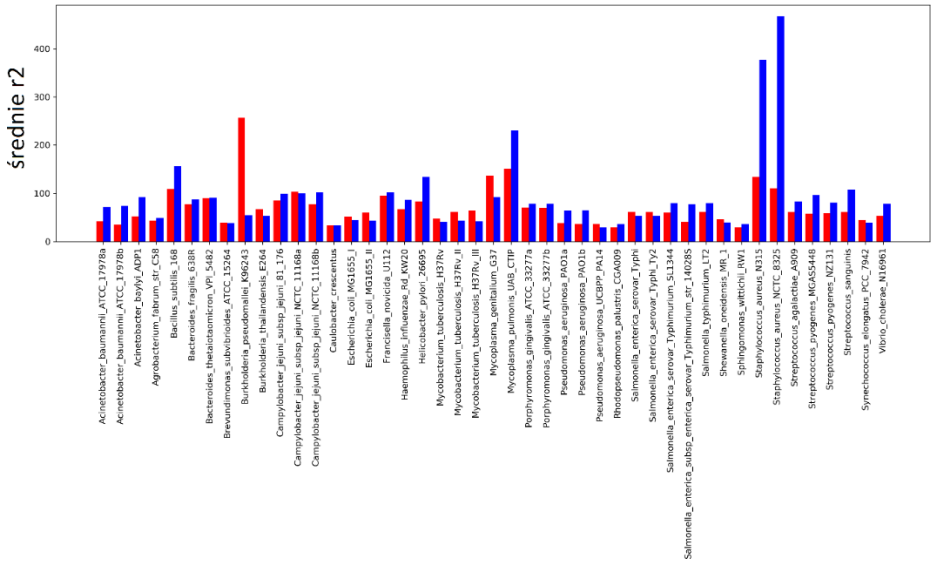
Podobnie jak poprzednio, punkty czerwone reprezentujące geny niezbędne mają zbliżony rozkład do tych niebieskich związanych z genami opcjonalnymi. Spójrzmy teraz na średnie poziomy $r_1$oraz r_2 w tych grupach, dla każdej z bakterii.



Rysunek 43. Średni poziom r1 dla kolejnych bakterii. Czerwonym kolorem oznaczono geny niezbędne, niebieskim opcjonalne.

Porównując te wyniki z poprzednimi (dotyczącymi R) można zaobserwować nieco mniejsze zróżnicowanie pomiędzy bakteriami dla tego współczynnika. Sprawdzono, że nadal jest on na znacznie wyższym poziomie niż dla sekwencji losowych. Dodatkowo, dla *Mycobacterium tuberculosis* średni poziom r_1 w grupie genów niezbędnych jest znacząco wyższy w stosunku do drugiej grupy. Analogiczny wykres wygenerowano dla r_2

61:3495407757



Rysunek 44. Średni poziom r2 dla kolejnych bakterii. Czerwonym kolorem oznaczono geny niezbędne, niebieskim opcjonalne.

Wyróżniają się tutaj szczególnie *Burkholderia pseudomallei* (wyższy średni poziom wśród genów niezbędnych) oraz *Staphylococcus aureus* (znacząco wyższy poziom wśród genów opcjonalnych). Średnie poziomy r_2 są nadal istotnie większe niż dla sekwencji losowych o tych samych długościach. Analogicznie jak w przypadku R, dla r_1 i r_2 zostały wygenerowane zależności pomiędzy zbiorami niezbędnymi oraz opcjonalnymi dla kolejnych bakterii. Współczynniki korelacji wskazują na słabszą zależność liniową i wynoszą odpowiednio 0.64 oraz 0.44. Związane jest to z mniejszą różnorodnością tych współczynników pomiędzy bakteriami oraz przypadkami, gdzie jeden ze zbiorów genów wykazuje istotnie wyższy średni poziom danego współczynnika. Przykładowo, po usunięciu bakterii o indeksach 8, 23, 37, 38 (czyli najbardziej odstających, porównaj z Tabela 4.) korelacja dla r_2 wzrasta z 0.44 do 0.68.

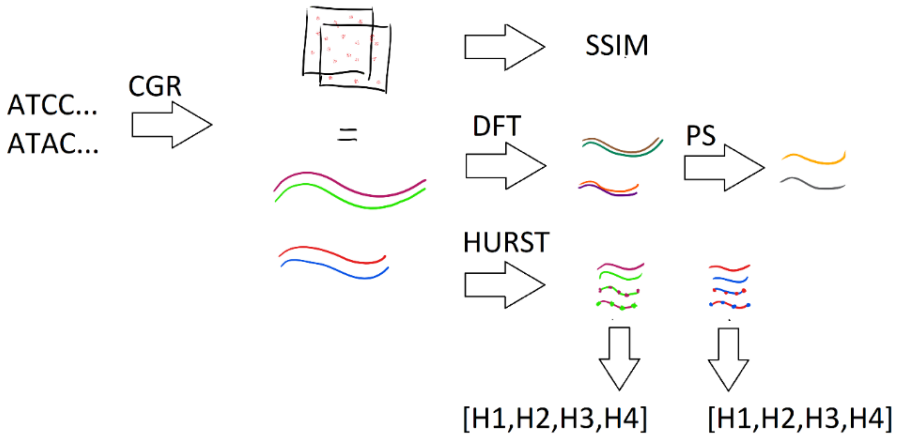
Badania te były potem jeszcze kontynuowane w ramach pracowni magisterskiej (P. Czarnota, *Zastosowanie sieci neuronowych do predykcji genów niezbędnych u bakterii*, magisterium na bioinformatyce, 2023.). Zbudowano m.in. klasyfikator do przyszłego rozróżnienia między pulą genów niezbędnych a opcjonalnych w oparciu o klasyczne metody nauczania maszynowego, gdzie jako cechy wybrano między innymi stosunki A/T czy G/C. W dalszej części, zastosowano też architekturę LSTM (*ang. The Long Short-Time Memory*) aby uwzględnić dalekozasięgowe korelacje. Dokładność wynosiła ok 70% (dla każdej z grup), co w praktycznych zastosowaniach może nie być wystarczające. Nieco lepsze wyniki osiągnięto przy użyciu konwolucyjnych sieci neuronowych (sama reprezentacja sekwencji była także zrealizowana z użyciem CGR) [124]. W naszych modelach posiłkowaliśmy się jedynie sekwencjami nukleotydowymi; warto w przyszłości wspomóc się chociażby homologią w stosunku do znanych sekwencji. W literaturze pojawiły się też inne deskryptory sekwencji genów niezbędnych oparte o reprezentację gry chaosu, m.in. wykładnik Hursta [26,27]. Stwierdzono, że ta pula genów różni się od pozostałych zarówno zakresem wartości jak i rozkładem tego parametru. W przypadku genów niezbędnych występuje rozkład normalny. Problem wyboru odpowiednich deskryptorów jak i niezbilansowania danych nadal pozostaje częściowo otwarty. W 2023 analizowane były też geny niezbędne z ludzkich linii komórkowych m.in. przy użyciu krzywej Z oraz metody wygładzania [125]. Ta ostatnia, istotnie poprawiła finalną dokładność.

62:1296613338

4.3. CGR może być stosowana w molekularnej analizie filogenetycznej do konstrukcji drzew filogenetycznych

4.3.1. Analiza sekwencji referencyjnych oraz genu neuraminidazy wirusów grypy

W tym rozdziale zweryfikowano czy i w jaki sposób CGR może być zastosowany do rekonstrukcji relacji ewolucyjnych pomiędzy grupą sekwencji. Wygenerowano w uproszczony sposób zestaw sekwencji o znanej historii ewolucyjnej, aby zbadać efektywność rozważanych metod i porównać je między sobą. Rozpoczęto od transformacji sekwencji biologicznych za pomocą CGR. W przypadku sekwencji białkowych stosowano dwa podejścia – redukcja do 4 znaków oraz zastosowanie hydrofobowości/hydrofilowości jako współrzędnych dla aminokwasów. Na temat drugiego podejścia (i innych parametryzacji) będzie nieco więcej w późniejszych rozdziałach. Po transformacji ciągu liter na ciąg par liczb nastąpiło ich dalsze przetwarzanie. Reprezentacje z CGR można traktować jak obrazy, a porównanie sekwencji zostaje w tym przypadku sprowadzone do porównania ze sobą tych obrazów. Stopień podobieństwa może być określony za pomocą metody SSIM [34], a końcowa odległość wyliczana jest jako $1 - SSIM$. Z drugiej strony, wartości po zastosowaniu CGR stanowią ciąg liczbowy. Jest to ciąg par, które mogą być traktowane jako liczby zespolone a zatem można do nich zastosować metodę DFT. Następnie wyznaczane są widma mocy (ang. *Power Spectra*, PS), które są ewentualnie wydłużane do najdłuższej sekwencji. Porównanie sekwencji polega na porównaniu ze sobą ich widm [35,105]. Jeszcze inne podejście polega na zastosowaniu wykładnika Hursta [36]. Ten liczony jest dla szeregu jednowymiarowego (występują wartości rzeczywiste). We wspomnianej pracy wyznaczono sumę współrzędnych na kolejnych pozycjach. Jest to istotna redukcja informacji, dlatego w tej pracy autor dokonał dekompozycji na 4 składowe. Są to kolejno: pierwsza współrzędna, druga współrzędna, średnia arytmetyczna ze współrzędnych oraz średnia geometryczna ze współrzędnych. W ostatnim przypadku, aby uniknąć ewentualnych wartości ujemnych zastosowano funkcję signum pod pierwiastkiem, a następnie przywrócono informację o znaku przed pierwiastkiem. Można by też zwyczajnie policzyć iloczyn wartości z kolejnych współrzędnych, pierwiastek ma jedynie na celu otrzymania wartości z zakresu porównywalnego do pozostałych składowych. Dla każdego z szeregów liczbowych wyznaczony został współczynnik Hursta. Poniżej znajduje się schemat rozważanych metod.



Rysunek 45. Schemat działania prezentowanych metod. Sekwencje najpierw przetwarzane są za pomocą CGR. Następnie, te reprezentacje traktujemy jako obrazy (w przypadku dalszego porównania za pomocą SSIM) lub ciągi liczbowe (DFT i wykładnik Hursta).

63:4700443657

Do budowy drzew wykorzystano metody UPGMA oraz Warda. UPGMA jest szybka oraz wydajna, jednak należy pamiętać o jej ograniczeniach. Zakładamy tutaj ultrametryczność danych (stałe tempo ewolucji) czy własności metryki – symetrię macierzy odległości oraz nierówność trójkąta. O ile macierz odległości między sekwencjami najczęściej jest symetryczna, to założenie nierówności trójkąta może w praktyce nie być spełnione. W tych przypadkach proponowana topologia drzewa może nie być prawidłowa. Z kolei metoda Warda minimalizuje wariancję w obrębie klas. Mimo, że stosunkowo rzadko jest stosowana w filogenetyce to ma swoje istotne zalety. Operuje dobrze na odległościach euklidesowych a takie są wyznaczane między widmami DFT czy wektorami wykładników Hursta, a także może być bardziej odporna na brak ultrametryczności w danych. Stosowana była m.in. w [126,127,128].

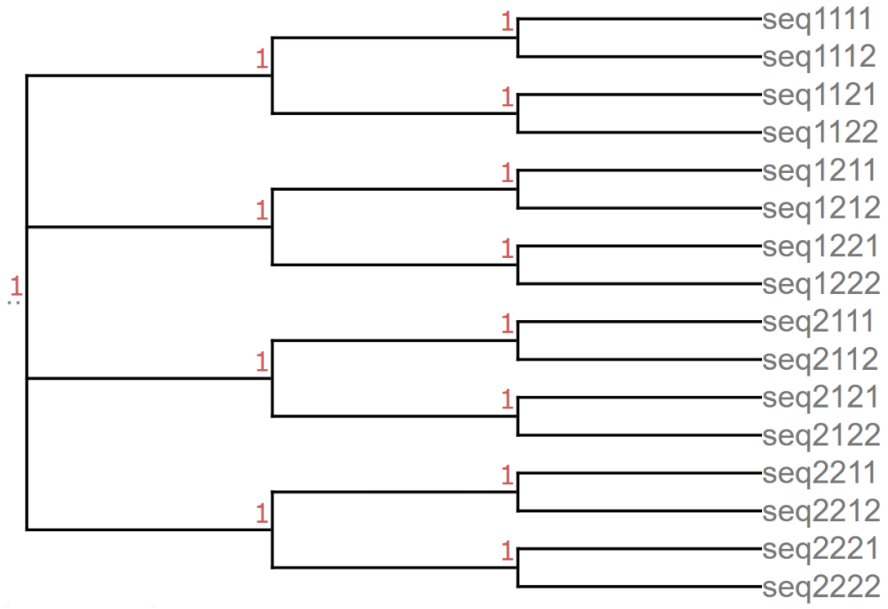
Oprócz metod opartych o CGR wybrano też inną, klasyczną metodę bez dopasowania opartą o złożoność Lempel-Ziv.

Tabela 10. Zestaw różnych metod używanych do generowania drzew filogenetycznych. W tabeli przedstawiono skróty nazw rozważanych metod. Wyróżniono w jaki sposób były analizowane sekwencje oraz jak wygenerowano ostateczne drzewo.

Metoda (Skrót)	Analiza sekwencji	Konstrukcja drzewa
CDU	CGR+DFT	UPGMA
CDW	CGR+DFT	Ward
CHU	CGR+Hurst	UGPMA
CHW	CGR+Hurst	Ward
CSU	CGR+SSIM	UGPMA
CSW	CGR+SSIM	Ward
LZU	Lempel–Ziv	UPGMA
LZW	Lempel–Ziv	Ward

Analizę rozpoczęto od sztucznie wygenerowanych zestawów sekwencji. Zasadniczą zaletą takiego podejścia jest znana historia i relacje ewolucyjne pomiędzy sekwencjami. Rozpoczęto od pojedynczej sekwencji (losowa sekwencja nukleotydowa o zawartości 60% nukleotydów GC lub sekwencja białka z bazy danych), a następnie generowano kolejne pokolenia. Za każdym razem z jednej sekwencji powstawały dwie sekwencje potomne. Po 4 takich generacjach otrzymano rodzinę 16 sekwencji. W rozszerzonej i bliższej rzeczywistej wersji, można by uwzględnić jeszcze dodatkowe motywy, reprezentowane przez zadane częstości dinukleotydów lub inne zależności dalekiego zasięgu. Celem było jednak zbadanie na ile proponowane metody są skalowane (jak działają dla różnych długości sekwencji) oraz jak duży wpływ ma poziom zróżnicowania sekwencji.

64:5368846284



Rysunek 46. Zestaw sekwencji powstałych z jednej, początkowej sekwencji wskutek mutacji. Numeracja wskazuje na kierunek powstania. W formacie Newick można je przedstawić jako (((seq1111,seq1112),(seq1121,seq1122)),((seq1211,seq1212),(seq1221,seq1222)),((seq2111,seq 2112), (seq2121,seq2122)),((seq2211,seq2212),(seq2221,seq2222))) i stanowi ono znaną referencję.

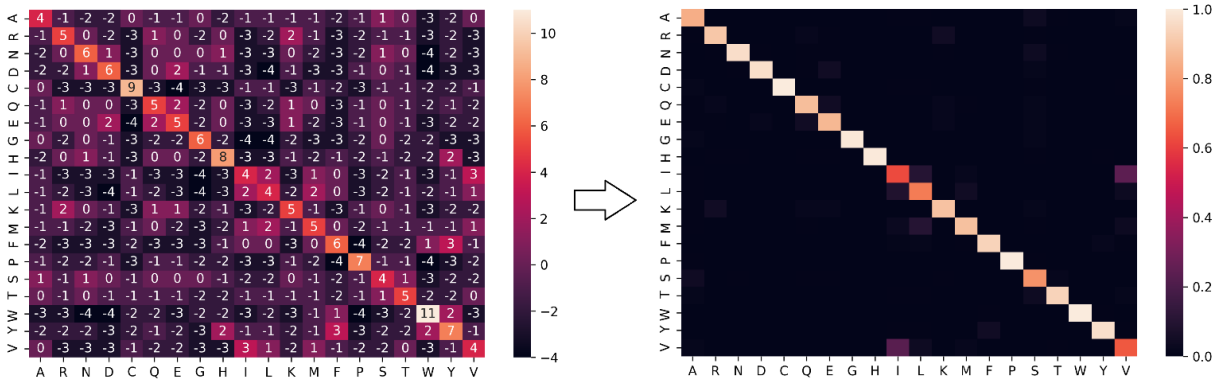
Takich zestawów rozważano kilka, zarówno dla sekwencji nukleotydowych jak i aminokwasowych, z różnym stopniem różnorodności.

Tabela 11. Rozważane zbiory sekwencji w zależności od typu sekwencji, długości sekwencji pierwotnej oraz stopnia podobieństwa kolejnych sekwencji potomnych.

Zbiór	Typ	Długość	Podobieństwo
N3H	nukleotydowa	1K	Wysokie
N3M	nukleotydowa	1K	Średnie
N3L	nukleotydowa	1K	Niskie
N4H	nukleotydowa	10K	Wysokie
N4M	nukleotydowa	10K	Średnie
N4L	nukleotydowa	10K	Niskie
N5H	nukleotydowa	100K	Wysokie
N5M	nukleotydowa	100K	Średnie
N5L	nukleotydowa	100K	Niskie
P393H	białkowa	393	Wysokie
P393M	białkowa	393	Średnie
P393L	białkowa	393	Niskie
P847H	białkowa	847	Wysokie
P847M	białkowa	847	Średnie
P847L	białkowa	847	Niskie
P26926H	białkowa	26926	Wysokie
P26926M	białkowa	26926	Średnie
P26926L	białkowa	26926	Niskie

65:1690785015

W przypadku sekwencji nukleotydowych rozważano początkowe sekwencje o długości 1 tys., 10 tys. oraz 100 tys. Kolejne pokolenia powstawały wskutek zmian zadanych poprzez tempo mutacji. Rozważano trzy prawdopodobieństwa: delecji, insercji oraz substytucji. W przypadku wysokiego podobieństwa, były to odpowiednio 0.1%, 0.1% oraz 0.5%. Dla średniego, 0.5%, 0.5% oraz 1%. Z kolei dla niskiego podobieństwa było to 1%, 1% oraz 10%. Wartości te należy traktować w odniesieniu na pozycję. Rozważając sekwencje aminokwasowe rozpoczęto od sekwencji biologicznych – białka p53 (NCB ID: BAC16799.1), IL-4 (NCBI ID: AAA57193.1) oraz tytyny (NCBI ID: CAA62188.1). Ich długości wynosiły kolejno 393, 847 oraz 26926. Podobnie jak dla sekwencji nukleotydowych wygenerowano sekwencje potomne z różnym stopień podobieństwa. Jednak w tym przypadku wykorzystano macierz BLOSUM62 w przypadku substytucji. Własności fizykochemiczne aminokwasów odgrywają zasadniczą rolę w funkcjonowaniu białka, dlatego zdecydowano się na takie podejście. Oczywiście wartości macierzy BLOSUM62 nie zostały wyznaczone w oparciu o własności fizykochemiczne aminokwasów, jednak te pośrednio są tam uwzględnione poprzez ewolucyjne preferencje substytucji, które wynikają z tych własności. Dlatego wysokie wartości w macierzy często odpowiadają substytucjom między aminokwasami o podobnych cechach. Wartości z macierzy BLOSUM62 zostały przekształcone na prawdopodobieństwa substytucji zgodnie z transformacją SoftMax ($\frac{e^{w_i}}{\sum_j e^{w_j}}$) dla każdego wiersza. Po tej operacji, wszystkie wartości znajdują się w przedziale [0,1] oraz mogą być interpretowane jako prawdopodobieństwa. Wysokie podobieństwo oznaczało, że w 1 na 10 przypadków zamieniamy aminokwas zgodnie z rozkładem (dla danego aminokwasu na kolejnej pozycji) a prawdopodobieństwo insercji/delecji wynosiło 0.1%. Przypadek średni oznaczał taką zmianę w 3 na 10 przypadków, z prawdopodobieństwami insercji/delecji 0.5%. Z kolei niskie podobieństwo oznaczało, że w 5 na 10 przypadków stosujemy rozkład a delecja i insercja występuje z prawdopodobieństwem 1%. Warto tutaj podkreślić, że fakt korzystania z rozkładu nie oznacza bezpośrednio zmianę (nadal najbardziej prawdopodobne jest zachowanie danego aminokwasu, jednak pojawia się tutaj możliwość zmiany).



Rysunek 47. Macierz BLOSUM62 oraz macierz powstała wskutek transformacji SoftMax, przekształcając te wartości na prawdopodobieństwa.

66:4372612501

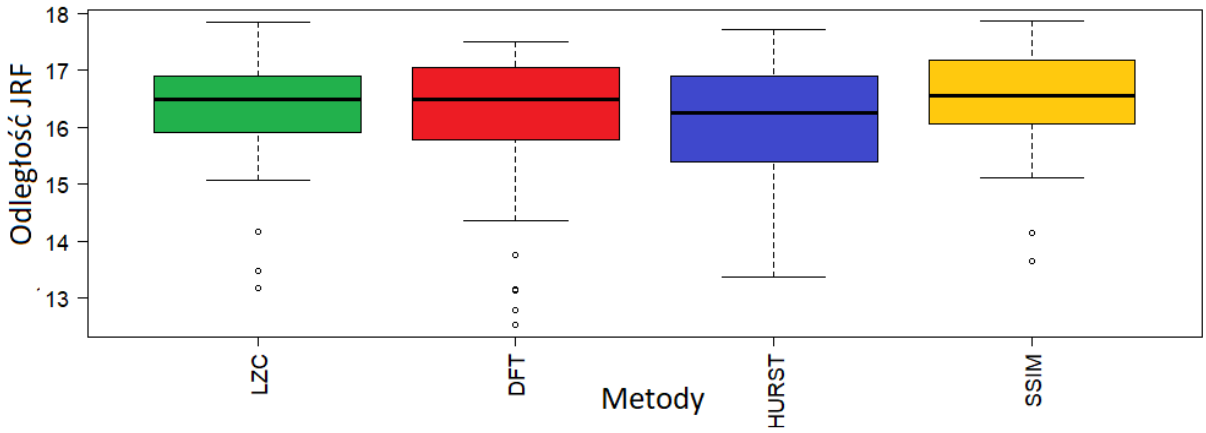
Dla każdego zestawu sekwencji oraz każdej z rozważanych metod wygenerowano drzewo filogenetyczne i porównano jego topologię z referencją za pomocą odległości Jaccarda-Robinsona-Fouldsa (JRF). Poniżej znajdują się otrzymane wartości.

Tabela 12. Nieunormowane odległości JRF dla różnych zestawów sekwencji oraz rozważanych metod. W przypadku sekwencji białkowych rozważano dwa typy CGR – oparty o hydrofobowość/hydrofilowość lub redukcja do 4 liter (oznaczone gwiazdką).

Zbiór/Metoda	CDU	CDW	CHU	CHW	CSU	CSW	LZU	LZW
N3H	6,77	6,17	11,98	11,98	3	2,5	13,04	12,57
N3M	6,53	5,33	12,03	11,32	2	1	12,98	12,94
N3L	5,23	3,4	15,87	15,34	1	1	17,68	16,16
N4H	4,2	4,53	11,7	12,11	1	1	14,83	14,9
N4M	4,5	4	12,32	12,61	1	1	10,47	8,33
N4L	3,17	3	11,34	10,84	1	1	9,89	10,27
N5H	7,37	6,61	7,97	7,38	1	1	9,6	8,41
N5M	6,17	6,93	12,22	12,02	1	1	3	3
N5L	1	3	11,82	11,82	2,67	1	5,94	5,03
P393H	5,17	5,67	13,4	13,66	10,45	11,29	12,24	9,57
P393M	5,83	4,33	15,36	15,36	12,89	13,12	14,25	13,14
P393L	2	2	14,57	14,37	7,5	7,43	10,4	8,06
P847H	9,11	8,6	15,03	14,83	9,54	8,57	13,52	13,13
P847M	7,01	6,87	13,89	14,22	11,62	12,04	1	1
P847L	6,04	3,9	15,87	14,48	10,2	9,72	9,76	10,4
P26926H	10,06	11,07	13,5	13,5	11,37	11,51	5	8,33
P26926M	7,17	7,31	17,6	15,4	10,93	8,13	1	1
P26926L	10,46	10,08	13,41	13,41	10,98	9,81	1	1
P393H*	2,5	3,33	9,71	10,05	4,53	1	-	-
P393M*	5,17	7,27	12,14	12,58	5,81	4,53	-	-
P393L*	6,53	4,67	16,37	16,37	6,44	3	-	-
P847H*	5,33	4,5	14,93	15,53	4,11	4,1	-	-
P847M*	4,36	4,17	13,88	13,36	3	1	-	-
P847L*	5,67	2,5	15,9	15,9	8,75	9,31	-	-
P26926H*	11,79	11,98	12,13	12,41	4,89	4,89	-	-
P26926M*	8,91	9,03	15,91	15,91	2	1	-	
P26926L*	6,4	5,09	14,44	13,7	2	1	-	

Obserwujemy dosyć wysokie wartości. Z powyższej tabeli widać, że najlepiej z rekonstrukcją poradziły sobie metody DFT oraz SSIM. Aby łatwiej zinterpretować te zakresy wartości, wygenerowano losowe topologie drzew w przypadku każdej z metod (dokonano losowej permutacji otrzymanego drzewa).

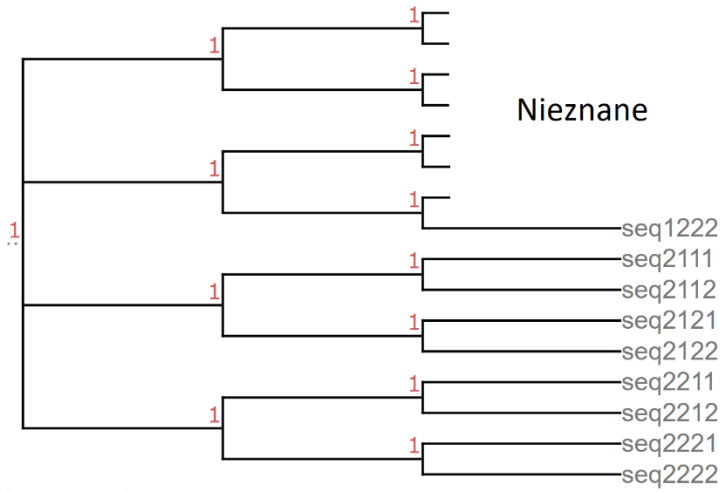
67:8989951186



Rysunek 48. Nieznormalizowane odległości JRF dla losowych topologii drzew.

Niezależnie od metody, średnie wartości odległości JRF wynoszą ok 16.5 z odchyleniem standardowym 1, a wartości maksymalne ok 18. Wracając do poprzedniej tabeli, dla metod opartych o CGR obserwujemy niższe wartości. Objawia się to szczególnie w połączeniu z DFT i SSIM. W przypadku sekwencji białkowych metoda SSIM działa lepiej dla sekwencji nukleotydowych, a w przypadku aminokwasowych, na zredukowanym alfabecie. Jest to niewątpliwie związane z dystrybucją punktów na płaszczyźnie. Ze względu na sposób liczenia cech obrazów, łatwiejszy do analizy jest zapewne kwadratowy niż nieregularny kształt. Najmniej korzystnie wypada współczynnik Hursta. Należy przypomnieć, że w tym przypadku sekwencja zostaje zredukowana do 4 współczynników, co może nie być wystarczające. Dodatkowo, w tym przypadku generowane sekwencje są raczej pozbawione długo dalekozasięgowych wzorców.

Analogicznie obliczenia wykonano dla niepełnego zbioru sekwencji. Usunięto część sekwencji, aby sprawdzić jak wprowadzone metody radzą sobie w takiej sytuacji. W rzeczywistych badaniach nie mamy dostępu do wszystkich sekwencji potomnych. Istotnym wydaje się pytanie czy taka niepełność informacji powoduje utratę skuteczności rozważanych metod. Poniżej przedstawiono schemat drzewa z numerami sekwencji, które usunięto z całej puli.



Rysunek 49. Zestaw sekwencji powstałych z jednej, początkowej sekwencji wskutek mutacji. Numeracja wskazuje na kierunek powstania. Część sekwencji jest nieznana (nieobserwowana). W formacie Newick można je przedstawić jako (((seq1222)),((seq2111,seq2112),(seq2121, seq2122)),((seq2211,seq2212),(seq2221,seq2222))) i stanowi ono referencję.

68:8766171307

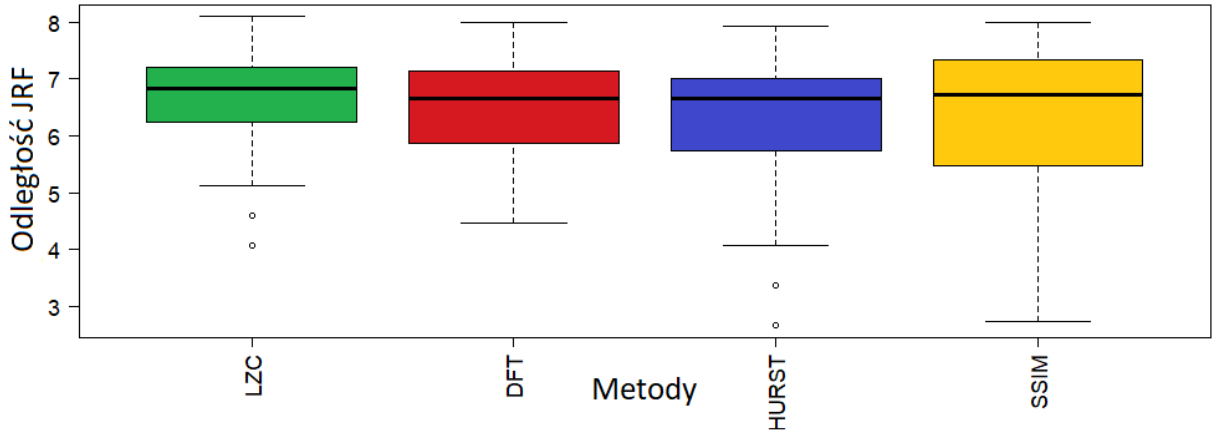
Podobnie jak wcześniej wyznaczono odległości JRF od referencji.

Tabela 13. Nieznormalizowane odległości JRF dla różnych zestawów sekwencji oraz rozważanych metod. W przypadku sekwencji białkowych rozważano dwa typy CGR – oparty o hydrofobowość/hydrofilowość lub redukcja do 4 liter (oznaczone gwiazdką).

Zbiór/Metoda	CDU	CDW	CHU	CHW	CSU	CSW	LZU	LZW
N3H	3,83	4,33	5,23	5,23	1,2	1,2	6	6
N3M	2,13	2,13	5,53	5,53	0	1,2	6,07	6,07
N3L	4,4	3,2	6,63	6,63	0	1,2	5,8	5,67
N4H	2,13	2,13	4,33	5,33	0	0	6,8	5,47
N4M	2,13	2,13	3,87	3,87	0	0	3,33	1,5
N4L	0	2,13	4,93	4,93	0	1,2	7,17	7,17
N5H	5,83	5,33	5,47	5,33	0	0	5,47	5,67
N5M	3,53	3,8	4,53	3,87	0	0	2	2
N5L	1,5	3,1	5,33	5,2	0	1,2	3,2	2
P393H	1,5	0	4,46	4,46	1,5	1,5	7,13	6,6
P393M	3	1,5	5,83	5,67	6,83	6,17	3,5	4,1
P393L	3	3	4,6	4,6	7,33	6,93	4	4,6
P847H	0	0	4,53	3,53	3,6	3,33	7,33	6,67
P847M	2,4	2,4	5,47	5,47	6	5,33	4,47	4,8
P847L	3,67	2	5,47	6	1,5	3	1,5	0
P26926H	3,1	3,6	6,87	6,87	2	2	3,8	2
P26926M	5,47	3,2	2,67	2,67	4	3,57	0	0
P26926L	2,47	2,73	7	7	2,13	1,2	0	0
P393H*	3,8	3,8	5,67	5,8	1,5	1,5	-	-
P393M*	0	1,2	5,33	5,33	0	0	-	-
P393L*	1,5	0	4,83	4,83	2,13	3,03	-	-
P847H*	2	3,2	3,87	2,47	2	2,4	-	-
P847M*	1,2	1,2	5,29	5,29	3,33	3,67	-	-
P847L*	0	0	6	6	1,5	1,5	-	-
P26926H*	3,87	4,13	6,33	6,33	0	0	-	-
P26926M*	3,2	3,7	6,83	6,83	0	0	-	-
P26926L*	2,13	2,13	6,53	6,83	2,67	2,8	-	-

Tutaj także wyznaczono wartości dla losowych topologii.

69:8768880856

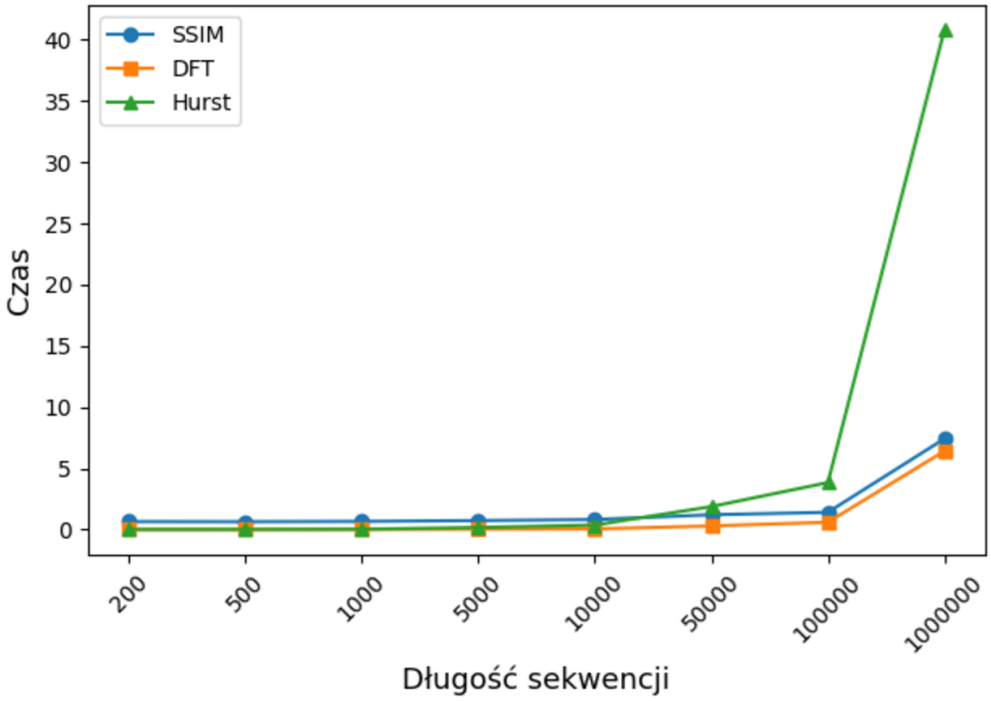


Rysunek 50. Nieznormalizowane odległości JRF dla losowych topologii drzew.

Oczywiście wartości są odpowiednio mniejsze, jednak same tendencje są analogicznie jak w poprzednim przypadku. Wszystkie rozważane metody prowadzą do nielosowych topologii relacji pomiędzy rozważanymi grupami sekwencji. Dotyczy to zarówno sekwencji nukleotydowych jak i aminokwasowych. W wielu przypadkach metoda SSIM doprowadziła do zgodnego z referencją drzewa.

W późniejszej części powyższe metody zostaną jeszcze przetestowane na rzeczywistych przypadkach. Rozważano sekwencje neuraminidazy wirusów grypy u ptaków oraz SPARC z organizmów wyższych. Zanim jednak do tego przejdziemy, spójrzmy jak szybkie są te metody. Rozważano porównanie ze sobą dwóch sekwencji o długości 200, 500, 1k, 5k, 10k, 50k 100k oraz 1 mln. Warto tutaj podkreślić ważną zaletę metod bez dopasowania. W przypadku tradycyjnego porównania sekwencji, dopasowanie wielu z nich nie może być w prosty sposób sprowadzone do porównania każdej z par sekwencji. Związane jest to głównie z możliwością wstawiania przerw, a te występują na innych pozycjach w różnych dopasowaniach. W metodach bez dopasowania sekwencje są przetwarzane na pewne cechy a porównanie sekwencji polega na porównaniu ze sobą tych cech. Z tego powodu, porównanie zbioru sekwencji może być często sprowadzone do porównywania każdej z par. Takich porównań jest $\frac{N(N-1)}{2}$, gdzie N oznacza liczbę sekwencji. Jest to zatem złożoność kwadratowa.

70:1151857333



Rysunek 51. Względne czasy (w sekundach) porównania dwóch sekwencji w zależności od ich długości i rozważanej metody.

Dla sekwencji do 100k dla każdej z metod opartej o CGR porównanie zajmuje od ułamka sekundy do kilku sekund. W przypadku sekwencji o długości 1 milion wyznaczenie wykładnika Hursta wymaga znacznie dłuższego czasu, podczas gdy dla SSIM/DFT to nadal kwestia ok 7 sekund. Samo stworzenie reprezentacji gry chaosu zajmuje ułamek sekundy dla wszystkich rozważanych długości. Należy zwrócić jeszcze uwagę, że w przypadku SSIM tworzone są obrazy dla każdej z sekwencji zatem mamy tutaj także dodatkowe obciążenie pamięciowe. Obliczenia zostały przeprowadzone na komputerze wyposażonym w procesor Intel Core i7-13700K (16 rdzeni, 24 wątki, taktowanie do 5,4 GHz), z architekturą 64-bitową oraz pamięcią podręczną L1 1,4 MiB, L2 24 MiB, L3 30 MiB.

Rozważmy teraz studium przypadku – gen neuraminidazy wśród ptasich wirusów grypy [129]. Ten zbiór sekwencji był wielokrotnie wykorzystywany do oceny nowych metod, m.in. w [35,130].

Tabela 14. Sekwencje neuraminidazy różnych wirusów grypy. Prócz nazwy wirusa podano również skróconą nazwę oraz identyfikator z NCBI.

ID	Wirus	Skrót (grupa)	NCBI ID
0	A/turkey/Ontario/FAV110-4/2009(H1N1)	H1N1v1	HM370969
1	A/mallard/Nova_Scotia/00088/2010(H1N1)	H1N1v2	CY138562
2	A/thick-billed_murre/Canada/1871/2011(H1N1)	H1N1v3	CY149630
3	A/duck/Guangxi/030D/2009(H1N1)	H1N1v4	KC608160
4	A/mallard/France/691/2002(H1N1)	H1N1v5	AM157358
5	A/duck/Hokkaido/w73/2007(H1N1)	H1N1v6	AB470663
6	A/pintail/Miyagi/1472/2008(H1N1)	H1N1v7	AB546159
7	A/mallard/Korea/KNU_YP09/2009(H1N1)	H1N1v8	HQ897966
8	A/mallard/Maryland/352/2002(H1N1)	H1N1v9	EU026046
9	A/mallard/Maryland/26/2003(H1N1)	H1N1v10	FJ357114

71:1117034219

10	A/dunlin/Alaska/44421-660/2008(H1N1)	H1N1v11	GQ411894
11	A/mallard/Minnesota/Sg-00620/2008(H1N1)	H1N1v12	CY140047
12	A/turkey/Virginia/4135/2014(H1N1)	H1N1v13	KM244078
13	A/chicken/Eastern_China/XH222/2008(H5N1)	H5N1v1	HQ185381
14	A/duck/Eastern_China/JS017/2009(H5N1)	H5N1v2	HQ185383
15	A/chicken/Yunnan/chuxiong01/2005(H5N1)	H5N1v3	EU635875
16	A/chicken/Germany/R3234/2007(H5N1)	H5N1v4	FM177121
17	A/domestic_duck/Germany/R1772/2007(H5N1)	H5N1v5	AM914017
18	A/wild_bird/Hong_Kong/07035-1/2011(H5N1)	H5N1v6	KF572435
19	A/Chicken/Hong_Kong/822.1/01(H5N1)	H5N1v7	AF509102
20	A/chicken/Miyazaki/10/2011(H5N1)	H5N1v8	AB684161
21	A/chicken/Korea/es/2003(H5N1)	H5N1v9	EF541464
22	A/mandarin_duck/Korea/K10-483/2010(H5N1)	H5N1v10	JF699677
23	A/turkey/VA/505477-18/2007(H5N1)	H5N1v11	GU186511
24	A/American_black_duck/NB/2538/2007(H7N3)	H7N3v1	EU500854
25	A/American_black_duck/New_Brunswick/02490/2007(H7N3)	H7N3v2	CY129336
26	A/American_green-winged_teal/California/44242-906/2007(H7N3)	H7N3v3	CY076231
27	A/avian/Delaware_Bay/226/2006(H7N3)	H7N3v4	CY039321
28	A/chicken/British_Columbia/GSC_human_B/04(H7N3)	H7N3v5	AY646080
29	A/chicken/Rizhao/713/2013(H7N9)	H7N9v1	KF259734
30	A/chicken/Jiangsu/1021/2013(H7N9)	H7N9v2	KF938945
31	A/duck/Jiangxi/3096/2009(H7N9)	H7N9v3	KF259688
32	A/wild_duck/Korea/SH19-47/2010(H7N9)	H7N9v4	KC609801
33	A/turkey/Minnesota/1/1988(H7N9)	H7N9v5	CY014788
34	A/mallard/Minnesota/AI09-3770/2009(H7N9)	H7N9v6	CY186004
35	A/mallard/Postdam/178-4/1983(H2N2)	H2N2v1	DQ017487
36	A/duck/Hong_Kong/319/1978(H2N2)	H2N2v2	CY005540
37	A/emperor_goose/Alaska/44297-260/2007(H2N2)	H2N2v3	JX081142

Występuje tutaj 13 wirusów z grupy H1N1, 11 z H5N1, 5 z H7N3, 6 z H7N9 oraz 3 z H2N2. Przetestowano wprowadzone wcześniej metody na tym zbiorze sekwencji i sprawdzono czy odpowiednie wirusy są zaklasyfikowane do tego samego kladu (grupy).

Tabela 15. Klasyfikacja wirusów grypy w oparciu o metody CDU i CDW odpowiednio.

Przewidywane/Faktyczne	H1N1	H5N1	H7N3	H7N9	H2N2
H1N1	12|12	0|0	0|0	0|0	0|0
H5N1	1|1	11|11	0|0	0|0	0|0
H7N3	0|0	0|0	5|5	0|0	0|0
H7N9	0|0	0|0	0|0	6|6	0|0
H2N2	0|0	0|0	0|0	0|0	3|3

72:2232554130

Tabela 16. Klasyfikacja wirusów grypy w oparciu o metody CHU i CHW odpowiednio.

Przewidywane/Faktyczne	H1N1	H5N1	H7N3	H7N9	H2N2
H1N1	8|8	0|0	0|0	0|0	1|1
H5N1	5|5	11|11	0|0	0|0	0|0
H7N3	0|0	0|0	5|5	0|0	0|0
H7N9	0|0	0|0	0|0	6|6	0|0
H2N2	0|0	0|0	0|0	0|0	2|2

Tabela 17. Klasyfikacja wirusów grypy w oparciu o metody CSU i CSW odpowiednio.

Przewidywane/Faktyczne	H1N1	H5N1	H7N3	H7N9	H2N2
H1N1	13|13	1|1	0|0	0|0	0|0
H5N1	0|0	10|10	0|0	0|0	0|0
H7N3	0|0	0|0	5|5	0|0	0|0
H7N9	0|0	0|0	0|0	6|6	0|0
H2N2	0|0	0|0	0|0	0|0	3|3

Tabela 18. Klasyfikacja wirusów grypy w oparciu o metody LZU i LZW odpowiednio.

Przewidywane/Faktyczne	H1N1	H5N1	H7N3	H7N9	H2N2
H1N1	10|13	0|3	0|0	0|0	1|0
H5N1	3|0	11|8	0|0	0|0	0|0
H7N3	0|0	0|0	5|5	0|0	1|0
H7N9	0|0	0|0	0|0	5|6	0|1
H2N2	0|0	0|0	0|0	1|0	1|2

Widzimy, że każda z metod dostarczyła w większości poprawnych relacji. Najbardziej dokładne są te oparte o SSIM oraz DFT. Tutaj tylko jeden z wirusów jest sklasyfikowany nieodpowiednio. Musimy jednak wziąć pod uwagę, że zamiana w obrębie grup H1N1 oraz H5N1 jest zrozumiała z powodu tego samego typu genu neuraminidazy (N1).

Sprawdźmy jeszcze na przykładzie relację pomiędzy drzewami zbudowanymi w oparciu o sekwencje nukleotydowe a korespondującymi z nimi sekwencje aminokwasowe z użyciem wprowadzonych metod. W tym celu wybrano mRNA genu SPARC (ang. *Secreted Protein Acidic Rich in Cysteine*) z kilu organizmów wyższych. Białko SPARC pełni rolę regulatora interakcji międzykomórkowych i macierzy pozakomórkowej, wpływając na procesy takie jak migracja komórek, remodelowanie tkanek oraz angiogeneza [131].

73:2681770724

Tabela 19. Rozważane organizmy, dla których rozważano sekwencje nukleotydowe oraz aminokwasowe SPARC z odniesieniami ID do NCBI. Pozostawiono angielskie nazwy.

ID	Organizm	NCBI ID
0	mouse	NM_009242.5
1	rat	XM_032914431.1
2	human	NM_001309443.2
3	chimpanzee	XM_024356783.1
4	dog	XM_014113053.2
5	glassy-fish	XM_028416359.1
6	greater-amberjack	XM_022760346.1

Otrzymane topologie drzew porównano ze sobą za pomocą metryki JRF. Uzyskano

Tabela 20. Porównanie topologii drzew skonstruowanych w oparciu o sekwencje nukleotydowe i aminokwasowe. W formacie Newick przedstawiono topologie obydwu z drzew.

	Format Newick	JRF
LZW	((((greater-amberjack,rat),glassy-fish),(chimpanzee,human)),(dog,mouse));	1.50
	(((((greater-amberjack,rat),glassy-fish),dog),mouse),(chimpanzee,human));	
LZU	((((greater-amberjack,rat),(chimpanzee,human)),glassy-fish),(dog,mouse));	2.20
	(((((greater-amberjack,rat),glassy-fish),dog),mouse),(chimpanzee,human));	
CDW	((((dog,rat),(greater-amberjack,glassy-fish)),mouse),(chimpanzee,human));	3.50
	((((dog,chimpanzee),human),(rat,mouse)),(greater-amberjack,glassy-fish));	
CDU	(((((greater-amberjack,glassy-fish),dog),rat),mouse),(chimpanzee,human));	3.17
	((greater-amberjack,glassy-fish),(((dog,chimpanzee),human),(rat,mouse)));	
CHW	((greater-amberjack,rat),((((chimpanzee,human),glassy-fish),mouse),dog));	4.73
	((((greater-amberjack,human),(rat,mouse)),(dog,chimpanzee)),glassy-fish);	
CHU	((((((chimpanzee,human),glassy-fish),mouse),dog),greater-amberjack),rat);	4.70
	((((greater-amberjack,human),(dog,chimpanzee)),(rat,mouse)),glassy-fish);	
CSW	((((greater-amberjack,glassy-fish),rat),(chimpanzee,human)),(dog,mouse));	1.20
	(((chimpanzee,human),(greater-amberjack,glassy-fish)),((dog,mouse),rat));	
CSU	((((greater-amberjack,glassy-fish),rat),(chimpanzee,human)),(dog,mouse));	1.20
	((((dog,mouse),rat),(greater-amberjack,glassy-fish)),(chimpanzee,human));	

W przypadku SSIM generowane są najbardziej zbliżone topologie. Z kolei dla współczynnika Hursta różnice są najbardziej istotne.

W ramach pracowni magisterskiej (J. Lewandowska, *Poszukiwanie relacji ewolucyjnych w oparciu o reprezentację gry chaosu*, magisterium na bioinformatyce, 2022.) testowane były też inne zbiory sekwencji jak i metody (m.in. k-merów oraz podejście oparte o dopasowanie za pomocą *Clustal Omega*). Metody z dopasowaniem działały istotnie wolniej. Najszybsze były z kolei metody oparte o słowa (metoda k-merów oraz oparta o złożoność Lempel-Ziv). Techniki oparte o CGR plasowały się pomiędzy. Połączenie DFT/CGR wypadało najczęściej lepiej niż Hurst/CGR i w wybranych przypadkach porównywalnie lub lepiej niż Clustal Omega (zwłaszcza w przypadku

74:8918181636

sekwencji o podobnej długości w zakresach od kilkaset do kilka tysięcy). W tych rozważaniach nie uwzględniano jeszcze podejścia CGR/SSIM.

W literaturze pojawiają się wciąż kolejne metody oparte o CGR. W 2024 autorzy wprowadzili tzw. dynamiczny model językowy do porównywania sekwencji [132]. Metoda ta łączy informacje o kontekście i częstości występowania k-merów. Testowano zarówno sekwencje DNA jak i białkowe wirusów. W drugim przypadku, redukując alfabet 20 liter do 4. Policzono odległości Robinsona-Fouldsa, porównując otrzymane drzewa filogenetyczne z referencją. Otrzymano satysfakcjonujące wyniki, w porównaniu z innymi metodami.

4.3.2. Rekonstrukcja relacji ewolucyjnych wśród wybranych genomów referencyjnych dostępnych przez AFproject: wskazanie ograniczeń

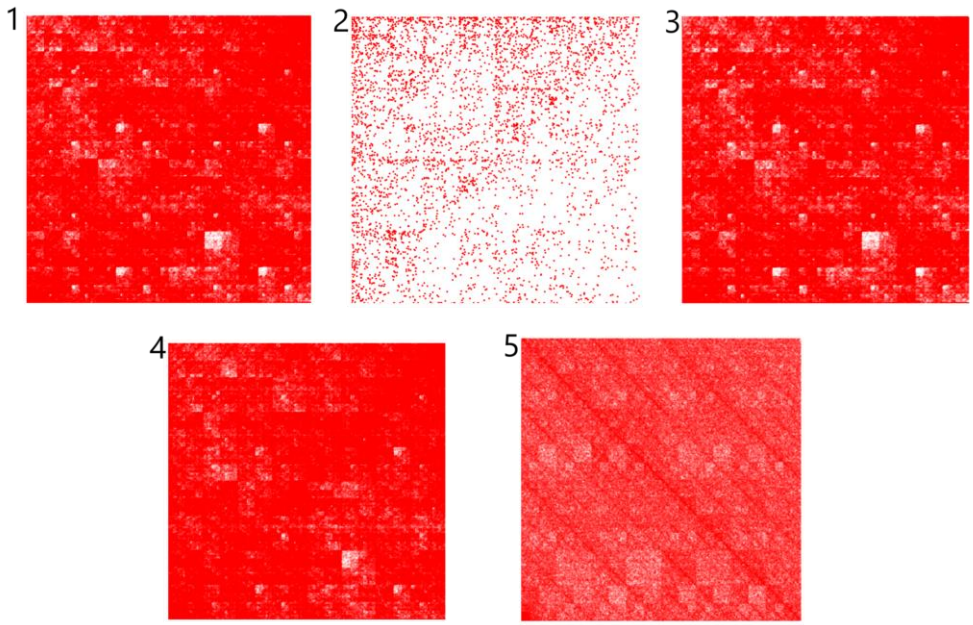
Analiza w poprzednim rozdziale miała na celu ocenę wprowadzonych metod względem siebie oraz referencji, którą utworzono w uproszczony sposób. Liczba rozważanych sekwencji była także stosunkowo niewielka. W tym rozdziale przeanalizowano kilka zbiorów sekwencji dostępnych przez AFproject [58] celem szerszej analizy wprowadzonych metod oraz ich weryfikacji na rzeczywistych przykładach. Wybrano 5 zestawów genomów. W przypadku ostatniego zestawu, został on przygotowany osobno przy różnych poziomach HGT (horyzontralnego transferu genów), określonych przez średnią liczbę zdarzeń HGT na iterację. Rozważano 0, 250, 500, 750 i 1000 zdarzeń.

Tabela 21. Rozważane zestawy referencyjne. W przypadkach 3., 4. oraz 5. rozważany był horyzontalny transfer genów. Ostatni z zestawów wygenerowano za pomocą EvolSimulator i składa się z 5 podzestawów.

ID	Organizm	Horyzontalny Transfer Genów	Źródło:
1	E. coli/Shigella strains	-	Yin and Jin (2013) [133]
2	Fish mtDNA	-	Fisher et al., 2013 [134]
3	E. coli/Shigella strains	+	Bernard et al. 2016 [135]
4	Yersinia strains (yersinia)	+	Bernard et al. 2016 [135]
5	Symulowane genomy (Evolsimulator)	+	Beiko, Charlebois (2007) [136]

Z użyciem reprezentacji gry chaosu zaprezentowano poniżej po jednej z sekwencji z każdego z zestawów. Obserwujemy różne wzorce pomiędzy sekwencjami. Drugi z zestawów zawiera najkrótsze sekwencje co widać po rozkładzie punków (genom mitochondrialny).

75:3340020363



Rysunek 52. CGR dla pierwszej sekwencji z każdego z zestawów. W zależności od długości sekwencji dobrano różny rozmiar kropek przy wizualizacji.

Dane są do pobrania w formacie FASTA. Wykonujemy porównanie sekwencji za pomocą proponowanej metody lokalnie i zwracamy wynik w postaci macierzy odległości (w formacie Phylip [137]) lub topologii wygenerowanego drzewa (w formacie Newick [138]). W pierwszym przypadku stosowany jest algorytm NJ (ang. *Neighbor-Joining*) a utworzone drzewo jest porównywane z referencją. Zwracana jest m.in. odległość Robinsona-Fouldsa (także w wersji unormowanej). Pobrano wspomniane zestawy sekwencji, w trzech z nich występuje horyzontalny transfer genów. Wyniki rozważanych metod opartych o CGR wraz z krótką charakterystyką dotyczącą liczby sekwencji i długości sekwencji przedstawiono w tabeli poniżej.

Tabela 22. Zestawienie podstawowych informacji podsumowujących zestawy takich jak liczba sekwencji czy średnia długość sekwencji w zestawie. Ostatnie trzy kolumny zawierają wartość unormowanej odległości Robinsona-Fouldsa (nRF). Dla ostatniego zestawu podano wynik średni z podzestawów.

Zestaw	L. sekwencji	Min. Długość	Max. Długość	Średnia długość	DFT	SSIM	Hurst
1	29	4.4 mln	5.5 mln	4.9 mln	0.73	0.54	-
2	25	16.4k	17k	16.6k	0.41	0.09	0.86
3	27	4.4 mln	5.5 mln	4.9 mln	0.79	0.54	-
4	8	4.5 mln	4.7 mln	4.6 mln	1	0.20	-
5	33	1.8 mln	2.6 mln	2.2 mln	0.78*	0.47*	-

Długość sekwencji jest ważna nie tylko z powodu złożoności czasowej, ale i pamięciowej. Sekwencje istotnie różne w długościach w obrębie zestawu mogą też być bardziej wymagające do analizy w przypadku metod szczególnie na to czułych (jak DFT). Z kolei w przypadku metody SSIM może być wymagana większa rozdzielczość w przypadku dłuższych sekwencji.

Podsumowując otrzymane wyniki dla zastosowanych technik należy wskazać na metodę opartą na SSIM za najbardziej obiecującą. Jest to konsystentne z wynikami z poprzedniego rozdziału.

76:7231205105

Wypada ona lepiej niż DFT w każdym z zestawów. Szczególnie satysfakcjonujący rezultat otrzymano dla porównań genomu mitochondrialnego. Prawdopodobnie związane jest to z możliwością uchwycenia wystarczającej liczby szczegółów na obrazie. W przypadku dłuższych sekwencji obrazy są gęściej wypełnione, a kropki nachodzą na siebie. W szczególności, parametr struktury, który mierzy podobieństwo lokalnych struktur może nie być w stanie uchwycić subtelnych różnic. Testowano inne rozdzielczości, ale nie przyniosło to istotnych zmian. W tym przypadku warto być może rozważyć metodę MSSIM (ang. *Multi-Scale SSIM*), operującą w różnych skalach [34]. W przypadku drzew skonstruowanych w oparciu o DFT obserwujemy pewne rzeczywiste relacje, jednak na potrzeby praktycznych zastosowań poziom różnic ze zbiorem referencyjnym wydaje się zbyt duży. Najlepszy wynik uzyskano również dla genomu mitochondrialnego. Według autora spowodowane jest to dwoma czynnikami: relatywnie krótkie sekwencje w zestawie oraz porównywalna długość sekwencji. Istotną wadą podejścia z DFT jest konieczność wydłużania widm do długości najdłuższej sekwencji. To zapewne negatywnie wpływa na ostateczne porównanie. Należałoby w przyszłości rozważyć jeszcze inne podejścia; być może zamiast porównywania całych widm warto byłoby zebrać pewne ich cechy i porównanie sekwencji sprowadzić do porównania ich ze sobą. W przypadku metody Hursta, wynik otrzymano jedynie dla sekwencji mitochondrialnych, ale i w tym przypadku okazał się istotnie dalece niesatysfakcjonujący (w porównaniu z analogicznymi wynikami dla SSIM czy nawet DFT). Redukcja sekwencji do 4 liczb powoduje ogromną utratę informacji, zwłaszcza dla dłuższych sekwencji. Autor uważa, że ten współczynnik nadal ma potencjał, jednak równocześnie ograniczone zastosowanie. Może być używany jako jedna z wielu charakterystyk do oceny podobieństwa sekwencji. W przypadku niewielkich różnic (drobne mutacje), wykładnik Hursta nie jest w stanie w poprawny sposób opisać te zmienności. Dla dłuższych sekwencji występowały problemy związane z implementacją i stabilnością wyznaczania tego wykładnika.

4.4. Możliwe są rozszerzenia CGR dla DNA, RNA oraz białek, zachowujące najważniejsze własności oryginalnej reprezentacji, ale wykazujące szerszy zakres potencjalnych zastosowań

Przedstawione zostaną teraz rozszerzenia klasycznej reprezentacji gry chaosu (CGR), jakie poczynił autor [111,139].

4.4.1. Przypadek DNA

Przypomnijmy raz jeszcze, że klasyczna reprezentacja DNA z użyciem CGR polega na umieszczeniu nukleotydów w wierzchołkach kwadratu. Startujemy z jego centrum, a każdy kolejny punkt znajduje się w połowie drogi między odpowiednim wierzchołkiem kwadratu a aktualnym położeniem. Bardziej formalnie

$$(x_{n+1}, y_{n+1}) = \frac{1}{2}\big(N_{n+1} + (x_n, y_n)\big), \tag{59}$$

gdzie N_{n+1} symbolizuje nukleotyd na $n+1$ pozycji. Działając rekurencyjnie, możemy rozpisać powyższą formułę

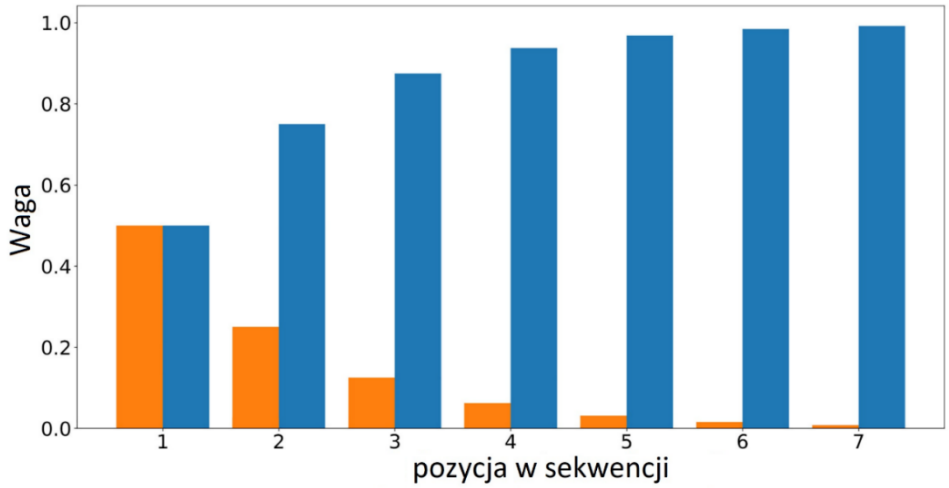
$$(x_{n+1}, y_{n+1}) = \frac{1}{2}\left(N_{n+1} + \frac{1}{2}\big(N_n + (x_{n-1}, y_{n-1})\big) =\right) = \frac{1}{2}N_{n+1} + \frac{1}{4}N_n + \frac{1}{4}(x_{n-1}, y_{n-1}). \tag{60}$$

77:8860908901

Postępując analogicznie otrzymujemy bezpośredni wzór na aktualne położenie punktu

$$(x_{n+1}, y_{n+1}) = \sum_{i=1}^{n+1} \frac{1}{2^{n-i+2}} N_i + \frac{1}{2^{n+1}} (x_0, y_0) \,. \tag{61}$$

Widać z niego, że aktualny punkt zależy oczywiście od wszystkich poprzednich punktów oraz od punktu startowego. Zależność od punktu startowego ma coraz mniejszy wpływ dla dalszych pozycji w sekwencji. Z kolei udział poprzednich reszt na aktualną pozycję maleje jako kolejna potęga liczby ½; aktualna pozycja ma udział ½, poprzednia ¼ itd.



Rysunek 53. Wpływ reszt w sekwencji (niebieski) oraz punktu startowego (pomarańczowy). Widzimy, że dla kolejnych pozycji wpływ punktu startowego maleje bardzo szybko. Rysunek opublikowano w Genomics w 2021.

Tabela 23. Wpływ wszystkich dotychczasowych nukleotydów oraz punktu startowego na wartość w danym punkcie. Tabelę opublikowano w Genomics w 2021.

Liczba nukleotydów	Skumulowany efekt reszt	Wpływ punktu startowego
1	0.50	0.50
2	0.75	0.25
3	0.875	0.125
4	0.9375	0.0625
5	0.96875	0.03125
6	0.984375	0.015625
7	0.9921875	0.0078125

Klasyczny CGR używa funkcji $a_n = \frac{1}{2^n}$ do ważenia. Warto jednak zastanowić się nad innymi ciągami o racjonalnych własnościach. Może się to przyczynić do lepszej reprezentacji sekwencji biologicznych o czym wspomnę w dalszej części. Autor proponuje trzy warunki jaki powinien spełniać ciąg a_n aby mógł być użyty do konstrukcji CGR:

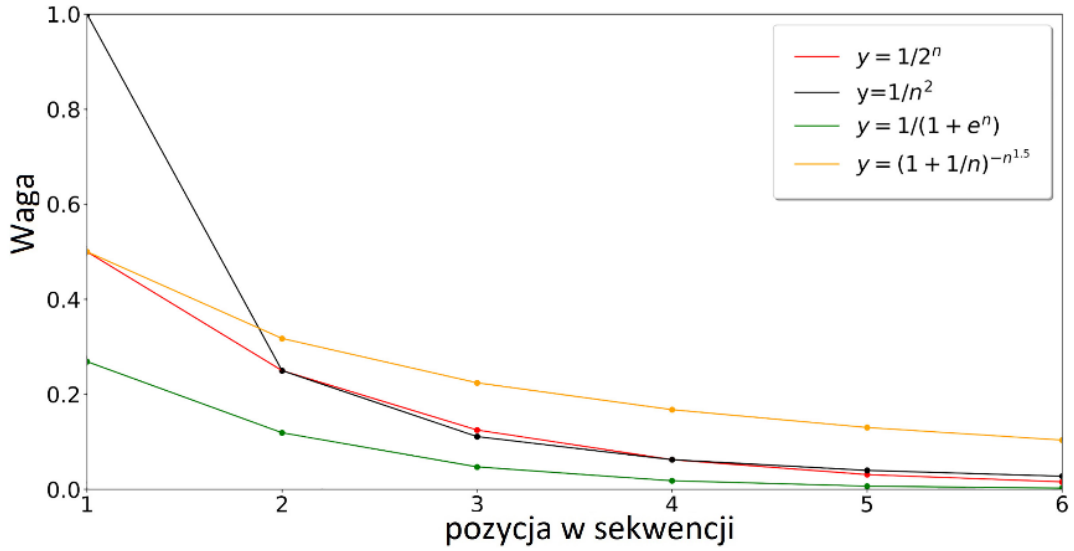
1) $a_n > 0$ dla każdego $n \in N$ (dodatniość),
2) $a_{n+1} < a_n$ (monotoniczność),
3) $\sum_{n=1}^{\infty} a_n < \infty$ (zbieżność).

78:3105526586

Pierwszy warunek gwarantuje, że każda poprzednia reszta ma dodatni wpływ na aktualną pozycję. Drugi prowadzi do tego, że im dalsza odległość w sekwencji względem rozważanej pozycji, tym mniejszy wpływ na jej aktualną wartość. Trzeci warunek zapewnia zbieżność i stabilność zakresu wartości, zwłaszcza dla dłuższych sekwencji. Wtedy, współrzędne kolejnych punktów wyznaczane są wg

$$(x_{n+1}, y_{n+1}) = \frac{\sum_{i=1}^{n+1} a_{n-i+2} N_i + a_{n+2}(x_0, y_0)}{\sum_{i=1}^{n+1} a_{n-i+2} + a_{n+2}}. \tag{62}$$

Czynnik w mianowniku ma na celu normalizację wartości. Spójrzmy teraz na kilka przykładów. Ciągami, które spełniają podane własności jest oczywiście oryginalny $\frac{1}{2^n}$ oraz np. $\frac{1}{n^2}, \frac{1}{1+e^n}$ czy $\left(1+\frac{1}{n}\right)^{-n^{1.5}}$.

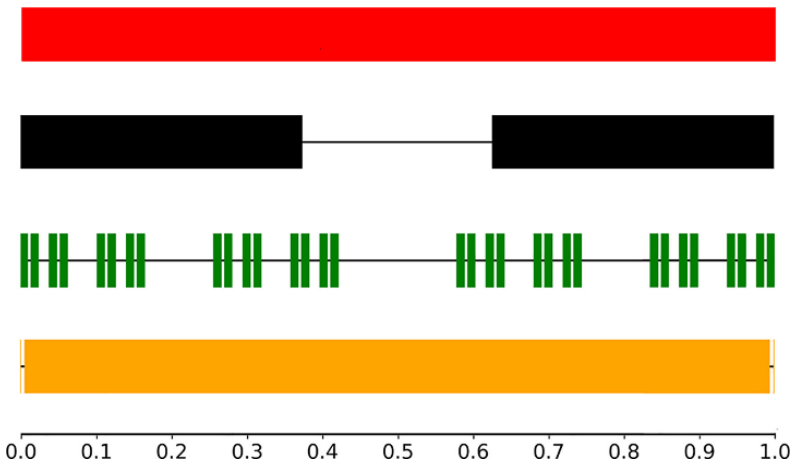


Rysunek 54. Waga obecnej i poprzednich pozycji dla różnych funkcji ważących. Przykładowo, 1 na osi OX oznacza tutaj poprzednią resztę względem aktualnej. Rysunek opublikowano w Genomics w 2021.

Druga z funkcji podnosi wpływ aktualnej pozycji, pozostawiając poprzednie na podobnym poziomie. Trzecia bardzo szybko maleje, co w praktyce oznacza, że pozycja 5 i wcześniejsze mają bardzo mały i zaniedbywalny wpływ na aktualną wartość. Ostatnia z funkcji startuje z tego samego poziomu co oryginalna, jednak jej dynamika jest dużo wolniejsza. Nawet bardzo odległe nukleotydy w sekwencji mają nadal istotny wpływ na aktualną pozycję. Modyfikując zatem funkcję ważącą możemy manipulować poziomem wpływu poprzednich i aktualnej pozycji na końcową reprezentację.

Wybór innej niż klasyczna funkcji ważącej prowadzi oczywiście też do odmiennych rozkładów punktów.

79:4194365213



Rysunek 55. Jednowymiarowe wykresy ukazujące rozkład punktów dla rozważanych funkcji ważących. u samej góry znajduje się oryginalna funkcja $\frac{1}{2^n}$. Rysunek opublikowano w Genomics w 2021.

O ile w przypadku klasycznej funkcji wiążącej możliwe jest równomierne wypełnienie przestrzeni, tak już np. w przypadku drugiej funkcji środkowa część jest nieosiągalna. Jeszcze ciekawszy efekt pojawia się w przypadku trzeciej funkcji. Tutaj występuje swego rodzaju wzór fraktalny. Ostatnia funkcja koncentruje punkty bliżej centrum, końce (ich okolice) w praktyce są nieosiągalne.

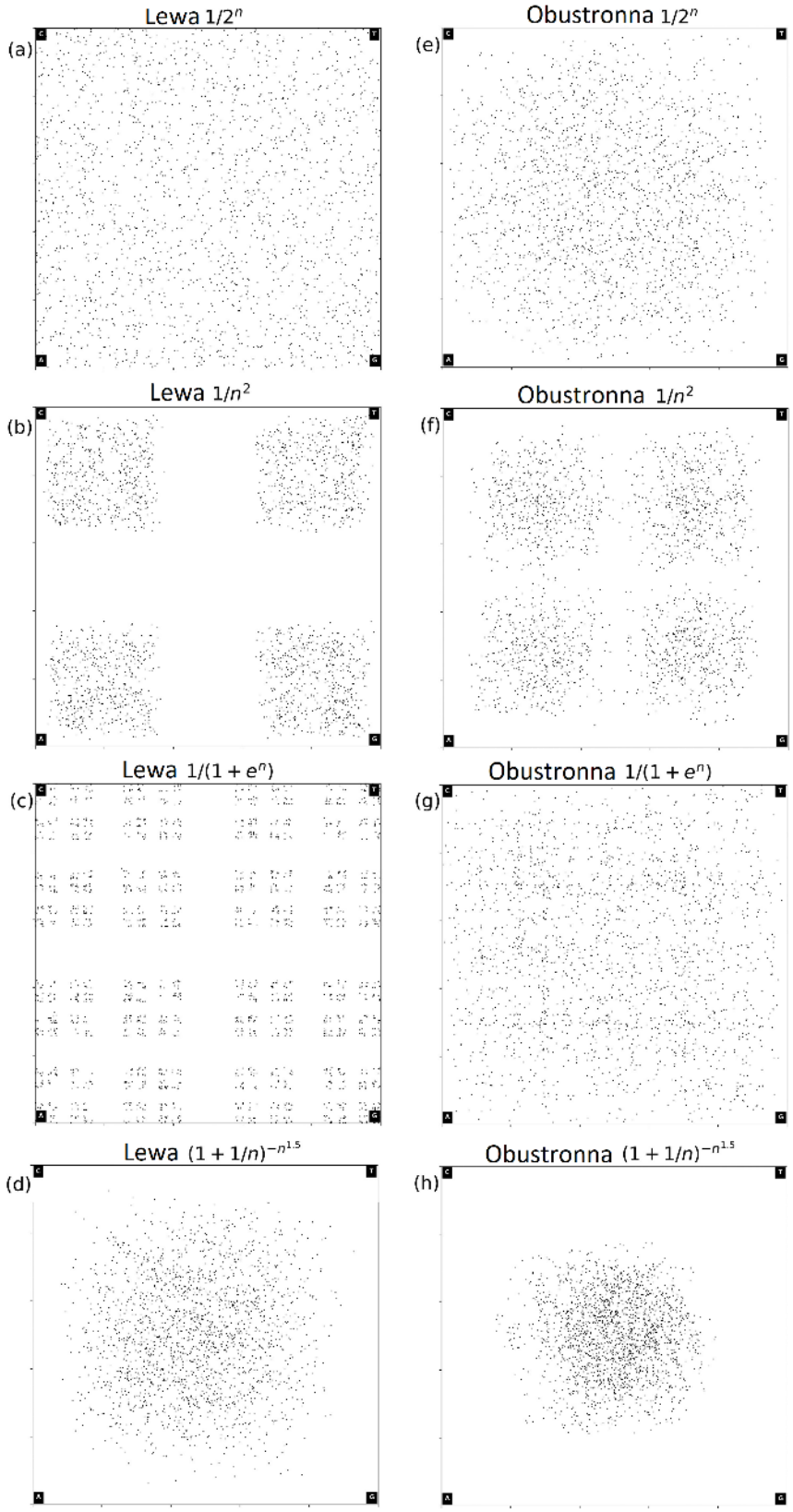
Prócz wpływu poprzednich reszt na aktualną wartość danego punktu można również uwzględnić te dalsze, wtedy wyrażenie na kolejny punkt prezentuje się następująco

$$(x_{n+1}, y_{n+1}) = \frac{K a_1 N_{n+1} + \sum_{i=1}^{n} a_{n-i+2} N_i + a_{n+2}(x_0, y_0) + \sum_{i=n+2}^{N} a_{i-n} N_i}{K a_1 + \sum_{i=1}^{n+1} a_{n-i+2} + a_{n+2} + \sum_{i=n+2}^{N} a_{i-n}}. \tag{63}$$

Wyszczególniono tutaj też aktualną resztę N_{n+1} i dodano parametr K dzięki któremu można by wzmocnić (lub osłabić) jej wpływ. Domyślnie $K = 1$. Mimo, że taka reprezentacja dwustronna ma kilka zalet, a najważniejszą jest uwzględnienie nie tylko przeszłych, ale i przyszłych reszt, nie jest pozbawiona pewnych negatywnych konsekwencji. Po pierwsze tracimy częściowo informację o kierunku i jednoznaczność. Przykładowo, dla tyminy w …AAA**T**GGG… oraz …GGG**T**AAA… otrzymamy te same wartości. Sekwencje biologiczne są zorientowane. Nie musi to jednak być problemem np. dla sekwencji aminokwasowych; tam istotną rolę odgrywają fizykochemiczne właściwości aminokwasów, a zarówno aminokwasy przed i po względem danej pozycji w sekwencji wydają się równie istotne.

Wygenerowano losową sekwencję o długości 2000, aby sprawdzić jej reprezentacje CGR. W przypadku klasycznym punkty równomiernie wypełniają wnętrze kwadratu. Kolejne funkcje ważące znacząco zmieniają rozkład punktów, prowadząc do obszarów nieosiągalnych lub większej koncentracji w centrum (w przypadku ostatniej funkcji). Po zastosowaniu reprezentacji obustronnej obserwujemy koncentrację punktów bliżej centrum. W przypadku funkcji $f(x) = \frac{1}{1+e^x}$ następuje rozmycie punktów. Może się to wiązać ze wzrostem niejednoznaczności i trudniejszym rozróżnieniem punktów względem pozycji w sekwencji.

80:3681545982



Rysunek 56. Przedstawienie losowej sekwencji o długości 2000 w reprezentacji CGR przy różnych funkcji ważących. Wygenerowano jednostronne (lewe) reprezentacje oraz te uwzględniające zarówno poprzednie jak i kolejne reszty (obustronne). Rysunek opublikowano w Genomics w 2021.

81:8219842183

Warto jeszcze przyjrzeć się kilku istotnym właściwościom omówionej wcześniej rodzinom reprezentacji. Przy wprowadzaniu nowych funkcji ważących wprowadzono 3 pożądane warunki jakie powinny zostać spełnione, aby uznać je za użyteczne. Nie gwarantują one jednak jednoznaczności reprezentacji (bijekcji pomiędzy reprezentacją a oryginalną sekwencją). Oczywiście taka jednoznaczność w większości przypadków funkcji dwustronnej nie będzie występować, jednak warto rozważyć co w przypadku lewostronnym. Weźmy dla przykładu ciąg $\{\frac{1}{2}, \frac{1}{4}, \frac{1.2}{8}, \frac{0.8}{8}, \frac{1}{16}, \frac{1}{32}, \dots\}$. Spełnia on oczywiście warunki 1-3, jednak prowadzi do tej samej reprezentacji ostatniej reszty w sekwencjach AATA oraz TTAA. Przypuśćmy, że mamy dwie sekwencje $s_1 = N_1 N_2 \dots N_n$ oraz $s_2 = N_1{}' N_2{}' \dots N_n{}'$. Przypuśćmy też, że dwa ostatnie punkty mają te same reprezentacje. Oznacza to, że

$$\sum_{i=1}^{n} a_{n-i+1} N_i = \sum_{i=1}^{n} a_{n-i+1} N_i{}', \tag{64}$$

czyli

$$\sum_{i=1}^{n} a_{n-i+1} N_i - \sum_{i=1}^{n} a_{n-i+1} N_i{}' = 0\,, \tag{65}$$

i dalej

$$\sum_{i=1}^{n} a_{n-i+1} (N_i - N_i{}') = 0. \tag{66}$$

Dla kwadratu jednostkowego, gdzie wierzchołki znajdują się w punktach $(0,0), (0,1), (1,0)$ oraz $(1,1)$ mamy $N_i - N_i' \in \{-1,0,1\}$. a zatem powyższe rozumowanie można podsumować następująco:

Funkcja wagowa $a_n$nie prowadzi do jednoznacznej reprezentacji, wtedy i tylko wtedy, gdy istnieją dwa rozłączne podzbiory indeksów $\{i_1, i_2, \dots., i_l\}, \{j_1, .j_2, \dots, j_k\}$ takie, że

$$a_{i_1} + a_{i_2} + \dots + a_{i_l} = a_{j_1} + a_{j_2} + \dots + a_{j_k}. \tag{67}$$

Mimo, że warunek wydaje się całkiem prosty, to w praktyce byłby bardzo trudny do sprawdzenia. Można jednak podać prostszy warunek wystarczający. Ciąg jest malejący zatem istnieje indeks $i^* \in \{i_1, i_2, \dots., i_l, j_1, .j_2, \dots, j_k\}$, dla którego $a_{i^*} = \max\{a_{i_1}, a_{i_2}, \dots, a_{i_l}, a_{j_1}, a_{j_2,\dots,} a_{j_k}\}$. Bez straty ogólności możemy założyć, że $i^* = i_1, i_1 = \min\{i_1, i_2, \dots., i_l, j_1, .j_2, \dots, j_k\}$. Gdyby spełniony był warunek $a_{i_1} > a_{j_1} + a_{j_2} + \dots + a_{j_k}$, to w oczywisty sposób powyższe równanie nie mogłoby być spełnione. Możemy podsumować nasze rozważania w postaci twierdzenia

Aby zapewnić jednoznaczną reprezentację wystarczy, że dla każdej liczby naturalnej k dla $n > k$ zachodzi $a_k > \sum_{i=k+1}^{n} a_i$ lub $a_k \geq \sum_{i=k+1}^{\infty} a_i$. Dla klasycznej funkcji wagowej $a_k = 2^{-k}$ oraz $\sum_{i=k+1}^{n} a_i = 2^{-k} - 2^{-n}$, czyli $\sum_{i=k+1}^{\infty} a_i = 2^{-k}$ więc warunek jest oczywiście spełniony. To samo kryterium może być zastosowane dla funkcji $a_k = \frac{1}{1+e^k}$; tutaj jednak trzeba wspomóc się obliczeniami komputerowymi. W przypadku dwóch pozostałych funkcji kryterium nie umożliwia rozstrzygnięcie problemu jednoznaczności.

82:8081710545

Na zakończenie przyjrzyjmy się jeszcze stabilności reprezentacji CGR. Podobne sekwencje powinny charakteryzować się podobnymi reprezentacjami, co oznacza, że jeżeli $||s_1 - s_2||$ przyjmuje niską wartość, to również $\left|\left| \left(x_n^{s_1}, y_n^{s_1}\right) - \left(x_n^{s_2}, y_n^{s_2}\right)\right|\right|$ będzie niewielkie. Mamy

$$\left|\left| \left(x_n^{s_1}, y_n^{s_1}\right) - \left(x_n^{s_2}, y_n^{s_2}\right)\right|\right| = \left|\left|\frac{\sum_{i=1}^{n} \tau_i a_{n-i+2}(N_i - N_i')}{\sum_{i=1}^{n} a_{n-i+2} + a_{n+2}}\right|\right|. \tag{68}$$

.

Zgodnie z nierównością Cauchy'ego-Schwarza mamy dalej

$$\left|\left| \left(x_n^{s_1}, y_n^{s_1}\right) - \left(x_n^{s_2}, y_n^{s_2}\right)\right|\right|^2 \leq \frac{\sum_{i=1}^{n} \tau_i a_{n-i+2}^2}{(\sum_{i=1}^{n+1} a_i)^2} \sum_{i=1}^{n} ||N_i - N_i'||^2 , \tag{69}$$

czyli

$$\left|\left| \left(x_n^{s_1}, y_n^{s_1}\right) - \left(x_n^{s_2}, y_n^{s_2}\right)\right|\right| \leq \sqrt{\frac{\sum_{i=1}^{n} \tau_i a_{n-i+2}^2}{(\sum_{i=1}^{n+1} a_i)^2}} \sum_{i=1}^{n} ||N_i - N_i'|| , \tag{70}$$

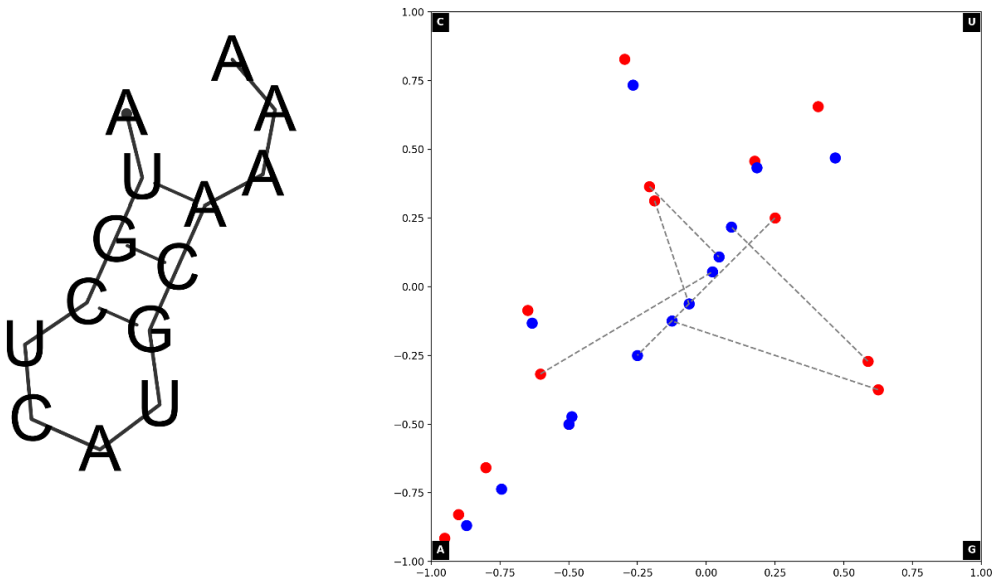
gdzie τ_i jest funkcją równą 0, jeżeli na i-tej pozycji znajduje się ten sam nukleotyd w obu sekwencjach oraz 1 w pozostałych przypadkach. Widzimy więc, że różnica $\left|\left| \left(x_n^{s_1}, y_n^{s_1}\right) - \left(x_n^{s_2}, y_n^{s_2}\right)\right|\right|$ może być rozważana jako efekt dwóch składowych – parametryzacji nukleotydów oraz funkcji ważących.

W ostatnim rozdziale tej pracy zostanie sprawdzone na ile wprowadzone tutaj funkcje ważące wpływają na generowanie drzewa filogenetyczne oraz w jakim stopniu są „odporne" na mutacje.

4.4.2. Przypadek RNA

Oczywiście klasyczny CGR może być również stosowany w przypadku sekwencji RNA. Alfabet w tym przypadku również składa się z 4 liter, a jedyna zmiana polega na zamianie parametryzacji T na U. Autor zaproponował modyfikację klasycznej CGR, która uwzględnia równocześnie strukturę drugorzędową [139]. Zostały wprowadzone również pewne miary oceniające. W przypadku sekwencji RNA, jej struktura odgrywa kluczową rolę w pełnionej przez nią funkcji. Uwzględnienie tej informacji wydaje się lepiej odzwierciedlać jej naturę. Standardowo, umieszczamy nukleotydy w wierzchołkach kwadratu. Przyjęto, że $A(-1,-1), C(-1,1), U(1,1)$ oraz $G(1,-1)$. Zauważmy, że komplementarne nukleotydy znajdują się naprzeciw siebie. Środek kwadratu znajduje się w punkcie $(0,0)$. Konstrukcja ciągu za pomocą CGR odbywa się w sposób standardowy dla niesparowanych nukleotydów. W przypadku tych które tworzą wiązanie, udajemy się nie w stronę odpowiedniego wierzchołka a środka kwadratu. Tę zmodyfikowaną wersję CGR będę dalej nazywał CG-RNA.

83:4289996694



Rysunek 57. Dla sekwencji AUGCUCAUGCAAAA przedstawiono jej strukturę drugorzędową oraz reprezentacje CGR (czerwony) oraz CG-RNA (niebieski). W formacie Vienna [140] strukturę drugorzędową można przedstawić tutaj jako .(((....)))..., gdzie nawias oznacza sparowany nukleotyd. Sparowane nukleotydy przesunięte są bliżej centrum w przypadku CG-RNA, co zostało zaznaczone przerywanymi liniami. Rysunek opublikowano w Comput Biol Med. w 2022.

Sposób konstrukcji powoduje przesunięcie sparowanych nukleotydów ku centrum względem klasycznego CGR. Zdefiniujmy teraz dwie wielkości naśladujące energię fizyczną cząsteczki

$$E_1 = \frac{1}{N}\sum_{i=1}^{N}(x_i'^2 + y_i'^2), \tag{71}$$

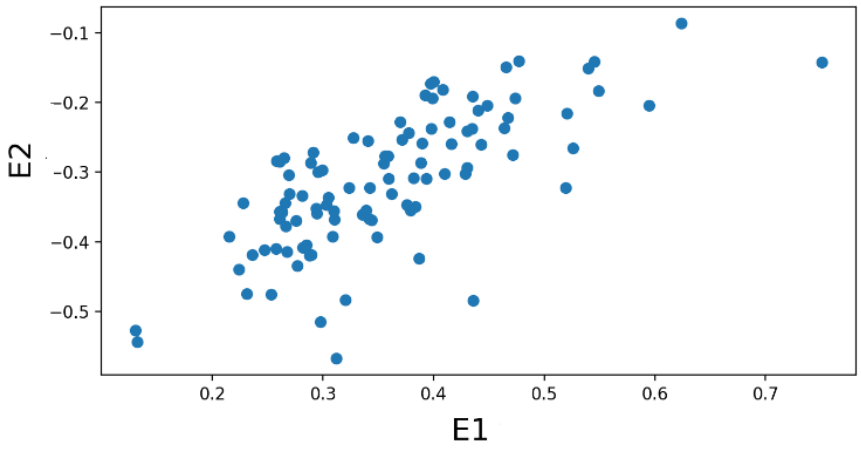
oraz

$$E_2 = -\frac{1}{N}\sum_{i=1}^{N}((x_i - x_i')^2 + (y_i - y_i')^2), \tag{72}$$

gdzie N oznacza długość sekwencji, (x_i, y_i) pary punktów z CGR, a (x_i', y_i') pary punktów z CG-RNA. Przyjrzyjmy się ich własnościom. W przypadku E_1 uwzględniona jest jedynie reprezentacja CG-RNA. Wielkość ta przyjmuje z definicji wartości nieujemne, jako że sumujemy nieujemne składniki. Po drugie, jej wartość wynosi zero tylko w przypadku, gdy wszystkie nukleotydy są sparowane. Licząc E_2 mierzymy różnicę pomiędzy obiema reprezentacjami. W przypadku braku obecności parowania któregokolwiek z nukleotydów reprezentacje CGR oraz CG-RNA pokrywają się i $E_2 = 0$. Parujące się nukleotydy zaburzają reprezentację i skutkujące tym różnice we współrzędnych powodują wzrost w różnicy CGR i CG-RNA (E_2 staje się ujemna). W dalszej części przyjrzymy się własnościom obu miar. Pożądanym byłoby np. żeby kolejne sparowane nukleotydy miały mniejszą energię. Dla przykładu, rozważając sekwencję AAAGGGGUUUUUUUU i potencjalne struktury ((((....))))... oraz ((((....)..).).), pierwsza powinna być bardziej preferowana. Kolejne sparowane nukleotydy będą układać się w CG-RNA coraz to bliżej centrum.

84:4968903311

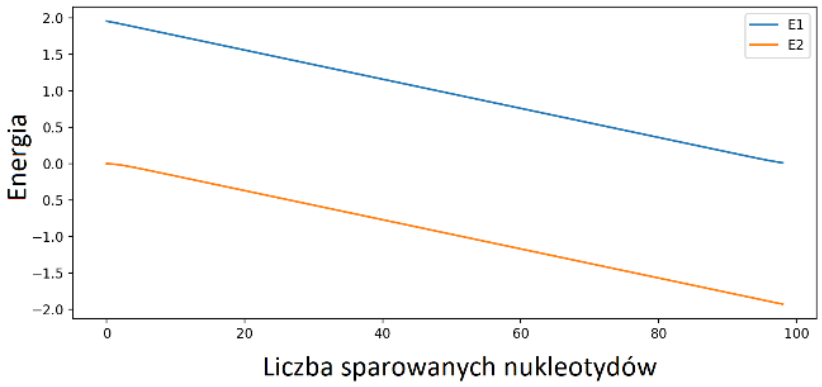
Rozpoczniemy od zbadania właściwości E_1 oraz E_2. Wygenerowano 100 losowych sekwencji RNA o różnym zakresie długości, a następnie dokonano predykcji ich struktury za pomocą algorytmu MFE (ang. *Minimum Free Energy*) [140,141,142] przy użyciu serwera RNAfold. Następnie, wyznaczono wartości E_1 oraz E_2.



Rysunek 58. Zależność pomiędzy E1 i E2 dla 100 losowo wygenerowanych sekwencji. Struktury drugorzędowe zostały przewidziane z użyciem algorytmu MFE. Rysunek opublikowano w Comput Biol Med. w 2022.

Obserwujemy umiarkowaną zależność liniową pomiędzy tymi zmiennymi. Współczynnik korelacji wynosi tutaj 0.73. Co najbardziej istotne, przyjmuje on dodatnią wartość, co potwierdza nasze wcześniejsze rozważania, że zachowana jest tutaj monotoniczność. Brak ścisłej zależności jest tutaj zaletą, bowiem oznacza to, że obie wielkości w istotnie różny sposób mierzą wpływ struktury drugorzędowej.

Rozważmy teraz sekwencję $s = \underbrace{AAA \ldots AAA}_{100\,nt}\underbrace{TTT \ldots TTT}_{100\,nt}$. Sprawdzimy jaki jest związek pomiędzy energią (E_1 i E_2) a liczbą sparowanych nukleotydów. W tym celu zaczniemy od struktury (… … …) każdorazowo ją zagnieżdżając, czyli przez ((… … …)), (((… … …))) aż do struktury zawierającej 4 niesparowane nukleotydy w pętli.

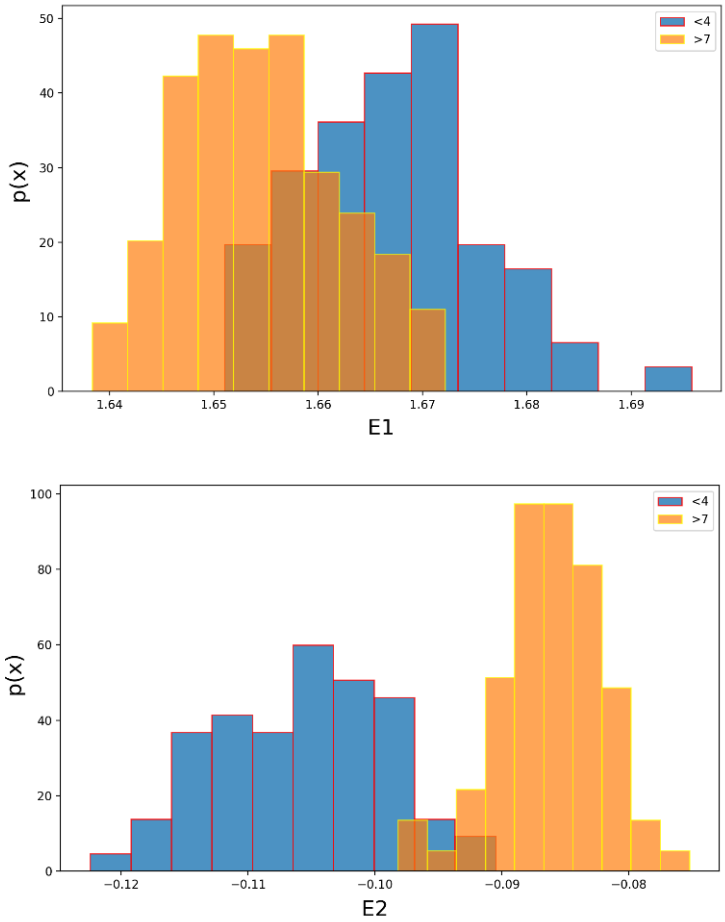


Rysunek 59. Zależność pomiędzy E1, E2 a ilością sparowanych nukleotydów dla kolejnych zagnieżdżonych struktur sekwencji zawierającej ciąg powtórzeń A oraz T. Rysunek opublikowano w Comput Biol Med. w 2022.

85:6886007697

Obserwujemy tutaj bardzo proste zależności. Okazuje się bowiem, że w obu przypadkach istnieje liniowa korelacja pomiędzy ilością sparowanych nukleotydów a wartościami E_1 i E_2. Graniczne wartości dla tych wielkości wynoszą odpowiednio 0 i -2. Widoczna zależność jest oczekiwana, bo kolejne sparowane nukleotydy stabilizują strukturę drugorzędową.

Co istotne, zarówno E_1 jak i E_2 zależą nie tylko od liczby sparowanych nukleotydów, ale i także od ich dystrybucji w sekwencji. W następnym kroku wygenerowano 200k losowych struktur dla sekwencji $s = \underbrace{AAA \dots AAA}_{100\,nt}\underbrace{TTT \dots TTT}_{100\,nt}$ gdzie 10 losowych par nukleotydów było sparowanych. Zgrupowano te struktury w dwa podzbiory. Pierwszy składał się z tych struktur, gdzie maksymalna odległość pomiędzy sparowanymi nukleotydami była mniejsza niż 4. W drugim zbiorze mamy te struktury, w których analogiczna odległość była większa niż 7.



Rysunek 60. Rozkłady energii E1 oraz E2 dla różnych struktur. Struktury podzielone zostały na te z dużą przerwą (>7) oraz małą (<4). Rysunek opublikowano w Comput Biol Med. w 2022.

Co ciekawe, obserwujemy tutaj przeciwne relacje. w przypadku E_1, niższe wartości energii są związane z wyższymi odległościami pomiędzy sparowanymi nukleotydami. Dzieje się tak ponieważ E_1 zależało tylko od CG-RNA i bardziej rozrzucone nukleotydy w sekwencji powodują wyższą koncentrację wszystkich punktów blisko centrum. Z kolei E_2 mierzy różnicę pomiędzy CG-RNA a CGR i kolejno sparowane nukleotydy powodują wyższe różnice pomiędzy korespondującymi punktami z obu reprezentacji. Pozostałe punkty są mniej podatne na zmianę położenia. Wydaje się, że E_2 lepiej odzwierciedla biologiczne i chemiczne właściwości kwasów nukleinowych.

86:1117877515

Na zakończenie rozważmy jeszcze przykład biologiczny, a dokładniej strukturę drugorzędową 3' końca RNA wirusa mozaiki lucerny 3 (AIMV-3) [143]. Wiele algorytmów do predykcji struktury drugorzędowej opiera swoje działanie głównie na liczbie sparowanych nukleotydów. Niektóre z nich uwzględniają również tzw. pseudowęzły, które tworzą dodatkowe pary zasad między odległymi fragmentami, co powoduje skrzyżowanie się różnych regionów i komplikuje strukturę. Jednakże, nie zawsze taka sama optymalizacja dokonuje się w komórce. Co więcej, zazwyczaj różne algorytmy prowadzą do różnych struktur. Problem jest wciąż otwarty i nowe narzędzia ciągle powstają lub są udoskonalane. Zastosujemy tutaj E_1 oraz E_2 do oceny wygenerowanych struktur. Zaczniemy od predykcji za pomocą algorytmu MFE, gdzie narzucono warunek, że sparowanych ma być 12 lub odpowiednio 9 par nukleotydów.

Tabela 24. Struktury dla ALMV-3 wygenerowane za pomocą algorytmu MFE oraz odpowiednie oceny E1 i E2 tych struktur. Narzucony został warunek parowania 12 lub 9 par nukleotydów. W pierwszym przypadku, średnie energie wynoszą 0.24 (±0.01) i -0.38 (±0.01) odpowiednio. W drugim przypadku, są to już 0.33 (±0.03) oraz -0.29 (±0.03). Tabelę opublikowano w Comput Biol Med. w 2022.

Liczba sparowanych	Struktura drugorzędowa	E_1	E_2
12	..(((((((((....)))))))..))(((...)))....	0.30	-0.39
12	....(((((((....)))))))((.(((....)))))..	0.28	-0.37
12	....(((((((....)))))))..(((((...)))..))	0.30	-0.39
12	..(((((((((....)))))))...(((....)))..))	0.29	-0.38
12	.((.(((((((....))))))).))(((....)))....	0.28	-0.37
12	..(((((((((....)))))))....(((...)))..))	0.30	-0.37
12	....(((((((....))))))).(((((....))).)).	0.28	-0.37
9	....(((((.....)))))..((((....))...))	0.37	-0.27
9	....((..((.....))..))((.(((....)))))..	0.39	-0.32
9	.....((((((....)))....))).(((...)))....	0.35	-0.30
9	.....(((.....(((....)))(((...))).)))	0.35	-0.33
9	((.....((((....))))...))..(((...)))....	0.35	-0.31
9	.(((....))).....((((.....((.....))))))	0.31	-0.24
9	.((..(((((.....)))))..)).((.....))....	0.37	-0.26
9	..((.((.(((....))).))...)).((...).....	0.41	-0.28
9	..((.((.((.....)).)).....(((...)))..))	0.45	-0.32
9	.....((((.....))))..((..(((...)))))..	0.36	-0.28

Zgodnie z oczekiwaniami, większa liczba sparowanych nukleotydów powoduje mniejsze wartości E_1 i E_2. Następnie porównano strukturę biologiczną (natywną) ze strukturą wygenerowaną z MFE [140,141,142] oraz za pomocą metody centroidów [144].

87:1798606918

Tabela 25. Wartości E_1 oraz E_2 dla struktur ALMV-3. Porównano strukturę biologiczną (referencja) oraz przewidywania za pomocą MFE oraz metody centroidów. Zauważmy, że MFE i referencją są porównywalne. Co ciekawe, jeżeli znormalizujemy wartości E2 (tzn. kolejne wartości zostaną jeszcze podzielone przez współrzędne CGR, oznaczone gwiazdką), to taka wielkość ma najniższą wartość dla referencji. Tabelę opublikowano w Comput Biol Med. w 2022.

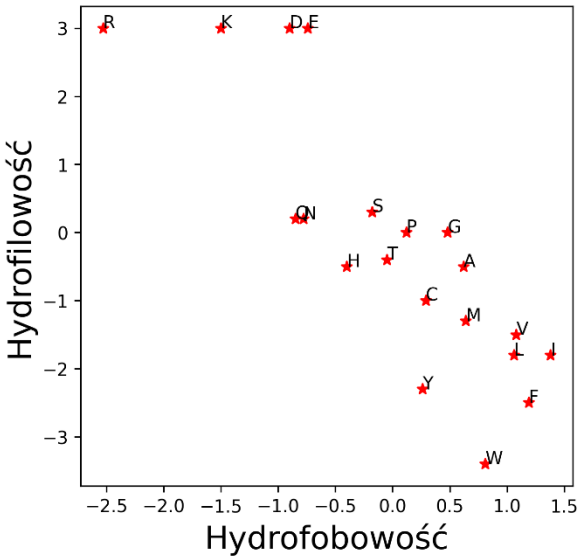
Metoda	Liczba sparowanych	Struktura drugorzędowa	E_1	E_2	E_{2*}
referencja	10	....(((((((....)))))))....(((...))))....	0.26	-0.33	-49.81
MFE	10	....(((((((....)))))))...(((....))))....	0.27	-0.33	-46.96
centroidy	7	....(((((((....))))))).....	0.46	-0.15	-39.43

Co ciekawe, struktura referencyjna ma najmniejsze wartości E_1 oraz E_2 w tym przypadku. Oznacza to, że być może wielkości te mogłyby być używano jako walidacja do istniejących już metod do predykcji struktur drugorzędowych. Pomysł uwzględnienia struktury drugorzędowej w CGR był już rozważany [145]. Autorzy we wspomnianej pracy zaproponowali reprezentację 3D jako z jednej strony metodę wizualizacji i szybkiego porównania wybranych struktur bez dopasowania, a z drugiej jeszcze inny sposób kodowania tej informacji. W tej pracy autor przedstawia reprezentację dwuwymiarową, co jest szczególnie ważne, jeżeli chcemy zastosować wprowadzone podejścia porównywania sekwencji dedykowane klasycznej reprezentacji gry chaosu; w tym momencie jest to jej proste uogólnienie. Warto poświęcić jej uwagę w przyszłych badaniach.

4.4.3. Przypadek białek

W przypadku sekwencji aminokwasowych sytuacja jest nieco trudniejsza [139]. Jest to spowodowane 20 literami w alfabecie. Mimo, że naturalnym wydaje się uogólnienie kwadratu do dwudziestościanu foremnego o którym wspomniano w rozdziale dotyczącym metod, to pojawiają się tutaj problemy natury technicznej i interpretacyjnej. O ile w przypadku nukleotydów arbitralne przypisanie do wierzchołków nie wydaje się odgrywać znaczącej roli, tak dla aminokwasów o różnych własnościach fizykochemicznych należałoby zachować ich relacje w przestrzeni 2D. Najprostsze podejście polega na podziale aminokwasów w 4 grupy: niepolarne AILMFPWV, ujemne polarne DE, nienaładowane polarne NCQGSTY, dodatnie polarne RHK i zastosowaniu klasycznej wersji CGR [17]. W tym przypadku następuje jednak istotna redukcja informacji, a same grupy są niezbilansowane. Innym z rozwiązań jest wybranie dwóch własności charakteryzującej aminokwasy i umiejscowienie ich wg nich. Naturalne wydają się tutaj hydrofobowość czy hydrofilowość. Właściwości te zostały wykorzystane m.in. w [146] a pochodzą z [147,148]. Mimo, że oczywiście są one silnie powiązane (ujemna korelacja), to odmienny sposób wyznaczenia powoduje, że ujmują one nieco inne cechy aminokwasów (patrz Rys. 60). Te cechy wybrano też przy konstrukcji drzew filogenetycznych w rozdziale, w którym testowano zastosowanie metod opartych o CGR do analizy filogenetycznej.

88:2246340525

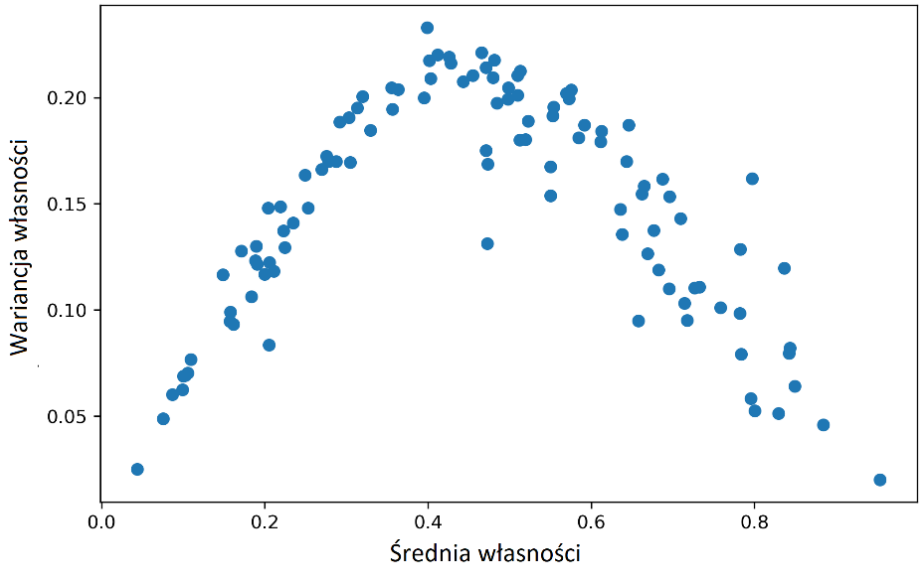


Rysunek 61. Hydrofobowość i hydrofilowość aminokwasów. Współrzędne punktu startowego określają odpowiednie średnie danej własności. Rysunek opublikowano w Comput Biol Med. w 2022.

W tym przypadku, kształt reprezentacji nie będzie już przyjmował postać figury foremnej, mimo to taka parametryzacja aminokwasów prowadziła do nielosowych i częściowo poprawnych relacji ewolucyjnych w przypadku omawianych wcześniej metod (w szczególności DFT).

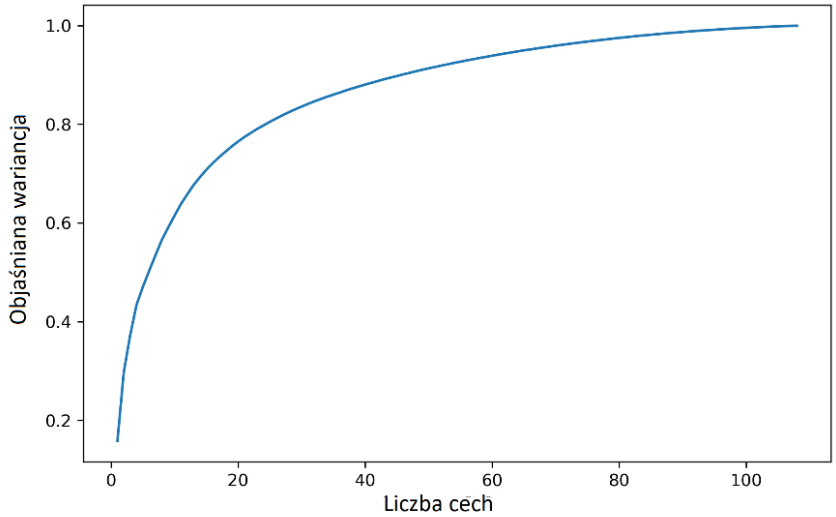
Poddano analizie również inne własności fizykochemiczne aminokwasów na rodzinach białek w bazie PROSITE [149]. Znajduje się tam 2162 rodzin i każda z nich składa się z sekwencji dzielących ze sobą ten sam motyw. Dodatkowo, znajduje się również informacja o dopasowaniu (MSA). Zostało sprawdzone które własności fizykochemiczne są w największym stopniu wykorzystywane przez kolejne rodziny białek. W tym celu wykorzystano kolejną bazę danych AAindex [150,151,152], gdzie znajduje się 566 własności aminokwasów. Na początku zredukowano tę liczbę do 109. Niektóre z własności są redundantne, więc na podstawie wyznaczonego współczynnika korelacji Pearsona ustalono próg na 0.7 i na tej podstawie dokonano redukcji. Zastosowano podejście zachłanne, tzn. usuwano cechy iteracyjnie. W kolejnym kroku, każda z 109 cech została znormalizowana do zakresu <0,1> przez transformacje min-max. Początkowe wartości były z różnych zakresów, co mocno mogłoby wpłynąć na wyniki. Następnie, dla każdej rodziny i każdej kolumny MSA, wyznaczono średnią wartość dla każdej z 109 własności. Następnie, wybrano wartości ekstremalne. Przez ekstremalne, rozumie się najbardziej odległe od średniej dla danej cechy. Zastosowano takie podejście, ponieważ celem było wskazanie, które własności aminokwasów są nadreprezentowane lub unikane. Miejsca aktywne białek bardzo często posiadają pozycje, gdzie kluczową rolę odgrywają właśnie pewne własności aminokwasów. Dla każdej ze 109 ekstremalnych wartości wyznaczono zależność pomiędzy średnią a wariancją.

89:4380743560



Rysunek 62. Zależność pomiędzy wariancją a średnią dla kolejnych własności dla rozważanych rodzin białkowych. Rysunek opublikowano w Comput Biol Med. w 2022.

Co ciekawe, obserwujemy tutaj kwadratową zależność z maksimum przy 0.5. Średnie znajdujące się blisko 0 i 1 wskazują, że dana cecha była raczej faworyzowana lub omijana przez większość rodzin białek. W tych przypadkach obserwujemy niskie wartości wariancji. Z drugiej strony, średnie blisko 0.5 wskazują, że cecha była częściowo omijana, a częściowo faworyzowana. Następnie, zastosowano algorytm zmniejszania wymiaru danych PCA (ang. *Principal Component Analysis*) [153] do cech, aby sprawdzić czy wytypowane 109 można jeszcze bardziej zredukować.

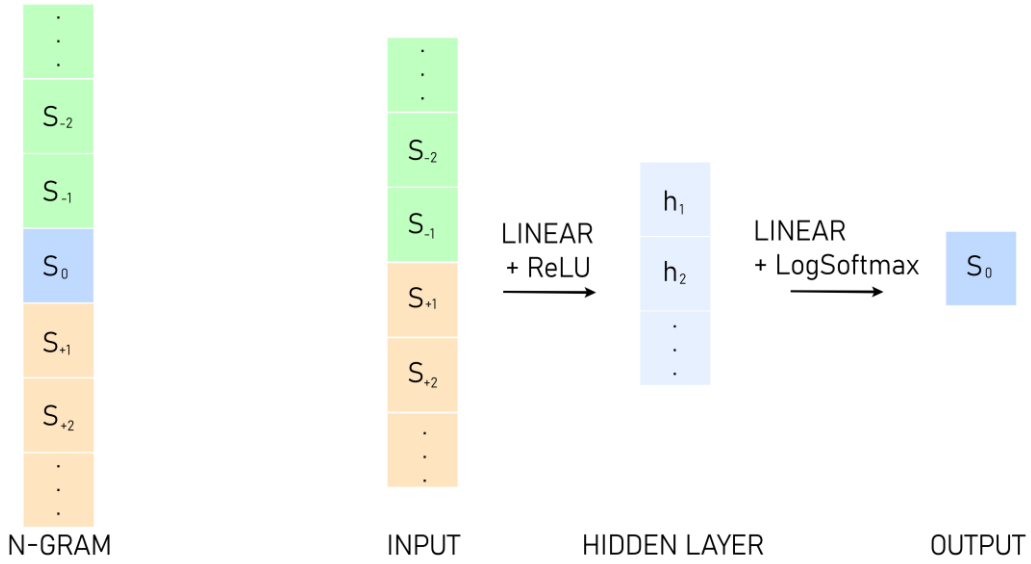


Rysunek 63. Zależność pomiędzy objaśnianą wariancją a pozostawioną liczbą cech. Rysunek opublikowano w Comput Biol Med. w 2022.

Jak widać na wykresie, wzrost objaśnianej wariancji z liczbą cech następuje bardzo powoli. Oznacza to, że redukcja do kilku/kilkunastu cech nie jest już możliwa. Około 50 cech jest wymaganych do pozostawienia 90% wariancji z danych. To oznacza, że cechy te są ważne i w pewnym sensie nieredukowalne.

90:4517733388

Powyższe obliczenia, chociaż pozornie nie związane z CGR miały na celu pokazanie, że chociaż wybór hydrofobowości czy hydrofilowości dostarcza dobrej reprezentacji aminokwasów, to istnieje obszerny zbiór innych cech, które mogą odgrywać także istotą rolę w niektórych rodzinach białkowych. Sprawdzanie każdej kombinacji jest niepraktyczne i dlatego postanowiono znaleźć inne rozwiązanie dla tego problemu. Reprezentacja aminokwasów miałaby być ustalana każdorazowo dla danej grupy sekwencji na podstawie technik przetwarzania języka naturalnego. Sekwencja aminokwasowa traktowana jest jak zdanie, a aminokwasy jak słowa. Budowana jest specjalna sieć neuronowa, celem nauczenia reprezentacji dla słów – tzw. wektorów osadzeń (ang. *word embedding*) [154]. Są to wektory o zadanym wymiarze (tutaj będzie to 2), mające tę własność, że słowa o podobnym znaczeniu powinny charakteryzować się podobnymi wektorami (np. w sensie odległości euklidesowej czy cosinusowej). Przedstawmy teraz schemat oraz proces uczenia takiej sieci. Rozważana grupa sekwencji dzielona jest na mniejsze części, tzw. n-gramy (zasadniczo jest to tożsame w tym kontekście z k-merami). Zadając parametr n dzielimy wszystkie sekwencje na mniejsze nakładające się ciągłe podsekwencje o długości n. Następnie mając do dyspozycji takie podsekwencje uczymy sieć rozpoznawania środkowego znaku na podstawie sąsiednich. Przykładowo, dla $n = 5$, aminokwas środkowy (trzeci) jest przewidywany w oparciu o pierwszy, drugi, czwarty oraz piąty aminokwas. Celem sieci jest z jednej strony predykcja żądanego aminokwasu (zagadnienie klasyfikacji), ale przy okazji uczone są też parametry reprezentujące same aminokwasy. Podsumowując, prócz typowych wag sieci, dodatkowo uczone są również reprezentacje (w tym przypadku wektory 2D) dla aminokwasów.



Rysunek 64. Sieć neuronowa do predykcji środkowego aminokwasu a także do nauki reprezentacji słów. Podana sieć składa się z jednej warstwy ukrytej (ang. hidden layer). Jako funkcje aktywacji wybrano ReLU oraz LogSoftmax. Rysunek opublikowano w Comput Biol Med. w 2022.

Powyższa prosta sieć składająca się tylko z jednej warstwy ukrytej może być oczywiście rozbudowana wg potrzeb. Należy jednak pamiętać, że liczba parametrów do nauczenia tej sieci obejmuje nie tylko same wagi sieci, ale i reprezentacje. Co więcej, twierdzenie Cybenko wskazuje, że jedna warstwa ukryta z odpowiednią liczbą neuronów jest wystarczająca (przynajmniej teoretycznie) do aproksymacji dowolnej funkcji ciągłej [155]. Testowano różne liczby neuronów

91:8785823031

(1, 5 oraz 10) w warstwie ukrytej i już 1 dostarczał satysfakcjonujących rezultatów (zwłaszcza dla niewielkiej liczby sekwencji). Należy tutaj wybrać możliwie małą (ale efektywną) sieć, która z jednej strony będzie szybko trenowana a z drugiej odporna na przetrenowanie. Sieć zaimplementowano w Python z użyciem biblioteki PyTorch. Krok uczenia został ustawiony na 0.001, jako funkcję kosztu wybrano NLLLoss (ang. *Negative Log Likelihood Loss*), a optymalizator SGD (ang. *Stochastic Gradient Descent*). Liczba epok 100 była wystarczająca do ustabilizowania funkcji kosztu.

A zatem, mając do dyspozycji zbiór sekwencji, dzieląc je na mniejsze podsekwencje i ucząc sieć przewidywania centralnego aminokwasu możemy przy okazji nauczyć się w jaki sposób powinny być reprezentowane te aminokwasy. Reprezentacje te oczywiście nie określają konkretnych właśności fizykochemicznych, ale bez wątpienia są ich pochodnymi. Co najważniejsze, aminokwasy występujące w podobnym otoczeniu będą miały podobne reprezentacje. Pamiętajmy, że mowa o przypisanych współrzędnych wierzchołków dla kolejnych aminokwasów. Przy późniejszym przetwarzaniu sekwencji na CGR stosowane są omówione wcześniej transformacje. Ze względu na zastosowanie wektorów osadzenia w tej technice, będę ją dalej oznaczał jako emb-CGR.

Porównajmy ją teraz ze standardowym CGR oraz techniką PseAAC. Wszystkie te techniki dostarczają wartości raczej łatwych w interpretacji i oparte są o proste formuły. W każdej z nich możemy też uwzględnić informację o kolejności występowania aminokwasów. W PseAAC jest to sterowane parametrem lambda. Można wskazać poziom uwzględnionych korelacji. PseAAC naturalnie też bierze pod uwagę pewne cechy aminokwasów, jednak muszą one być zadane na początku. W technice emb-CGR są one dostosowywane do każdej grupy sekwencji i nie stanowią prostego podzbioru znanych własności, chociaż z pewnością są ich kombinacją. Spowodowane jest to chociażby funkcyjną konserwatywnością białek. Jednakże mogą być one bardziej złożone. Nie muszą być natomiast arbitralnie zadawane – są dostosowywane do listy zadanych sekwencji. Dalsze generowanie reprezentacji jest już deterministyczne i w łatwy sposób interpretowalne.

Tabela 26. Porównanie techniki emb-CGR z klasycznym CGR oraz PseAAC. Tabelę opublikowano w Comput Biol Med. w 2022.

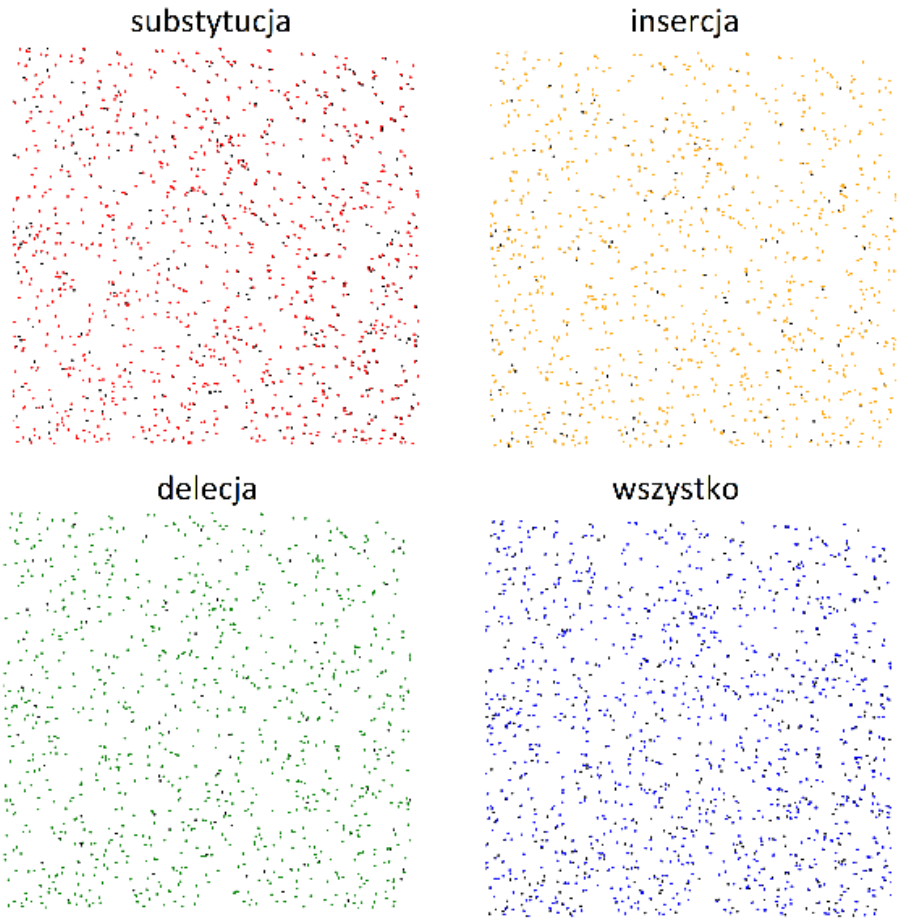
Cecha	emb-CGR	CGR	PseAAC
1. Łatwość interpretacji	TAK	TAK	TAK
2. Informacja o kolejności	TAK	TAK	CZĘŚCIOWE
3. Uwzględnienie własności fizykochemicznych	CZĘŚCIOWE	CZĘŚCIOWE	TAK
4. Dostosowane (dynamiczne) cechy	TAK	NIE	NIE

Istotną zaletą stosowania zarówno CGR jak i emb-CGR jest fakt, że otrzymujemy ciąg par punktów, które mogą być interpretowane jako liczby zespolone, a to umożliwia ich dalsze przetwarzanie, np. przez Dyskretną Transformatę Fouriera. W następnym rozdziale zostanie ukazany wpływ takiego dynamicznego cechowania na konstrukcję drzew filogenetycznych.

92:5429041977

4.5. Efektywność kodowania sekwencji i rekonstrukcji relacji ewolucyjnych w oparciu o CGR może być optymalizowana

Rozpocznijmy od prostej analizy wpływu mutacji na CGR. Wygenerowano losową sekwencję nukleotydową o długości 1000. Następnie poddano jej kolejno: substytucjom, inercjom, delecjom oraz wszystkim z tych zdarzeń. Przyjęto prawdopodobieństwo wystąpienia zdarzenia mutacji na 0.02 na pozycję. Celem było sprawdzenie w jakim stopniu CGR dla zmienionej sekwencji różni się od oryginalnej. Taka sytuacja występuje np. gdy porównujemy sekwencje homologiczne.



Rysunek 65. Wpływ różnych mutacji na reprezentację CGR. Czarne kropki wskazują na oryginalną sekwencję. w ostatnim przypadku zastosowano wszystkie rodzaje mutacji na raz.

Substytucje powodują przesunięcie punktów względem CGR dla sekwencji niezmienionej. Przy insercji czy delecji zmiany dotyczą głównie tych miejsc, gdzie to zdarzenie nastąpiło. W rzeczywistości, w sekwencjach występują wszystkie z tych zdarzeń. To prowadzi oczywiście do największego zaburzenia początkowego układu punktów.

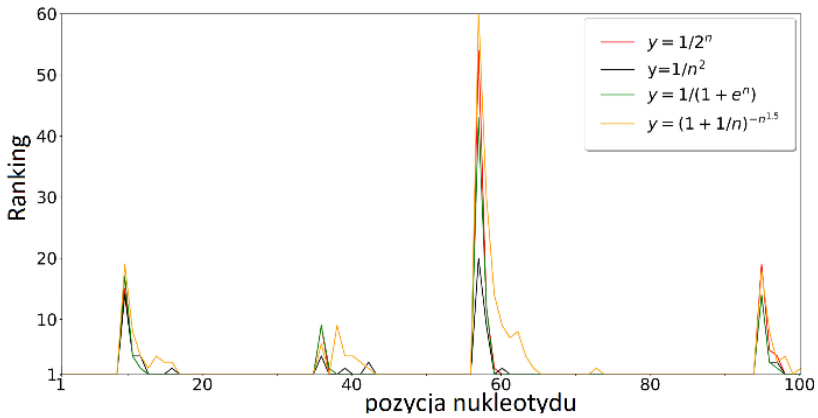
Sprawdźmy teraz w jaki wybór funkcji ważących związany jest z występującymi w sekwencjach mutacjami. Wygenerowano 1000 losowych sekwencji, każda o długości 100 nt. Następnie, każda z tych sekwencji została zmutowana. Rozważano substytucje i delecje. W pierwszym przypadku badano 1 ,2, 3, 5 oraz 10 substytucji na sekwencję. Oznacza to, że stworzono pięć zbiorów sekwencji, które zawierały zmodyfikowane sekwencje z oryginalnego zbioru. Następnie

93:7388330545

sprawdzono w jaki sposób powstałe mutacje zaburzają oryginalne CGR. Idea polega na tym, że podobne sekwencje powinny dzielić podobne reprezentacje. Oznacza to, że odpowiednie punkty powinny być względem siebie relatywnie blisko (CGR oryginalnej i zmutowanej sekwencji). Aby to sprawdzić zbadano rozkład odległości pomiędzy odpowiadającymi sobie punktami i porównano je względem różnych punktów. Załóżmy, że mamy dwie sekwencje oraz niech N_c oznacza liczbę punktów odpowiadających sobie. Niech s_1 (z pierwszej sekwencji) oraz odpowiadający mu s_2 (z drugiej sekwencji) oznaczają dwa wybrane punkty. Wyznaczamy odległość pomiędzy s_1 a wszystkimi punktami z drugiej sekwencji (w tym s_2). Odległości te ustawiamy w kolejności rosnącej i sprawdzamy który na tej liście jest s_2 – nazywamy tę wielkość rankingiem. Mniejszy indeks wskazuje, że zajmowane pozycje są podobne dla oryginalnej i zmutowanej sekwencji (na danej pozycji). Porównano też trzy wprowadzone funkcje wagowe z tradycyjnym CGR używając funkcji oceny

$$S_c = \sum_{i=1}^{N_c} s_{c_i}, \tag{73}$$

gdzie s_{c_i} jest różnicą pomiędzy rankingiem na i-tej pozycji dla klasycznego CGR oraz rozważanej funkcji wagowej. Dodatkowo, autor zastosował analogiczną miarę, redukując jednak wartości s_{c_i} do -1, jeżeli ranking jest wyższy niż w klasycznym CGR, 0, jeżeli jest ten sam i 1, jeżeli jest niższy. Im wyższa wartość S_c, tym bardziej odporna na mutacje jest dana reprezentacja (związana z zadaną funkcją wagową). Warto tutaj podkreślić, że wprowadzone wyżej metryki są odpowiednie do porównywania sekwencji, gdzie odpowiadające sobie pozycje są znane, więc metoda nie jest bezpośrednio użyteczna w kontekście technik bez dopasowania.



Rysunek 66. Wartość rankingu po zastosowaniu różnych funkcji wagowych dla przykładowej sekwencji o długości 100 oraz zmienionej. Wprowadzono 5 substytucji na pozycjach 9, 35, 56, 57 oraz 94. Rysunek opublikowano w Genomics w 2021.

Ranking mierzy stopień podobieństwa pomiędzy odpowiadającymi sobie pozycjami dla dwóch homologicznych sekwencji. Wartość 1 oznacza, że na CGR punkt z oryginalnej sekwencji jest bardzo blisko do odpowiadającego mu punktu w zmienionej sekwencji. Mutacje powodują zaburzenie w rozkładzie punktów, w ich okolicy obserwujemy zwiększone wartości (piki). Zastosowanie funkcji $f(x) = \left(1 + \frac{1}{x}\right)^{-x^{1.5}}$ prowadzi do największych zmian, a tym samym wpływ mutacji jest znaczący. Z kolei $f(x) = \frac{1}{1+e^x}$ wydaje się ograniczać ich oddziaływanie. Jest to oczywiście bezpośrednio związane z wpływem sąsiadujących pozycji na aktualną dla różnych funkcji wag. Podobna analiza została też przeprowadzona dla delecji; rozważano pojedynczą

94:1098026130

delecję o długości 1,2,3,5 oraz 10 pozycji na sekwencję. Wyniki (zaokrąglone do części setnych) zaprezentowano w poniższych tabelach.

Tabela 27. Porównanie klasycznego CGR z trzema funkcjami ważącymi przy pomocy S_c. Średnie zostały policzone tylko dla wartości niezerowych. Dodatkowo, w nawiasach podano średnie dla dodatnich i ujemnych wartości. Tabelę opublikowano w Genomics w 2021.

	$f_1 = \frac{1}{x^2}$	$f_2 = \frac{1}{1+e^x}$	$f_3 = \left(1+\frac{1}{x}\right)^{-x^{1.5}}$
	substytucje		
średnia różnica w rankingu	2,24 (+4,51|-1,24)	2,36 (+2,77,-1,49)	-4,19 (+3,57|-4,82)
procent wartości dodatnich	7,07	7,86	1,98
procent wartości równych zero	88,31	91,31	73,78
procent wartości ujemnych	4,62	0,83	24,24
	delecje		
średnia różnica w rankingu	1,28 (+2,88|-1,29)	1,15 (+2,26|-1,64)	-2,26 (+3,29|-3,48)
procent wartości dodatnich	3,71	3,05	1,78
procent wartości równych zero	93,98	95,74	85,64
procent wartości ujemnych	2,31	1,21	12,58

Tabela 28. Porównanie klasycznego CGR z obustronnym CGR z czterema funkcjami ważącymi przy pomocy S_c. Średnie zostały policzone tylko dla wartości niezerowych. Dodatkowo, w nawiasach podano średnie dla dodatnich i ujemnych wartości. Tabelę opublikowano w Genomics w 2021.

	$f_0 = \frac{1}{2^x}$	$f_1 = \frac{1}{x^2}$	$f_2 = \frac{1}{1+e^x}$	$f_3 = \left(1+\frac{1}{x}\right)^{-x^{1.5}}$
	substytucje			
średnia różnica w rankingu	-3,50 (+4,06|-7,23)	0,13 (+6,36|-4,30)	-1,82(+4,36|-7,66)	-6,02 (+5,98|-7,08)
procent wartości dodatnich	7,06	9,29	9,24	3,50
procent wartości równych zero	78,68	77,67	80,99	56,75
procent wartości ujemnych	14,26	13,04	9,77	39,75
	delecje			
średnia różnica w rankingu	-2,31 (+3,20|-5,08)	0,00 (+4,54|-3,37)	-1,37 (+3,35|-5,41)	-3,73 (+4,70|-4,57)
procent wartości dodatnich	3,97	5,14	4,86	2,37
procent wartości równych zero	88,11	87,94	89,47	73,87
procent wartości ujemnych	7,92	6,92	5,67	23,76

Dwie pierwsze funkcje prowadzą do lepszych rezultatów niż klasyczny CGR (dodatnia wartość średniej różnicy rankingu pomiędzy CGR a zadana funkcją wagi). W tym przypadku, zarówno

95:7533004333

substytucje jak i delecje mają mniejszy wpływ na reprezentacje sekwencji (szczególnie $f(x) = \frac{1}{1+e^x}$). Generalnie, delecje powodują większe różnice niż substytucje. Trzecia z funkcji ($f(x) = \left(1+\frac{1}{x}\right)^{-x^{1.5}}$) jest bardziej podatna na zmiany genetyczne i potencjalnie będzie mniej użyteczna. Testowano też reprezentację obustronną (uwzględniającą wcześniejsze jak i późniejsze pozycje). W każdym z przypadków prowadziło to jednak do wzrostu rankingu, a zatem trudniejszej rozróżnialności punktów. Dodatkowo, sprawdzony został wpływ parametru K (dla obustronnego CGR). Testowano $K = 1,2,\dots,10$. Wyższe jego wartości prowadziły do korzystniejszych rezultatów. Jednakże, żadnej z przypadków nie był lepszy od jednostronnego CGR. Wynika to głównie z niejednoznaczności takiej reprezentacji i większej koncentracji punktów przy centrum. Punkty mogą zajmować podobne miejsca, mimo że pochodzą z różnych miejsc. W konsekwencji, rozróżnialność spada a ranking rośnie. Dalsze analizy zostaną ograniczone tylko do jednostronnego CGR.

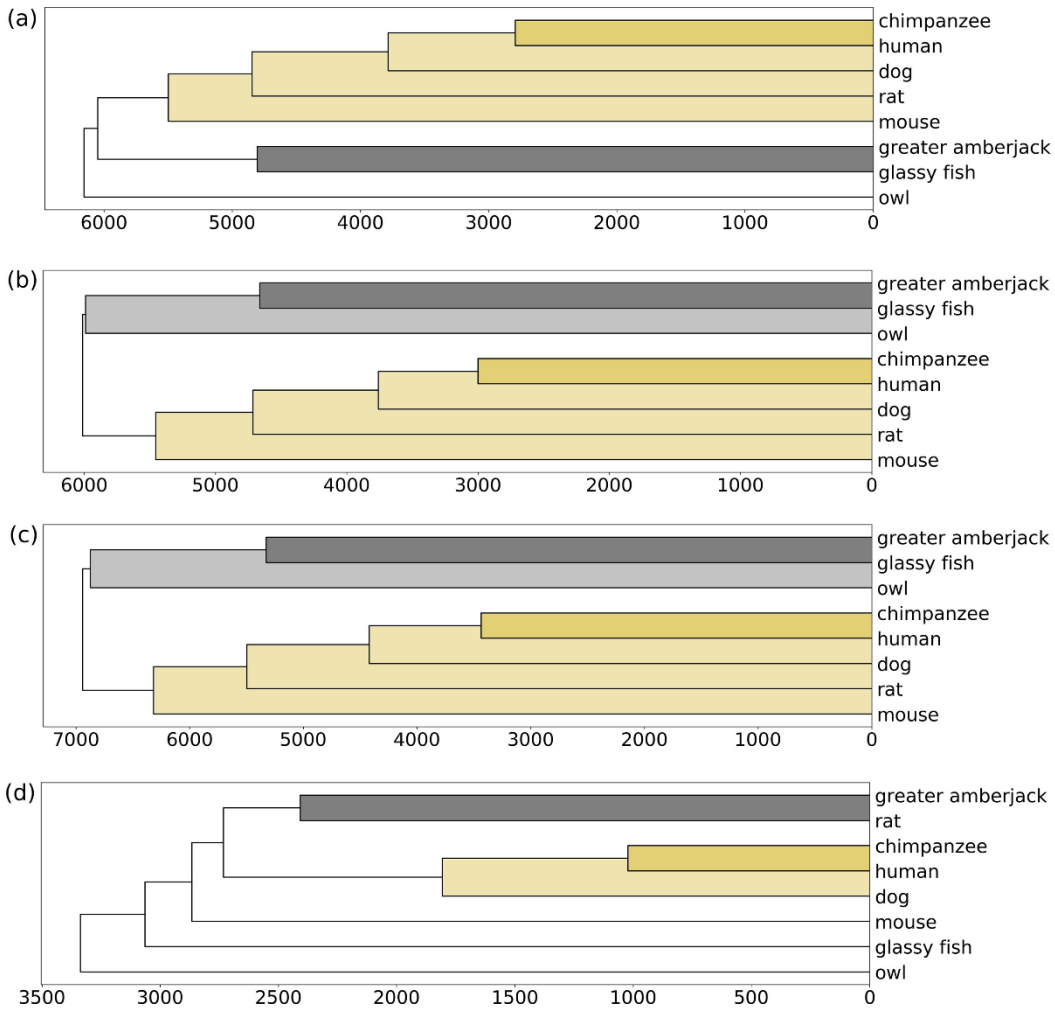
Sprawdzimy teraz jak wybór funkcji wagowej wpływa na rekonstrukcję drzew filogenetycznych. Rozważmy mRNA dla genu SPARC z 8 organizmów. Gen SPARC koduje białko bogate w cysteinę odgrywające ważną rolę m.in. w interakcji międzykomórkowej i jest szeroko rozpowszechnione u organizmów wyższych [131]. Sekwencje pobrano z NCBI.

Tabela 29. Rozważane organizmy wraz z NCBI ID sekwencji SPARC. Tabelę opublikowano w Genomics w 2021.

Organizm	Identyfikator NCBI
Burrowing owl (owl)	XM_026858278.1
Mus musculus (mouse)	NM_009242.5
Rattus norvegicus (rat)	FQ227727.1
Homo sapiens (human)	NM_001309443.2
Pan troglodytes (chimpanzee)	XM_024356783.1
Canis lupus (dog)	XM_014113053.2
Parambassis ranga (glassy fish)	XM_028416359.1
Seriola dumerili (greater amberjack)	XM_022760346.1

Te zostały przekształcone wg CGR z odpowiednią funkcją wagową. Następnie poddane dyskretnej transformacie Fouriera (DFT). Drzewo zostało zbudowane za pomocą metody UPGMA.

96:7036818628

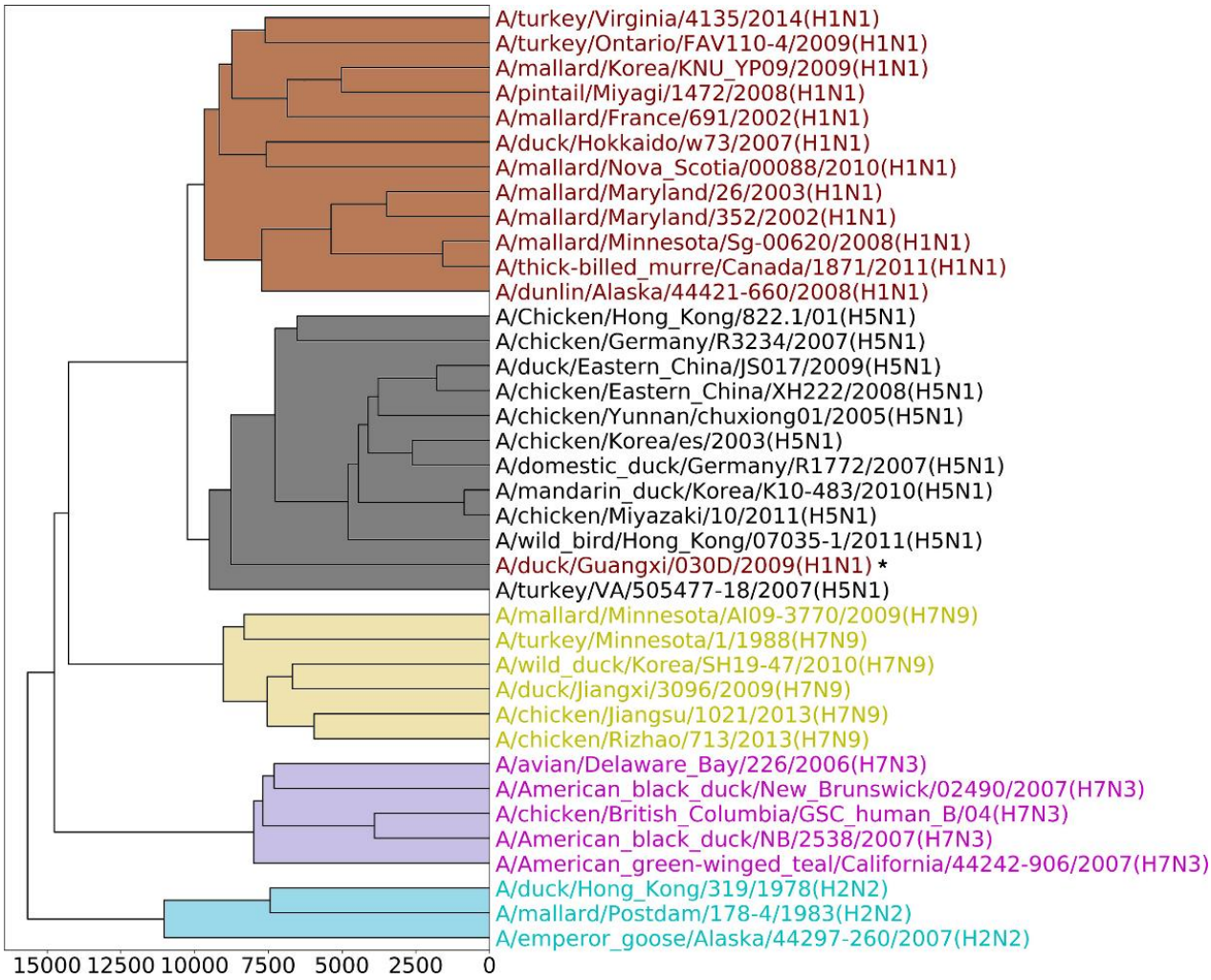


Rysunek 67. Drzewa filogenetyczne w oparciu o CGR dla 8 sekwencji mRNA genu SPARC. Rozważano 4 funkcje wagowe (a-d), klasyczną $1/2^x$, $1/x^2$, $1/(1+e^x)$ oraz $(1+1/x)^{-x\wedge 1.5}$ odpowiednio. Rysunek opublikowano w Genomics w 2021.

Tradycyjny CGR dostarcza całkiem rozsądne relacje pomiędzy organizmami. Zastosowanie funkcji $1/x^2$ oraz $1/(1+e^x)$ prowadzi do umieszczenia sowy do tego samego kladu co ryby. Ostatnia funkcja mocno zaburza relacje pomiędzy organizmami i wydaje się prowadzić do niepoprawnych konfiguracji.

Rozważmy jeszcze analizowane już wcześniej sekwencje neuraminidazy wirusów grypy. Zbiór tych sekwencji jest często używany do testowania nowych metod [35, 130]. Wirusy grypy są klasyfikowane w oparciu o gen neuraminidazdy (N) oraz hemaglutyniny (H) [129]. Sekwencje pochodzą z ptasich wirusów azjatyckich. Pierwsze drzewo powstało w oparciu o klasyczny CGR. Obserwujemy tutaj bardzo dobre rozróżnienie pomiędzy różnymi typami. Występuje jedynie jedna sekwencja, która wydaje się powinna znajdować się w innej grupie.

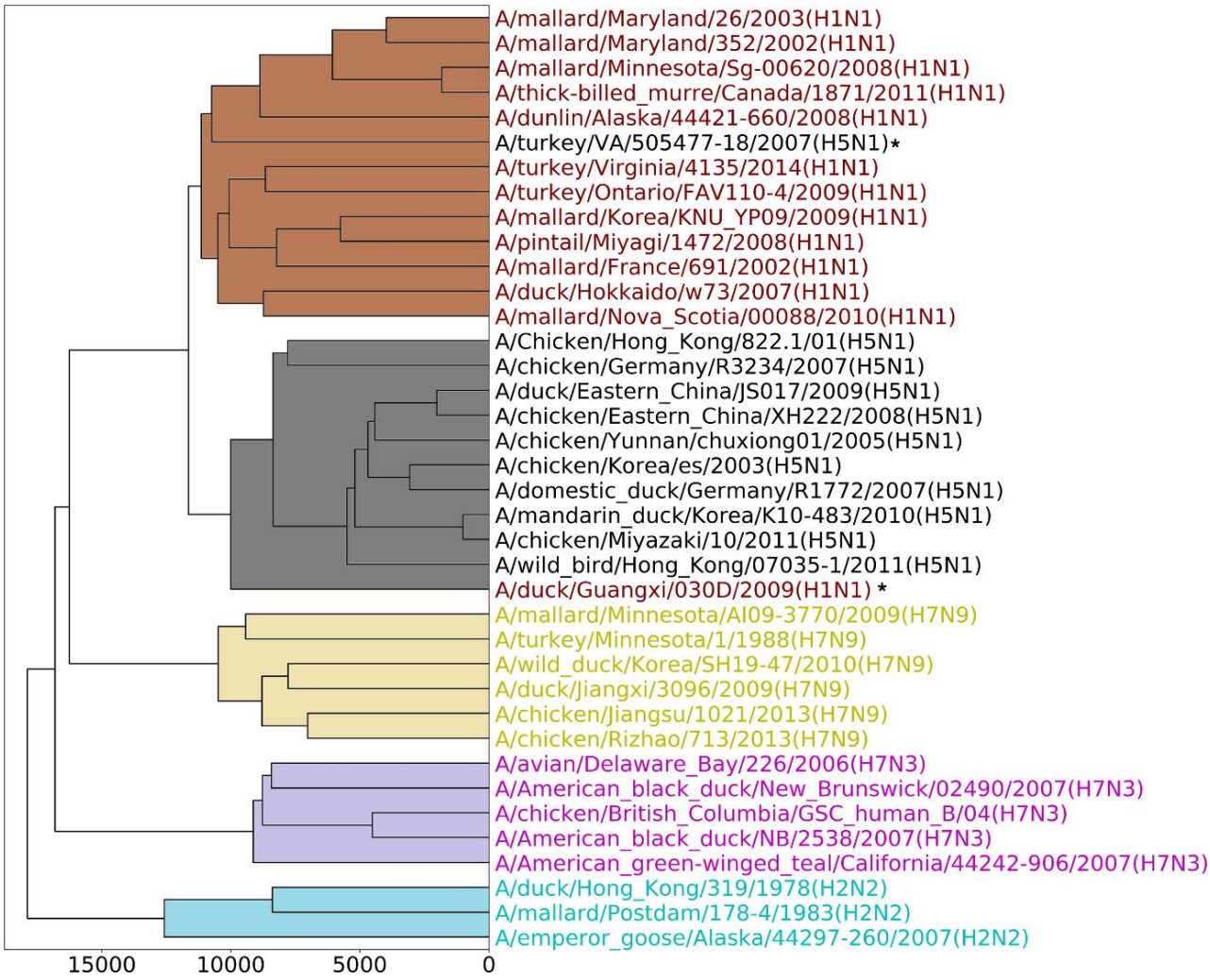
97:1078011195



Rysunek 68. Drzewo filogenetyczne skonstruowane w oparciu o CGR oraz DFT za pomocą metody UPGMA. Wybrano klasyczną funkcję wagową $f(x)=1/2^x$. Rysunek opublikowano w Genomics w 2021.

Kolejne drzewo powstało w oparciu o CGR z funkcjami $f(x) = 1/x^2$ oraz $f(x) = 1/(1+e^x)$ (w obu przypadkach otrzymano te same topologie drzew). Widzimy, że w stosunku do poprzedniego drzewa wirus *turkey* H5N1 został przeniesiony do grupy wirusów H1N1, natomiast wirus *duck* H1N1 pozostał w H5N1. W dalszej części zostanie to jeszcze poddane pewnej dodatkowej analizie.

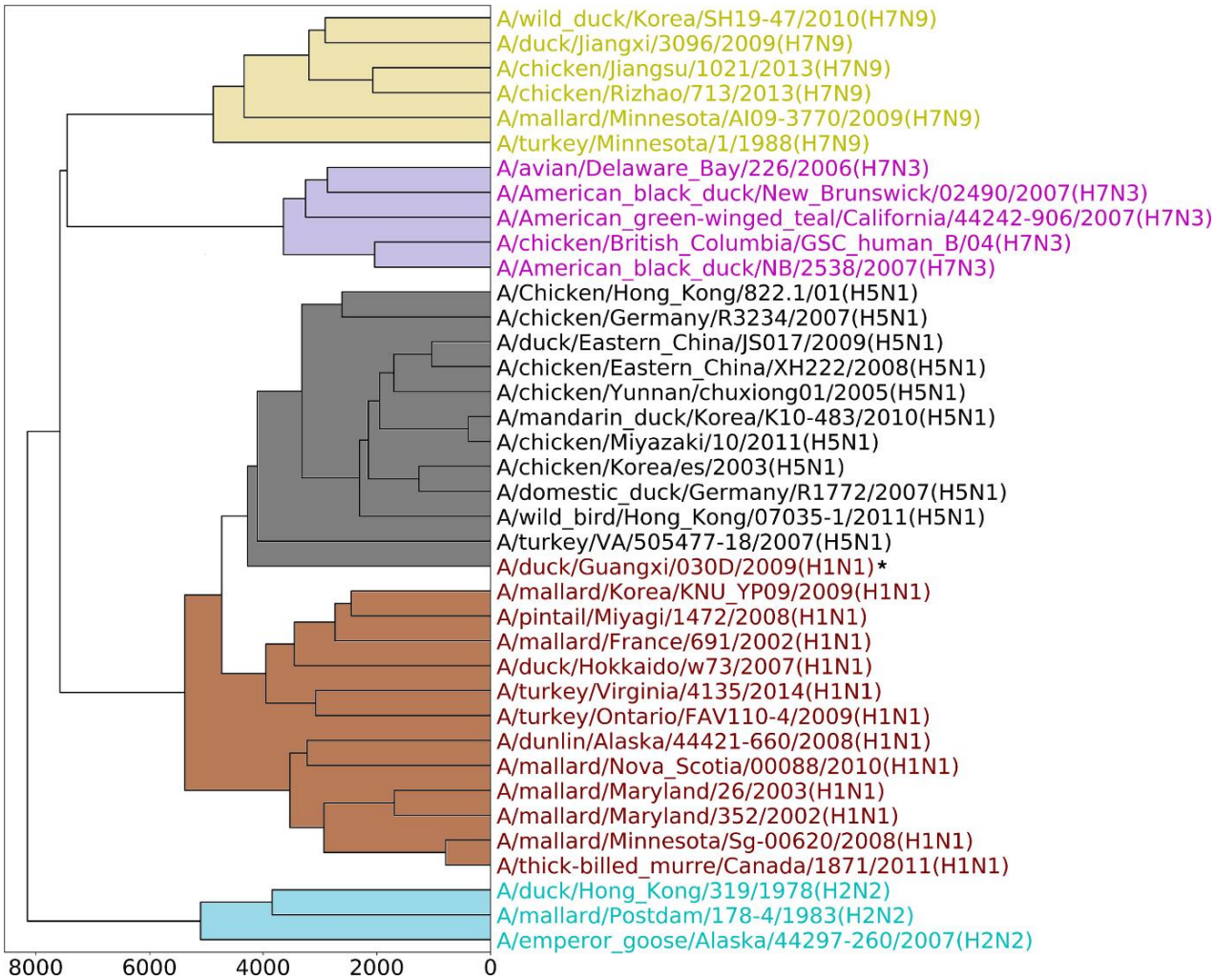
98:6806954340



Rysunek 69. Drzewo filogenetyczne skonstruowane w oparciu o CGR oraz DFT za pomocą metody UPGMA. Jako funkcje wagowo wybrano $f(x)=1/x^2$($f(x) = 1/(1+e^x)$). Rysunek opublikowano w Genomics w 2021.

Ostatnie drzewo powstało w oparciu o funkcję $f(x) = (1+1/x)^{-x^{\wedge}1.5}$. Może się wydawać, że ma ono najlepszą topologię ze wszystkich rozważanych drzew. Tutaj m.in. wirus *duck* H1N1 nadal znajduje się w grupie H5N1, jednak stanowi on wyraźnie najbardziej odległy takson w tej grupie. Zgodnie z tym drzewem, wirus *turkey* H5N1 jest bardziej zbliżony do H5N1 (wg oczekiwań).

99:5981146004



Rysunek 70. Drzewo filogenetyczne skonstruowane w oparciu o CGR oraz DFT za pomocą metody UPGMA. Jako funkcję wagową wybrano $f(x) = (1+1/x)^{-x^{1.5}}$. *Rysunek opublikowano w Genomics w 2021.*

Aby zweryfikować dwa problematyczne wirusy – *turkey* H5N1 oraz *duck* H1N1 użyto dodatkowo narzędzia BLAST [156,157]. W analizie wykorzystano sekwencje białkowe neuraminidazy dla obu wirusów i porównano je z innymi wirusami należącymi do grupy H1N1 oraz H5N1.

100:3095483978

Tabela 30. Procent identyczności sekwencji białka neuraminidazy pomiędzy A/turkey/VA/505477–18/2007(H5N1) a innymi wybranymi wirusami H5N1/H1N1. Tabelę opublikowano w Genomics w 2021.

Wirus	Procent identyczności
A/chicken/Germany/R3234/2007(H5N1)	90.42 %
A/duck/Eastern China/JS017/2009(H5N1)	89.98 %
A/chicken/Eastern China/XH222/2008(H5N1)	90.42 %
A/chicken/Yunnan/chuxiong01/2005(H5N1)	89.53 %
A/chicken/Korea/es/2003(H5N1)	90.87 %
A/domestic duck/Germany/R1772/2007(H5N1)	90.20 %
A/mandarin duck/Korea/K10-483/2010(H5N1)	88.86 %
A/chicken/Miyazaki/10/2011(H5N1)	88.64 %
A/wild bird/Hong Kong/07035-1/2011(H5N1)	90.42 %
A/Chicken/Hong Kong/822.1/01 (H5N1)	91.09 %
A/dunlin/Alaska/44421-660/2008(H1N1)	92.11 %
A/thick-billed-murre/Canada/1871/2011(H1N1)	92.54 %
A/mallard/Minnesota/Sg-00620/2008(H1N1)	92.75 %
A/mallard/Maryland/352/2002(H1N1)	92.54 %
A/mallard/Maryland/26/2003(H1N1	92.32 %
A/duck/Guangxi/030D/2009(H1N1)*	92.87 %

101:7823536781

Tabela 31. Procent identyczności sekwencji białka neuraminidazy pomiędzy A/duck/Guangxi/030D/2009(H1N1) a innymi wybranymi wirusami H5N1/H1N1. Tabelę opublikowano w Genomics w 2021.

Wirus	Procent identyczności
A/chicken/Germany/R3234/2007(H5N1)	95.32 %
A/duck/Eastern China/JS017/2009(H5N1)	94.88 %
A/chicken/Eastern China/XH222/2008(H5N1)	95.32 %
A/chicken/Yunnan/chuxiong01/2005(H5N1)	95.32 %
A/chicken/Korea/es/2003(H5N1)	95.99 %
A/domestic duck/Germany/R1772/2007(H5N1)	95.10 %
A/mandarin duck/Korea/K10-483/2010(H5N1)	94.88 %
A/chicken/Miyazaki/10/2011(H5N1)	94.65 %
A/wild bird/Hong Kong/07035-1/2011(H5N1)	95.10 %
A/Chicken/Hong Kong/822.1/01 (H5N1)	95.77 %
A/turkey/VA/505477-18/2007(H5N1)*	92.87 %
A/dunlin/Alaska/44421-660/2008(H1N1)	89.13 %
A/thick-billed-murre/Canada/1871/2011(H1N1)	89.77 %
A/mallard/Minnesota/Sg-00620/2008(H1N1)	89.77 %
A/mallard/Maryland/352/2002(H1N1)	89.98 %
A/mallard/Maryland/26/2003(H1N1	89.77 %
A/duck/Hokkaido/w73/2007(H1N1)	92.54 %
A/mallard/Nova Scotia/00088/2010(H1N1)	89.13 %
A/turkey/Virginia/4135/2014(H1N1)	84.65 %
A/turkey/Ontario/FAV110-4/2009(H1N1)	85.07 %
A/mallard/France/691/2002(H1N1)	91.90 %
A/mallard/Korea/KNU YP09/2009(H1N1)	92.54 %
A/pintail/Miyagi/1472/2008(H1N1)	92.96 %

Okazało się, że biorąc pod uwagę procent identyczności, wirus *turkey* H5N1 jest bardziej podobny do wirusów H1N1. Z kolei wirus *duck* H1N1 wykazuje większe podobieństwo do grupy H5N1. Uruchamiając BLAST dla wirusa *duck* H1N1, pierwsze 98 wyników należy do wirusów z grupy H5N1. Co więcej, jego sekwencja posiada charakterystyczną dla H5N1 delecję 60 nt (czyli 20 aminokwasów), która powoduje większą wirulencję tych wirusów [158]. Co ciekawe, wirus *turkey* H5N1 nie posiada tej delecji. Powyższe analizy skłaniają nas do przynależności sekwencji neuraminidazy wirusa *duck* H1N1 do H5N1 zamiast H1N1. Powodem, dla którego wirus *turkey* H5N1 został sklasyfikowany do H5N1 (na pierwszym i trzecim drzewie) jest jego wysoki stopień podobieństwa do wirusa *duck* H1N1. Jednakże, uwzględniając też inne wirusy ma on większe powinowactwo do H1N1. Na podstawie tego zestawu sekwencji trudno jednak w sposób decydujący rozstrzygnąć jego prawdziwe pochodzenie. Przydatna byłaby w tym celu druga sekwencja markerowa (hemaglutyniny). Zwróćmy bowiem uwagę, że wirusy te nadal należą do pierwszego typu neuraminidazy, a to jest zgodne z ich klasyfikacją.

Przejdźmy teraz do sekwencji aminokwasowych. Sprawdzimy na ile początkowy wybór parametryzacji wpływa na końcowe drzewo filogenetyczne. Zastosujemy 3 zestawy parametrów: losowy, hydrofobowość/hydrofilowość oraz wytrenowane (emb-CGR). Rozważmy na początku ND5 (*NADH-ubiquinone oxidoreductase chain 5),* która stanowi podjednostkę dehydrogenazy NADH i jest zlokalizowana w mitochondrium [159]. Często wykorzystywana jest do oceny nowych

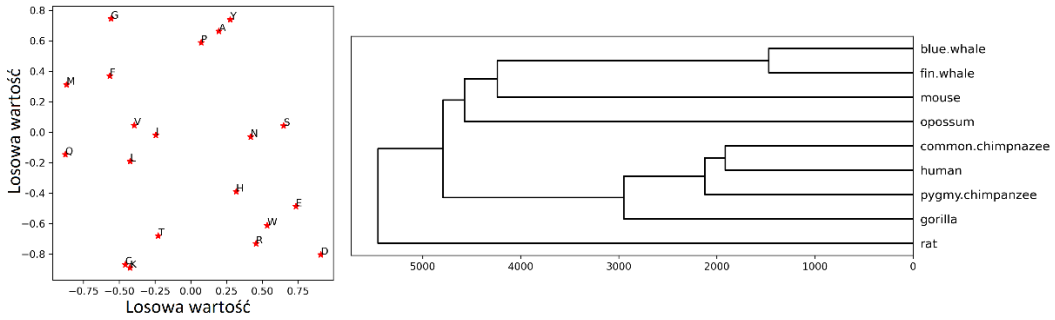
102:1366561984

metod do rekonstrukcji drzew ewolucyjnych (m.in. w [98]). Poniżej znajdują się ID oraz organizmy, których sekwencje zostały ponownie pobrane z NCBI (pozostawiono angielskie nazwy).

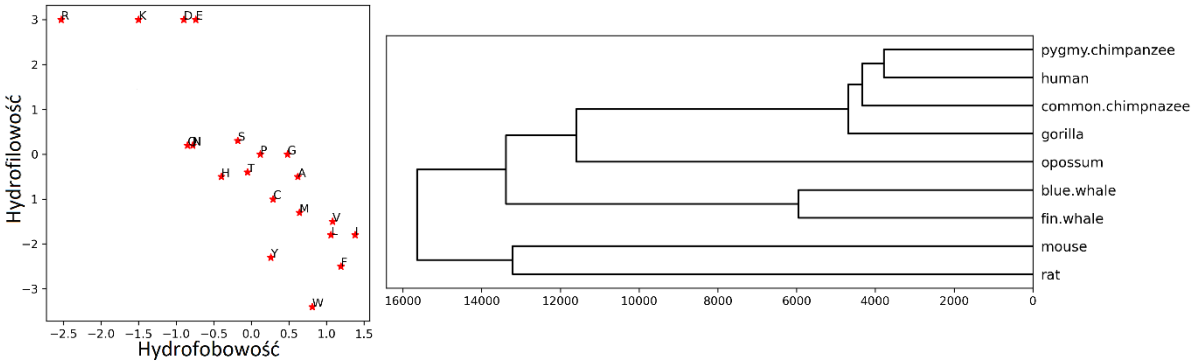
Tabela 32. Rozważane organizmy. Podano również ID z NCBI. Tabelę opublikowano w Comput Biol Med. w 2022.

Nr	Organizm	NCBI ID
1	Human (Homo sapiens)	QRC37232.1
2	Gorilla (Gorilla gorilla)	NP_008222
3	Pygmy chimpanzee (Pan paniscus)	NP_008209
4	Common chimpanzee (Pan troglodytes)	NP_008196
5	Fin whale (Balenoptera physalus)	NP_006899
6	Blue whale (Balenoptera musculus)	NP_007066
7	Rat (Rattus norvegicus)	AP_004902
8	Mouse (Mus musculus)	NP_904338
9	Opossum (Didelphis virginiana)	NP_007105

Poniżej przedstawiono kolejne wygenerowane drzewa razem z reprezentacją dla aminokwasów.

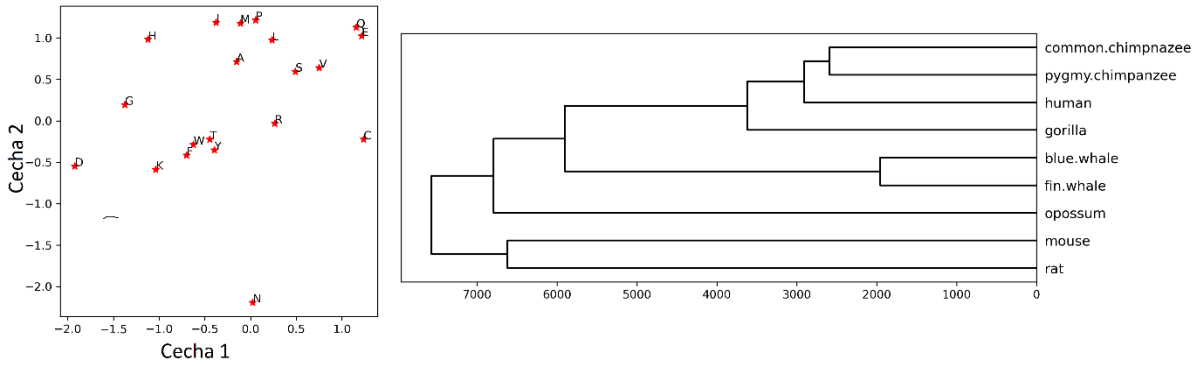


Rysunek 71. Drzewo filogenetyczne dla sekwencji ND5 w oparciu o CGR i DFT. Aminokwasy sparametryzowano w sposób losowy. Rysunek opublikowano w Comput Biol Med. w 2022.



Rysunek 72. Drzewo filogenetyczne dla sekwencji ND5 w oparciu o CGR i DFT. Aminokwasy sparametryzowano za pomocą ich hydrofobowości i hydrofilności. Rysunek opublikowano w Comput Biol Med. w 2022.

103:1031377798



Rysunek 73. Drzewo filogenetyczne dla sekwencji ND5 w oparciu o CGR i DFT. Aminokwasy sparametryzowano za pomocą techniki wektorów osadzenia. Rysunek opublikowano w Comput Biol Med. w 2022.

Widzimy, że w pierwszym przypadku występuje kilka rozbieżności. Dla przykładu, mysz znajduje się blisko wielorybów a szczur wydaje się najbardziej odległym organizmem. Po uwzględnieniu własności fizykochemicznych aminokwasów (hydrofobowość i hydrofilowość) otrzymane drzewo jest bardziej wiarygodne. Tutaj mysz jest zgrupowana razem ze szczurem. Po uwzględnieniu dostosowanej reprezentacji obserwujemy kolejne, mniejsze, ale istotne różnice. Wtedy, szympansy są określone jako bliskie organizmy i ulokowane w jednej grupie. Potwierdza to korzyści ze stosowania dynamicznej parametryzacji aminokwasów.

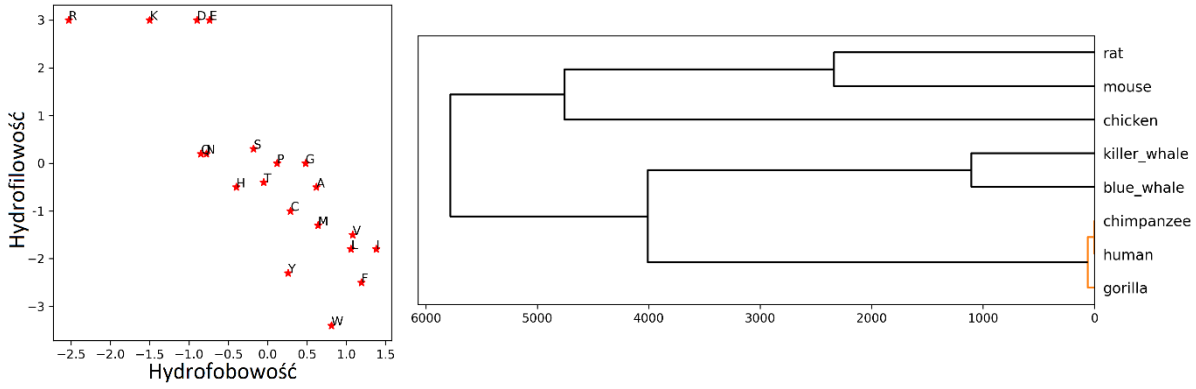
Drugim przypadkiem będzie analizowane już białko SPARC. W tabeli znajdują się odpowiednie nazwy organizmów i ID do rekordów z NCBI.

Tabela 33. Rozważane organizmy. Podano również ID z NCBI. Tabelę opublikowano w Comput Biol Med. w 2022.

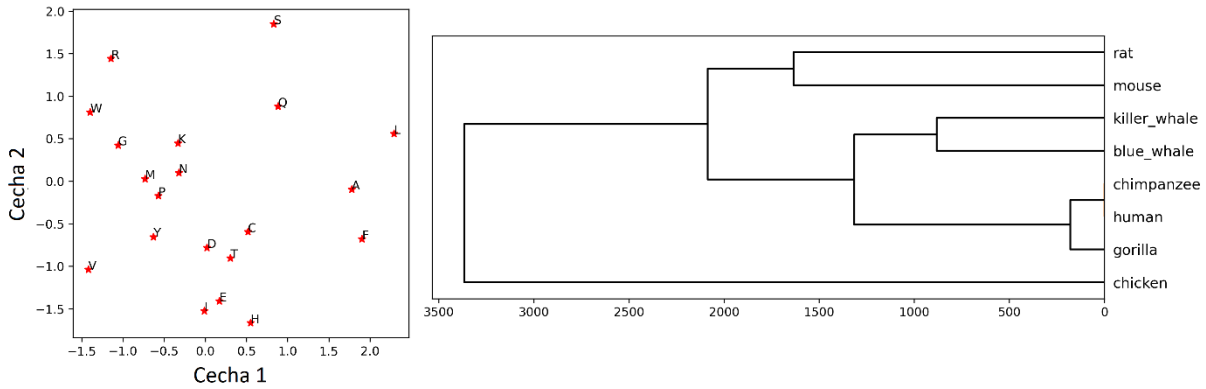
Nr	Organizm	NCBI ID
1	Human (Homo sapiens)	CAG33080.1
2	Gorilla (Gorilla gorilla)	XP_004042902.1
3	Common chimpanzee (Pan troglodytes)	XP_024212551.1
4	Killer whale (Ornicus orca)	XP_033262575.1
5	Blue whale (Balenoptera musculus)	XP_036704335.1
6	Rat (Rattus rattus)	XP_032770322.1
7	Mouse (Mus musculus)	CAJ18514.1
8	Chicken (Gallus gallus)	XP_040509250.1

Zamieszczono tutaj m.in. kurczaka, który powinien stanowić grupę zewnętrzną w stosunku do ssaków. Wygenerowano drzewo w oparciu o własności fizykochemiczne oraz dostosowane cechy.

104:1061126545



Rysunek 74. Drzewo filogenetyczne dla sekwencji SPARC w oparciu o CGR i DFT. Aminokwasy sparametryzowano za pomocą ich hydrofobowości i hydrofilności. Rysunek opublikowano w Comput Biol Med. w 2022.



Rysunek 75. Drzewo filogenetyczne dla sekwencji SPARC w oparciu o CGR i DFT. Aminokwasy sparametryzowano za pomocą techniki wektorów osadzenia. Rysunek opublikowano w Comput Biol Med. w 2022.

W pierwszym przypadku kurczak znajduje się w tej samej grupie co mysz czy szczur, co nie jest oczekiwane. Poprzez zastosowanie dynamicznej reprezentacji dla aminokwasów, stanowi on grupę zewnętrzną w stosunku do ssaków.

Widzimy więc, że parametryzacja aminokwasów stanowi istotny element do końcowego CGR, dlatego warto uwzględnić przynajmniej cechy fizykochemiczne aminokwasów.

105:1101755069

Rozdział 5. Podsumowanie

W pracy przedstawiono różne aspekty analizy sekwencji przy zastosowaniu reprezentacji gry chaosu. Autor postawił kilka hipotez i celów z tym związanych. Na początku zastosowano tę reprezentację do analizy puli ludzkich mikroRNA; zarówno całego zbioru jak i podzielonego na dwie grupy, w zależności od energii oddziaływania z transkryptem. W tej pracy zamieszczono fragment szerszych badań, których częścią była właśnie analiza motywów występujących w tych cząsteczkach. Wyniki zostały zaprezentowane graficznie i okazało się m.in., że w grupie niskoenergetycznie związanych mikroRNA obserwujemy deficyt dinukleotydów UA czy AU, z kolei dla drugiej grypy było to przede wszystkim CG. Potwierdzono to klasycznymi wykresami częstości i skonfrontowano z literaturą. Należy tutaj podkreślić, że CGR zawiera pełną informację o rozważanych sekwencjach (nie tylko częstościach określonych k-merów). W przyszłości warto być może zastosować to narzędzie do innych organizmów i porównać ich reprezentacje między sobą, aby sprawdzić czy podobne relacje występują. W kolejnym rozdziale poddano analizie pulę genów niezbędnych u bakterii pobranych z bazy danych DEG. Rozpoczęto od standardowej analizy statystycznej. Na prosty, ale szczególnie interesujący rezultat należy wskazać wykresy A vs. T oraz G vs. C, gdzie występuje odmienny charakter dla puli genów niezbędnych oraz uzupełniających. Zastosowanie reprezentacji gry chaosu do analizy miało na celu uwzględnienie dalekozasięgowych relacji w obrębie sekwencji. Wyznaczono spacer DNA oparty o CGR i policzono dla niego różne miary. Po pierwsze, okazało się, że różne bakterie mają inne, charakterystyczne dla nich reprezentacje (oczywiście nie wliczając tych najbardziej spokrewnionych). Po drugie, że sekwencje z puli genów niezbędnych są dosyć charakterystyczne dla kolejnych bakterii. Oczywiście istnieje pula genów, które są bardzo podobne, a często i w znacznej części niezmienione między rożnymi bakteriami. Wykreślając zależność pomiędzy pulą genów niezbędnych a opcjonalnych dla miary R (odległość od początku układu współrzędnych od ostatniego punktu reprezentacji) uzyskano umiarkowanie silną zależność. Nie jest to prosta konsekwencja długości sekwencji czy składu pojedynczych nukleotydów, co potwierdzono stosując jako referencję wygenerowane sekwencje losowe o zadanych parametrach. W kolejnym rozdziale rozważano zastosowanie CGR do rekonstrukcji relacji ewolucyjnych w połączeniu z innymi metodami – DFT, wykładnik Hursta oraz SSSIM. W pierwszych dwóch reprezentację sekwencji traktujemy jako ciąg liczb, w trzeciej skupiamy się na jej postaci graficznej. Metody te były wcześniej zaproponowane w literaturze, jednak często publikacje odnoszą wyniki do dwóch/trzech zestawów sekwencji. W tej pracy, analizie poddano kilka zestawów referencyjnych takich jak gen neuraminidazy czy genomy bakteryjne. Autor wprowadził też pewne dodatkowe modyfikacje, np. w podejściu z wykładnikiem Hursta wyznaczono ostatecznie 4 współczynniki charakteryzujące szereg. W uproszczony sposób wygenerowano też kilka ścieżek ewolucyjnych i za pomocą odległości JRF sprawdzono topologię otrzymanych drzew porównując je z referencją. Zdecydowanie najkorzystniej wypadła tutaj metoda oparta o SSIM. W dalszej części ta związana z DFT oraz wykładnikiem Hursta. Mimo, że wszystkie podejścia prowadziły do nielosowych relacji ewolucyjnych, to potencjalne zastosowanie praktyczne będą miały zapewne właśnie CGR/SSIM oraz CGR/DFT. Porównano czasy obliczeń i te były porównywalne dla obydwu metod; w przypadku wykładnika Hursta obliczenia trwały znacznie dłużej, a sam algorytm zachowywał się wręcz niestabilnie dla bardzo długich sekwencji. Należy jednak dodać, że w przypadku wybranych zbiorów sekwencji (jak gen neuraminidazy u wirusów grypy) wszystkie metody doprowadziły do wiarygodnych relacji ewolucyjnych. Z użyciem platformy AFproject przeanalizowano kolejnych 5 zestawów

106:2883819729

referencyjnych – 4 rzeczywiste oraz 1 symulowany. W części przypadków rozważany był także horyzontalny transfer genów. Wyniki były konsystentne z tymi z poprzedniego rozdziału. Najlepsze rezultaty otrzymano dla genomów mitochondrialnych różnych gatunków ryb. W przypadku SSIM unormowany współczynnik RF wynosił 0.09. Wprowadzone metody poddano tutaj krytycznej analizie zwracając uwagę na ich słabe strony. Dla DFT jest to przede wszystkim konieczność wydłużania widm, co znacząco obniża efektywność, szczególnie dla dłuższych i bardziej zróżnicowanych pod tym względem sekwencji. W przypadku SSIM, prócz złożoności czasowej, należy uwzględnić tutaj złożoność pamięciową związaną z przechowywaniem obrazów. Dla dłuższych sekwencji utrudnione może być rozpoznawanie charakterystycznych wzorców. Wreszcie, w przypadku wykładnika Hursta następuje redukcja informacji do 4 liczb, które następnie porównywane są między sekwencjami. Dla bardzo długich sekwencji jest to dalece niewystarczające, jednak według autora może stanowić jeden z parametrów do szybkiej oceny podobieństwa zwłaszcza w bardziej zróżnicowanej grupie sekwencji. W kolejnym rozdziale zaproponowano uogólnienie metody CGR, rozważając inne funkcje ważące oraz parametryzacje reszt. Sprawdzono własności matematyczne oraz konsekwencje biologiczne tych zmian, m.in. wpływ na widma DFT jak i późniejszą konstrukcję drzew filogenetycznych. W przypadku sekwencji białkowych zaproponowano dynamiczną reprezentację aminokwasów w oparciu o techniki przetwarzania języka naturalnego. Ta prowadziła do korzystnych modyfikacji generowanych relacji ewolucyjnych. Zaproponowano także CGR dla sekwencji RNA uwzględniającą informację o strukturze drugorzędowej. Przyszłe jej zastosowanie może obejmować m.in. odpowiednie osadzenie tego typu sekwencji przy ich dalszym przetwarzaniu, np. za pomocą technik uczenia maszynowego. Badania związane z CGR jak i zagadnieniami powiązanymi z tej pracy były też rozważane/rozwijane w ramach pracowni magisterskich. Dla przykładu, porównywano rekonstrukcję drzew filogenetycznych metodami opartymi o CGR z klasycznymi (jak tymi opartymi o dopasowanie czy metodą k-merów). Uwzględniając część deskryptorów, które rozważano w przypadku puli genów niezbędnych zbudowano model klasyfikacyjny. Podsumowując, reprezentacja gry chaosu dostarcza użyteczny sposób kodowania przy analizie sekwencji biologicznych. Część technik powinna jednak zostać jeszcze dalej udoskonalona, aby realnie konkurować z obecnymi rozwiązanymi chociażby przy analizie filogenetycznej.

107:1144861406

Dodatki

Implementacja wybranych metod do porównywania sekwencji opartych o CGR znajduje się tutaj

https://github.com/adkwazar/CGR

Literatura

1. R. M. Durbin, S. Eddy, **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids**, Cambridge, 1998.
2. J. Watson, **The Secret of Life**, Cornerstone, 2017.
3. D. Deshpande i inni, **RNA-seq data science: From raw data to effective interpretation**, Front Genet., 2023, doi: 10.3389/fgene.2023.997383.
4. V. A. Jisna, P. B. Jayaraj, **Protein Structure Prediction: Conventional and Deep Learning Perspectives**, Protein J., 522-544, 2021, doi: 10.1007/s10930-021-10003-y.
5. J. W. Prokop i inni, **Genome sequencing in the clinic: the past, present, and future of genomic medicine**, Physiol Genomics., 50(8), 563-579, 2018, doi: 10.1152/physiolgenomics.00046.2018.
6. S.H. Ye i inni, **Benchmarking Metagenomics Tools for Taxonomic Classification**, Cell, 178(4), 779-794, 2019, doi: 10.1016/j.cell.2019.07.010.
7. R. Guigo, **Genome annotation: From human genetics to biodiversity genomics**, Cell Genom., 3(8), 2023, doi: 10.1016/j.xgen.2023.100375.
8. A. David i inni, **DNA Sequencing Technologies and DNA Barcoding**, Methods Mol Biol., 2744:139-154, 2024, doi: 10.1007/978-1-0716-3581-0_8.
9. S. Morganti i inni, **Next Generation Sequencing (NGS): a Revolutionary Technology in Pharmacogenomics and Personalized Medicine in Cancer**, Adv Exp Med. Biol., 9-30, 2019, doi: 10.1007/978-3-030-24100-1_2.
10. D. Xia i inni, **Applications of single-cell RNA sequencing in atopic dermatitis and psoriasis**, Front Immunol., 2022, doi: 10.3389/fimmu.2022.1038744.
11. M. Mirza i inni, **Assessing the Cost-Effectiveness of Next-Generation Sequencing as a Biomarker Testing Approach in Oncology and Policy Implications: A Literature Review, Value Health**, 27(9):1300-1309, 2024 doi: 10.1016/j.jval.2024.04.023.
12. J. Xiong, **Essential Bioinformatics**, Cambridge University Press, 2006.
13. S. Vinga, J. Almeida, **Alignment-free sequence comparison—a review, Bioinformatics**, 19(4), 513-23, 2003, doi: 10.1093/bioinformatics/btg005.
14. O. Bonham-Carter i inni, **Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis**, Brief Bioinform., 15(6), 890-905, 2014, doi: 10.1093/bib/bbt052.
15. A. Zielezinski i inni, **Alignment-free sequence comparison: benefits, applications, and tools**, Genome Biol., 18(1), 2017, doi: 10.1186/s13059-017-1319-7.
16. H.J. Jeffrey, **Chaos game representation of gene structure**, Nucleic Acids Res., 18(8), 2163-70. 1990, doi: 10.1093/nar/18.8.2163.
17. Z.G. Yu, V. Anh, K.S. Lau, **Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analysis**, J Theor Biol, 226(3), 341-8, 2004, doi: 10.1016/j.jtbi.2003.09.009.

108:1116088021

18. H.F. Lochel i inni, **Deep learning on Chaos Game representation for proteins,** Bioinformatics, 36(1), 272-279, 2020, doi: 10.1093/bioinformatics/btz493.
19. P.J. Deschavanne i inni, **Genomic signature: characterization and classification of species assesed by chaos game reprezentation of sequences**, Mol. Biol. Evol., 16(10), 1391–1399, 1999, doi: 10.1093/oxfordjournals.molbev.a026048.
20. Y. Wang i inni, **Spectrum of genomic signatures: from dinucleotides to chaos game representation,** Gene, 346, 173–185, 2005, doi: 10.1016/j.gene.2004.10.021.
21. Y.Z. Zhang, S. Imoto, **Genome analysis through image processing with deep learning models**, J Hum Genet., 69(10):519-525, 2024, doi: 10.1038/s10038-024-01275-0.
22. M. Itaya, **An estimation of minimal genome size required for life**, FEBS Lett., 362(3):257-60, 1995 doi: 10.1016/0014-5793(95)00233-y.
23. G. Giaever i inni, **Functional profiling of the Saccharomyces cerevisiae genome**, Nature., 418(6896):387-91, 2002 doi: 10.1038/nature00935.
24. Y. Veeranagouda Y i inni, **Identification of genes required for the survival of B. fragilis using massive parallel sequencing of a saturated transposon mutant library**, BMC Genomics., 15(1):429, 2014, doi: 10.1186/1471-2164-15-429.
25. E.M. i inni, **Molecular mechanisms of antibiotic resistance revisited**, Nat Rev Microbiol., 21(5):280-295, 2023, doi: 10.1038/s41579-022-00820-y.
26. Q. Zhou, Y. Yu**, Comparative analysis of bacterial essential and nonessential genes with Hurst exponent based on chaos game representation**, Chaos, Solitons and Fractals, 69, 209-216, 2014, doi: 10.1016/j.chaos.2014.10.003.
27. X. Liu, B. Wang, L. Xu, **Statistical Analysis of Hurst Exponents of Essential/Nonessential Genes in 33 Bacterial Genomes**, PLoS One., 12;10(6), 2015, doi: 10.1371/journal.pone.0129716
28. T. Hoang, C. Yin, S.S. Yau, **Splice sites detection using chaos game representation and neural network**, Genomics, 112(2), 1847-1852, 2020, doi: 10.1016/j.ygeno.2019.10.018.(28)
29. A. Arsiccio, L. Stratta, T. Menzen, **Evaluating the chaos game representation of proteins for applications in machine learning models: prediction of antibody affinity and specificity as a case study,** J Mol Model., 29(12):377, 2023, doi: 10.1007/s00894-023-05777-0.
30. Y. Ren i inni, **Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning, Bioinformatics**, 38(2):325-334, 2022, doi: 10.1093/bioinformatics/btab681.
31. M.H. Olyaee, A. Khanteymoori, K. Khalifeh, **Application of Chaotic Laws to Improve Haplotype Assembly Using Chaos Game Representation**, Sci Rep., 9(1), 10361, 2019, doi: 10.1038/s41598-019-46844-y.
32. B. Hewelt i inni, **The DNA walk and its demonstration of deterministic chaos-relevance to genomic alterations in lung cancer**, Bioinformatics, 35(16), 2738-2748, 2019, doi: 10.1093/bioinformatics/bty1021.
33. X. Wu i inni, **miCGR: interpretable deep neural network for predicting both site-level and gene-level functional targets of microRNA**, Brief Bioinform., 26(1), 2024, doi: 10.1093/bib/bbae616.
34. H. M. Ni, D.W. Qi, **Applying MSSIM combined chaos game representation to genome sequences analysis**, Genomics, 1110(3), 180–190, 2018, doi: 10.1016/j.ygeno.2017.09.010.

109:3884019172

35. T. Hoang, C. Yin, S.S.T. Yau, **Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison**, Genomics, 108(3-4), 134-142, 2016, doi: 10.1016/j.ygeno.2016.08.002.
36. W. Deng, Y. Luan, **Analysis of Similarity/Dissimilarity of DNA Sequences Based on Chaos Game Representation**, Abstr. Appl. Anal., 1-6, 2013, doi: 10.1155/2013/926519.
37. K. Sarapata, A. Kania, **Revealing miRNAs patterns by employing matrix representations and energy analysis**, J Mol Graph Model., 132, 2024, doi: 10.1016/j.jmgm.2024.108835.
38. S. Griffiths-Jones, **miRBase: microRNA sequences and annotation,** Curr. Protoc. Bioinformatics, 12, 2010, doi: 10.1093/nar/gky1141.
39. A. Kozomara, S. Griffiths-Jones, **miRBase: annotating high confidence microRNAs using deep sequencing data**, Nucleic Acids Research, 42, D68–D73, 2014, doi: 10.1093/nar/gkt1181.
40. CH, Chih-Hung i inni, **miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions**, Nucleic Acids Res., 46, D296–D302, 2017, doi: 10.1093/nar/gkx1067.
41. A. Zabalza i inni, **MiRNA-based therapeutic potential in multiple sclerosis**, Front Immunol., 15:1441733, 2024, doi: 10.3389/fimmu.2024.1441733.
42. H. Luo, Y. Lin, F. Gao, C.T. Zhang and R. Zhang, **DEG 10, an update of the Database of Essential Genes that includes both protein-coding genes and non-coding genomic elements,** Nucleic Acids Res., 42, D574-D580, 2014, doi: 10.1093/nar/gkt1131.
43. S. Bocker, S. Canzar, G.W. Klau, **The generalized Robinson-Foulds metric**, Algorithms in Bioinformatics, 8126, 156-169, 2013, doi:10.1007/978-3-642-40453-5_13.
44. M. Llabres i inni, **The Generalized Robinson-Foulds Distance for Phylogenetic Trees**, J Comput Biol., 28(12), 1181-1195, 2021, doi: 10.1089/cmb.2021.0342.
45. H. Ghanbarian i inni, **MicroRNA Targeting**, Methods Mol Biol., 2257:105-130, 2022 doi: 10.1007/978-1-0716-1170-8_6.
46. C. Chen i inni, **Protein Bioinformatics Databases and Resources**, Methods Mol Biol., 3-39, 2017, doi: 10.1007/978-1-4939-6783-4_1.
47. D.A. Benson i inni, **GenBank**, Nucleic Acids Res., 25-30, 2008, doi: 10.1093/nar/gkm929.
48. T. Kulikova i inni, **EMBL Nucleotide Sequence Database in 2006**, Nucleic Acids Res., 16-20, 2007, doi: 10.1093/nar/gkl913.
49. UniProt Consortium, **UniProt: a hub for protein information**, Nucleic Acids Res., 204-12, 2015, doi: 10.1093/nar/gku989.
50. R.D. Finn i inni, **The Pfam protein families database**, 36, 281-8, 2008, doi: 10.1093/nar/gkm960.
51. T. Hubbard i inni, **The Ensembl genome database project**, Nucleic Acids Res., 30(1), 38-41, 2002, doi: 10.1093/nar/30.1.38.
52. L.A. Gallagher, J. Bailey, C. Manoil, **Ranking essential bacterial processes by speed of mutant death**, Proc. Natl. Acad. Sci. U. S. A., 117(30), 18010-18017, 2020, doi: 10.1073/pnas.2001507117.
53. F.R. Fields, S. W. Lee, M.J. McConnell, **Using bacterial genomes and essential genes for the development of new antibiotics**, Biochem Pharmacol., 15(134), 74-86, 2017, doi: 10.1016/j.bcp.2016.12.002
54. T. Gánti, **The Principles of Life by Tibor Gánti,** Oxford University Press, 2003.

110:7278387999

55. D. Meinke, R. Muralla, C. Sweeney, A. Dickerman, **Identifying essential genes in Arabidopsis thaliana**, Trends Plant Sci., 13(9), 483-91, 2008, doi: 10.1016/j.tplants.2008.06.003
56. S.M. Wong, B.J. Akerley, **Identification and analysis of essential genes in Haemophilus influenzae**, Methods Mol Biol., 416, 27-44, 2008, doi: 10.1007/978-1-59745-321-9_3.
57. P. Rice, I. Longden, A. Bleasy, **EMBOSS: the European Molecular Biology Open Software Suite**, Trends Genet., 16(6), 276-7, 2000, doi: 10.1016/s0168-9525(00)02024-2.
58. A. Zielezinski i inni, **Benchmarking of alignment-free sequence comparison methods**, Genome Biol., 20(1), 144, 2009, doi: 10.1186/s13059-019-1755-7.
59. W.M. Fitch, **Toward defining the course of evolution: minimum change for a specified tree topology**, Systematic Zoology, 20(4), 406-416, 1971, doi: 10.2307/2412116.
60. A. de Bruyn i inni, **Phylogenetic reconstruction methods: an overview**, Methods Mol Biol., 257-77, 2014, doi: 10.1007/978-1-62703-767-9_13.
61. J.P. Huelsenbeck, F. Ronquist, **MRBAYES: Bayesian inference of phylogenetic trees, Bioinformatics**, 17(8), 754-5, 2001, doi: 10.1093/bioinformatics/17.8.754.
62. A. Das, X. Huang, **HPC: Hierarchical phylogeny construction,** PLoS One, 14(8), 2019, doi: 10.1371/journal.pone.0221357.
63. R.R. Sokal, C.D. Michener**, A statistical method for evaluating systematic relationships**, University of Kansas Science Bulletin. 38: 1409–1438, 1958.
64. N. Saitou, M. Nei, **The neighbor-joining method: a new method for reconstructing phylogenetic trees**, Mol Biol Evol., 4(4), 406-25, 1987, doi: 10.1093/oxfordjournals.molbev.a040454.
65. J.H. Ward, **Hierarchical Grouping to Optimize an Objective Function**, Journal of the American Statistical Association, 58, 236–244, 1963.
66. T. Horiike i inni, **HGT-Gen: a tool for generating a phylogenetic tree with horizontal gene transfer**, Bioinformation., 7(5), 211-3, 2011, doi: 10.6026/97320630007211.
67. T.R. Gregory, **Understanding Evolutionary Trees**, Evolution: Education and Outreach, 1, 121-137, 2008.
68. K.M. Kuhner, J. Yamato, **Practical Performance of Tree Comparison Metrics**, Systematic Biology, 64(2), 205-2014, 2015, doi:10.1093/sysbio/syu085.
69. D.F. Robinson, L.R. Foulds, **Comparison of phylogenetic trees**, Mathematics Biosciences, 53(1-2), 131-147, 1981, doi:10.1016/0025-5564(81)90043-2.
70. J.D. Wuitschick, K.M. Karrer, **Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in Tetrahymena thermophila**, J. Eukaryot Microbiol., 46(3), 239-47, 1999, DOI: 10.1111/j.1550-7408.1999.tb05120.x.
71. J. Romiguier i inni, **Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes**, Genome Res., 20(8), 1001-9, 2010, doi: 10.1101/gr.104372.109.
72. D. Prinbow, **Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter**, Proc Natl. Acad. Sci. USA, 72(3), 784-8, 1975, doi: 10.1073/pnas.72.3.784.
73. H Schaller, C. Gray, K. Hermann, **Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd**, Proc Natl. Acad. Sci. USA, 72(2), 737-41, 1975, doi: 10.1073/pnas.72.2.737.
74. D. Takai, P.A. Jones, **Comprehensive analysis of CpG islands in human chromosomes 21 and 22**, Proc Natl. Acad. Sci. USA, 99(6), 3749-5, 2002, doi: 10.1073/pnas.052410099.

111:8735039748

75. A.M. Deaton, A. Bird, **CpG islands and the regulation of transcription**, Genes Dev., 25(10), 1010-22, 2011, doi: 10.1101/gad.2037511.
76. X. Liao i inni, **Repetitive DNA sequence detection and its role in the human genome**, Commun Biol., 6(1), 2023, doi: 10.1038/s42003-023-05322-y.
77. F. Crick, **The Origin of the Genetic Code**, J. Mol. Biol., 1968.
78. B.G. Barrell, A.T. Bankier, J. Drouin, **A different genetic code in human mitochondria**, Nature, 282(5735), 189-94, 1979, doi: 10.1038/282189a0.
79. S. Petoukhov, **The genetic code, algebra of projection operators and problems of inherited biological ensembles**, Biology Mathematics, 2013, doi: 10.48550/arXiv.1307.7882.
80. B. Dragovich, N.Z. Misic, **The genetic code and its p-adic ultrametric modeling**, Biosystems, 246, 2024, doi: 10.1016/j.biosystems.2024.105322.
81. S.T. Parvathy i inni, **Codon usage bias**, Mol Biol Rep., 49(1), 539-565, 2022, doi: 10.1007/s11033-021-06749-4.
82. P.M. Sharp, W.H. Li, **The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications**, Nucleic Acids Res., 15(3), 1281-95, 1987, doi: 10.1093/nar/15.3.1281.
83. W. Szybalski, H. Kubinski, P. Sheldrick, **Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis**, Cold Spring Harb Symp Quant Biol., 31, 123-7, 1966, doi: 10.1101/sqb.1966.031.01.019.
84. D. Elson, E. Chargaff, **On the deoxyribonucleic acid content of sea urchin gametes**, Experientia, 8(4), 143-145, 1952, doi: 10.1007/BF02170221.
85. P. Fariselli i inni, **DNA sequence symmetries from randomness: the origin of the Chargaff's second parity rule**, Brief Bioinform., 22(2), 2172-2181, 2021, doi: 10.1093/bib/bbaa041.
86. C.E. Shannon, W. Weaver, **The Mathematical Theory of Communication**, Univ of Illinois Press, 1949.
87. T.D. Schneider, R.M. Stephens, **Sequence Logos: a New Way to Display Consensus Sequences**, Nucleic Acids Res., 18(20), 6097-100, 1990, doi: 10.1093/nar/18.20.6097.
88. G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, **WebLogo: a sequence logo generator**, Genome Research, 14(6), 1188-1190, 2004, doi: 10.1101/gr.849004.
89. U. Hesse, **K-Mer-Based Genome Size Estimation in Theory and Practice**, Methods Mol Biol., 2672:79-113, 2023, doi: 10.1007/978-1-0716-3226-0_4.
90. H.N. AlEisa, S. Hamad, A. Elhadad, **K-Mer Spectrum-Based Error Correction Algorithm for Next-Generation Sequencing Data**, Comput. Intell. Neurosci. (2022), doi: 10.1155/2022/8077664.
91. S. Das i inni, **Optimal choice of k-mer in composition vector method for genome sequence comparison**, Genomics., 110(5), 263-273, 2018, 10.1016/j.ygeno.2017.11.003.
92. E.P.C. Rocha, **The replication-related organization of bacterial genomes**, Microbiology, 150(Pt 6), 1609-1627, 2004, doi: 10.1099/mic.0.26974-0.
93. R. Zhang, C.T. Zhang, **Z Curves, An Intutive Tool for Visualizing and Analyzing the DNA Sequences**, J Biomol Struct Dyn., 11(4), 767-82, 1994, doi: 10.1080/07391102.1994.10508031.
94. R. Zhang, C.T. Zhang, a **Brief Review: The Z-curve Theory and its Application in Genome Analysis**, Curr Genomics., 15(2), 78-94, 2014, doi: 10.2174/1389202915999140328162433.

112:3015859462

95. R. Zhang, C.T. Zhang, **Single replication origin of the archaeon Methanosarcina mazei revealed by the Z curve method**, Biochem Biophys Res Commun., 297(2), 396-400, 2002, doi: 10.1016/s0006-291x(02)02214-3.
96. W.X. Zheng i inni, **Coronavirus phylogeny based on a geometric approach**, Mol Phylogenet Evol., 36(2), 224-32, 2005, doi: 10.1016/j.ympev.2005.03.030.
97. J.L. Caulfield, J.S. Wishnok, S.R. Tannenbaum, **Nitric Oxide-induced Deamination of Cytosine and Guanine in Deoxynucleosides and Oligonucleotides**, J Biol Chem., 273(21), 12689-95, 1998, doi: 10.1074/jbc.273.21.12689.
98. A. Czarniecka A., D. Bielińska-Wąż, P. Wąż, T. Clark, **20D-dynamic representation of protein sequences**, Genomics, 107(1), 16-23, 2016, doi: 10.1016/j.ygeno.2015.12.003.
99. K.C. Chou, **Prediction of protein cellular attributes using pseudo-amino acid composition**, Proteins, 43(3), 246-55, 2001, doi: 10.1002/prot.1035.
100. K.C. Chou, H.B. Shen, **PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition**, Anal Biochem., 373(2), 386-8, 2008, doi: 10.1016/j.ab.2007.10.012.
101. X. Xiao i inni, **Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor**, J. Comput. Chem., 27(4), 478-82, 2006, doi: 10.1002/jcc.20354.
102. W. Zhao i inni, a **Brief Review on Software Tools in Generating Chou's Pseudo-factor Representations for All Types of Biological Sequences**, Protein Pept. Lett., 25(9), 822-829, 2018, doi: 10.2174/0929866525666180905111124.
103. J.O. Smith, **Mathematics of the Discrete Fourier Transform (DFT)**, 2 ed., W3K Publishing, 2007.
104. C. Yin, Y. Chen, S.S.T. Yau, **A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering**, J Theor Biol., 18-28, 2014, doi: 10.1016/j.jtbi.2014.05.043.
105. C. Yin, S.S. Yau, **An improved model for whole genome phylogenetic analysis by Fourier transform**, J Theor Biol., 382, 99-110, 2015 doi: 10.1016/j.jtbi.2015.06.033.
106. H.E. Hurst, **Long-term storage capacity of reservoirs**, Transactions of the American Society of Civil Engineers, 116, 770, 1951, doi: 10.1061/TACEAT.0006518.
107. N. Marwan i inni, **Recurrence plots for the analysis of complex systems**, Phys. Reports, 438(5-6), 237-329, 2007, doi: 0.1016/j.physrep.2006.11.001.
108. J. Mitra, P. Mundra, B.D. Kulkarni, V.K. Jayaraman, **Using recurrence quantification analysis descriptors for protein sequence classification with support vector machines**, J. Biomol. Struct. Dyn., 25(3), 289-298, 2007, doi: 10.1080/07391102.2007.10507177.
109. S. Schultz, N. Spiegel, S. Marwan, Albayrak, **Approximation of diagonal line based measures in recurrence quantification analysis**, Phys. Lett., 379(14–15), 997-1011, 2015, doi: 10.1016/j.physleta.2015.01.033
110. S. Wallot, **Multidimensional Cross-Recurrence Quantification Analysis (MdCRQA) – a Method for Quantifying Correlation between Multivariate Time-Series,** Multivar. Behav. Res., 54(2), 173-191, 2018, doi: 10.1080/00273171.2018.1512846.
111. A. Kania, K. Sarapata, **The robustness of the chaos representation to mutations and its application in free-alignment methods**, Genomics, 113(3), 1428-1437, 2021, doi: 10.1016/j.ygeno.2021.03.015.
112. P. Wąż, D. Bielińska-Wąż, **Non-standard similarity/dissimilarity analysis of DNA sequences**, Genomics, 104(6), 464-471, 2014, doi: 10.1016/j.ygeno.2014.08.010.

113:6960925888

113. A. Nandy, S. Dey, S.C. Basak, D. Bielińska-Waz, P. Waz, **Characterizing the Zika Virus Genome - a Bioinformatics Study**, Curr Comput Aided Drug Des., 12(2), 87-97, 2016, doi: 10.2174/1573409912666160401115812.
114. A.J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, A.J. Marks, **MicroRNA targets in Drosophila**, Genome Biol., 5(1), 2003, doi: 10.1186/gb-2003-5-1-r1.
115. A.M. Mustoe i inni, **Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing**, Cell., 173(1), 181-195, 2018, doi: 10.1016/j.cell.2018.02.034.
116. D. Hong, S. Jeong, **3'UTR Diversity: Expanding Repertoire of RNA Alterations in Human mRNAs**, Mol Cells, 46(1), 48-56, 2023, doi: 10.14348/molcells.2023.0003.
117. C. Barreau, L. Paillard, H.B. Osborne, **AU-rich elements and associated factors: are there unifying principles?**, Nucleic Acids Res., 33(22), 7138-50, 2006, doi: 10.1093/nar/gki1012.
118. F.M. Gratacós, G. Brewer, **The role of AUF1 in regulated mRNA decay**, Wiley Interdiscip Rev RNA, 1(3), 457-73, 2010, doi: 10.1002/wrna.26.
119. M. Plass, S.H. Rasmussen, A. Krogh **, Highly accessible AU-rich regions in 3' untranslated regions are hotspots for binding of regulatory factors**, PLoS Comput Biol., 13(4), 2017, doi: 10.1371/journal.pcbi.1005460.
120. J. Qing Q i inni, **Involvement of microRNA in AU-rich element-mediated mRNA instability**, Cell., 120(5):623-34, 2005, doi: 10.1016/j.cell.2004.12.038.
121. O. Binas i inni, **Structural basis for the recognition of transiently structured AU-rich elements by Roquin**, Nucleic Acids Res., 48(13):7385-7403, 2020, doi: 10.1093/nar/gkaa465.
122. K. Rolle i inni, **The Sequence and Structure Determine the Function of Mature Human miRNAs**, PLoS One., 11(3), 2016, doi: 10.1371/journal.pone.0151246.
123. A. Kania, **Harnessing the information theory and chaos game representation for pattern searching among essential and non-essential genes in Bacteria**, J Theor Biol., 531, 110917, 2021, doi: 10.1016/j.jtbi.2021.110917.
124. Q. Zhou, S. Qi, C. Ren, **Gene essentiality prediction based on chaos game representation and spiking neural networks**, Chaos Solitons Fractals, 144, 2021, doi: 10.1016/j.chaos.2021.110649.
125. H. Shi i inni, **Identify essential genes based on clustering based synthetic minority oversampling technique**, Comput Biol Med., 153:106523, 2023 doi: 10.1016/j.compbiomed.2022.106523.
126. Y. Nakano, Y. Domon, K. Yamagishi, **Phylogenetic trees of closely related bacterial species and subspecies based on frequencies of short nucleotide sequence**, PLoS One, 18(4), 2023, doi: 10.1371/journal.pone.0268847
127. D. Jacobson, Y. Zheng, M.M. Plucinski, Y. Qvarnstrom, J.L.N. Baratt, **Evaluation of various distance computation methods for construction of haplotype-based phylogenies from large MLST datasets**, Mol Phylogenet Evol., 177, 2022, doi: 10.1016/j.ympev.2022.107608.
128. K. Sorimachi, **Visible Evolution from Primitive Organisms to Homo sapiens**, Cheminformatics and its Applications, 2020, doi: 10.5772/intechopen.91170.
129. S. Pleschka, **Overview of influenza viruses**, Curr. Top. Microbiol. Immunol., 370, 1-20, 2013, doi: 10.1007/82_2012_272.
130. G.S. Randhawa, A.H. Katleen, L. Kari, **ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels**, BCM Genomics, 20(267), 2019, doi: 10.1186/s12864-019-5571-y.

114:3672080734

131. S.L. Wong, M.B. Sukkar, **The SPARC protein: an overview of its role in lung cancer and pulmonary fibrosis and its potential role in chronic airways disease**, Br. J. Pharmacol., 174(1), 3-14, 2017, doi: 10.1111/bph.13653.
132. T. Wang, Z.G. Yu, J. Li, **CGRWDL: alignment-free phylogeny reconstruction method for viruses based on chaos game representation weighted by dynamical language model**, Front Microbiol., 15:1339156, 2024, doi: 10.3389/fmicb.2024.1339156.
133. H. Yi, L. Jin, Yi H, Jin L. **Co-phylog: an assembly-free phylogenomic approach for closely related organisms**, Nucleic Acids Res., 41(7), 2013, doi: 10.1093/nar/gkt003.
134. C. Fisher i inni, **Complete mitochondrial DNA sequences of the threadfin cichlid (Petrochromis trewavasae) and the blunthead cichlid (Tropheus moorii) and patterns of mitochondrial genome evolution in cichlid fishes**, PLoS One, 8(6), 2013, doi: 10.1371/journal.pone.0067048.
135. G. Bernard, C.X. Chan, M.A. Ragan, **Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer**, Sci Rep., 6, 2016, doi: 10.1038/srep28970.
136. R.G. Beiko, R.L. **Charlebois, a simulation test bed for hypotheses of genome evolution**, Bioinformatics, 2007, 23(7), 825-31, doi: 10.1093/bioinformatics/btm024.
137. R.D.M. Page, **Visualizing phylogenetic trees using TreeView**, Curr Protoc Bioinformatics, 6, 2002, doi: doi: 10.1002/0471250953.bi0602s01.
138. G. Cardona, F. Rossello, G. Valiente, **Extended Newick: it is time for a standard representation of phylogenetic networks**, BMC Bioinformatics, 9, 2008, doi: 10.1186/1471-2105-9-532.
139. A. Kania, K. Sarapata, **Multifarious aspects of the chaos game representation and its applications in biological sequence analysis**, Comput Biol Med., 151(PtA), 2022, doi: 10.1016/j.compbiomed.2022.106243.
140. R. Lorenz i inni, **ViennaRNA Package 2.0**, Algorithms Mol Biol., 6(26), 2011, 10.1186/1748-7188-6-26.
141. M. Zuker, P. Stiegler, **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information**, Nucleic Acids Res., 9(1), 133-148, 1981, doi: 10.1093/nar/9.1.133
142. S. Wolfsheimer, A.K. Hartmann, **Minimum-free-energy distribution of RNA secondary structures: Entropic and thermodynamic properties of rare events**, Phys. Rev. E, 82, 2010, doi: 10.1103/PhysRevE.82.021902
143. E.C. Koper-Zwarthoff, F.T. Brederode, P. Walstra, J.F. Bol**, Nucleotide sequence of the 3'-noncoding region of alfalfa mosaic virus RNA 4 and its homology with the genomic RNAs**, Nucleic Acids Res., 7(7), 1887-1900, 1979, doi: 0.1093/nar/7.7.1887.
144. M. Hamada, H. Kiryu, K. Sato, T. MItuyama, K. Asai, **Prediction of RNA secondary structure using generalized centroid estimators**, Bioinformatics, 25(4), 465-473, 2009, doi: 10.1093/bioinformatics/btn601.
145. J. Feng, T. Wang, **A 3D graphical representation of RNA secondary structures based on chaos game representation**, Chemical Physics Letters, 454(4-6), 355-361, 2008, doi: 10.1016/j.cplett.2008.01.041.
146. J. Sankalp, S. Dipankar, R. Piyush, K.N. Pradeep, **TpPred: a Tool for Hierarchical Prediction of Transport Proteins Using Cluster of Neural Networks and Sequence Derived Features**, International Journal for Computing Biology, 1(1), 46-60, 2012, doi: 10.34040/IJCB.1.1.2012.18.

115:3096819160

147. C. Tanford, **Contribution of hydrophobic interactions to the stability of the globular conformation of proteins**, J. American Chem. Society, 84: 4240-4247, 1962, doi: 10.1021/ja00881a009.
148. T .Tp. Hopp, K.R. Woods, **Prediction of protein antigenic determinants from amino acid sequences**, Proc. Natl. Acad. Science, 78, 3824-3828, 1981, doi: 10.1073/pnas.78.6.3824
149. C.J.A. Sigrist i inni, **New and continuing developments at PROSITE**, Nucleic Acids Res., 41, D344-7, 2013, doi: 10.1093/nar/gks1067.
150. S. Kawashima, H. Ogata, M. Kanehisa, **Aaindex: amino acid index database**, Nucleic Acids Res., 27, 368-369, 1999, doi: 10.1093/nar/27.1.368.
151. S. Kawashima, M. Kanehisa, **Aaindex: amino acid index database**, Nucleic Acids Res., 28, 374, 2000, doi: 10.1093/nar/28.1.374.
152. S. Kawashima i inni, **AAindex: amino acid index database, progress report** 2008, Nucleic Acids Res., 36, D202-D205, 2008, 10.1093/nar/gkm998.
153. S. Osowski, **Metody i narzędzia eksploracji danych**, BTC, 2017.
154. F. Almeida, G. Xexeco, **Word embeddings: a survey**, Computer Science, 2019, doi: 10.48550/arXiv.1901.09069.
155. G. Cybenko, **Approximation by superpositions of a sigmoidal function**, Mathematics of Control, Signals, and Systems. 2 (4), 303–314, 1989, doi: 10.1007/BF02551274.
156. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, **Basic local alignment search tool**, J. Mol. Biol., 215, 403-410, 1990, doi: 10.1016/S0022-2836(05)80360-2
157. G.M. Boratyn, C. Camacho i inni, **BLAST: a more efficient report with usability improvements**, Nucleic Acids Res., 41, 2013, doi: 10.1093/nar/gkt282.
158. D. Blumenkrantz, K.L. Roberts, H. Shelton, S. Lycett, W.S. Barclay, **The short stalk length of highly pathogenic avian influenza H5N1 virus neuraminidase limits transmission of pandemic H1N1 virus in ferrets**, J. Virol., 87, 10539-10551, 2013, doi: 10.1128/JVI.00967-13.
159. S. Kerscher, S. Drose, V. Zickermann, U. Brandt, **The three families of respiratory NADH dehydrogenases**, Results Probl. Cell Differ., 45, 185-222, 2008, 10.1007/400_2007_028.

116:2337885472

Spis rysunków

Rysunek 1. Logo wygenerowano dla 200 sekwencji nukleotydowych o długości 3. Wszystkie sekwencje posiadały A na pierwszej pozycji. Na drugiej pozycji 100 sekwencji posiadało G, a pozostałe T. Na trzeciej pozycji C wystąpiło w 100 sekwencjach, T w 50, A w 25 oraz G w 25.

Rysunek 2. Zawartość GC (GC %) oraz parametr asymetrii G/C (GC skew) dla kolejnych podsekwencji ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT. Przyjęto długość okna 10.

Rysunek 3. Krzywa Z reprezentująca genom Methanosarcina mazei (NCBI ID: NC_003901.1). Na osiach OX, OY oraz OZ przedstawione są kolejne pozycje w genomie w skali tysiąca par zasad (kb).

Rysunek 4. Reprezentacja gry chaosu dla sekwencji AACTG. Startujemy z punktu $(0,0)$, a każdy kolejny kładziemy w połowie odległości pomiędzy poprzednim punktem a aktualnym nukleotydem (symbolizowany przez wierzchołek kwadratu).

Rysunek 5. Zasady azotowe występujące w DNA. Struktury pochodzą z bazy danych PubChem.

Rysunek 6. CGR dla losowej sekwencji nukleotydowej o długości 2000 nt. Obserwujemy równomierny rozkład punktów.

Rysunek 7. CGR dla losowej sekwencji nukleotydowej o długości 2000 nt. która nie zawiera C.

Rysunek 8. CGR dla losowej sekwencji nukleotydowej o długości 10000 nt. która nie zawiera AT.

Rysunek 9. CGR dla losowej sekwencji nukleotydowej o długości 10000 nt. która nie zawiera AG.

Rysunek 10. Reprezentacje gry chaosu dla losowej sekwencji z alfabetu 3, 4 oraz 5 literowego. Tylko w przypadku 4 liter losowa sekwencja wypełnia jednorodnie wnętrze wielokąta.

Rysunek 11. Genomacierz wygenerowana dla sekwencji AACTG.

Rysunek 12. Genomacierz wygenerowana dla losowej sekwencji o długości 2000 nt. Kolejne punkty są na tyle blisko, że różnice przestają być widoczne.

Rysunek 13. Genomacierz przedstawiająca 1000 losowych sekwencji o długości 100 nt. każda, gdzie nie występuje podsekwencja AT.

Rysunek 14. Genomacierz przedstawiająca 1000 losowych sekwencji o długości 100 nt. każda, gdzie nie występuje podsekwencja AG.

Rysunek 15. Genomacierz przedstawiająca 1000 losowych sekwencji o długości 100 nt. każda, gdzie AT występuje na trzeciej pozycji.

Rysunek 16. Spacer DNA dla sekwencji GATAC. Kolejnymi liczbami naturalnymi zaznaczono punkty odpowiadające kolejnym resztom w sekwencji (pozycja). Zauważmy, że dwa z położeń zajmowane są dwukrotnie.

Rysunek 17. Spacer DNA dla sekwencji GGGGGGCGCGCAAAA. Dla przejrzystości rysunku nie zaznaczono tym razem numeracji kolejnych punktów. Punkty zajmowane wielokrotnie zwiększono.

Rysunek 18. Spacer DNA dla dwóch losowych sekwencji nukleotydowych o długości 1000 każda.

117:8582030007

Rysunek 19. Spacer DNA dla mRNA genu SPARC (NCBI ID: NM_001309443.2).

Rysunek 20. Wygenerowano dwa ciągi liczbowe według tej samej formuły – 200 wartości (czerwone punkty) oraz 20 wartości (zielone punkty). Zbyt małe próbkowanie może spowodować brak możliwości rekonstrukcji oryginalnego ciągu czy jego własności.

Rysunek 21. Wartości ciągu x_t określonego zależnością $sin(50 * 2\pi t) + \frac{1}{2} sin(250 * 2\pi t) + \varepsilon_t$, gdzie rozważano czas z przedziału [0,1].

Rysunek 22. Wykres znormalizowanych modułów z wartości ciągu Y_k. Pojawiają się dwie charakterystyczne częstości przy 50 oraz 250. Wykres przycięto w połowie z powodu symetrii.

Rysunek 23. Widma mocy o długości 800 (rysunek z lewej, przycięto w połowie) wydłużono do 1200 (rysunek z prawej, przycięto w połowie). Na pierwszym rysunku pik znajduje się na pozycji 50, czyli w 50/400 = 0.125 części; podobnie na drugim wykresie na 75 pozycji, czyli 75/600= 0.125 części.

Rysunek 24. Przykład szeregu losowego, dla którego współczynnik Hursta wynosi 0.47. Następują tutaj losowe zmiany w kolejnych chwilach czasowych.

Rysunek 25. Przykład szeregu losowego dla którego współczynnik Hursta wynosi 0.82 (tzw. szereg persystentny). Pewne trendy utrzymują się przez określony okres.

Rysunek 26. Przykład szeregu losowego dla którego współczynnik Hursta wynosi 0.26 (tzw. szereg antypersystentny). Obserwujemy tutaj nagłe zmiany i odwracanie trendu.

Rysunek 27. CGR dla wszystkich sekwencji ludzkich miRNA. Taki obraz pozwala na globalną ocenę składu tych cząsteczek. Rysunek opublikowano w J Mol Graph Model. w 2024.

Rysunek 28. Częstość występowania dinukleotydów pośród ludzkich miRNA. Wyróżniono przy tym podział miRNA na tzw. region seed (pierwsze 7 nt., oznaczono ciemniejszym kolorem) oraz pozostałą część (oznaczono jaśniejszym kolorem). Rysunek opublikowano w J Mol Graph Model. w 2024.

Rysunek 29. CGR dla wszystkich sekwencji ludzkich miRNA. Te, po analizie energetycznej, podzielono ten zbiór na dwa podzbiory – niskoenergetyczne (po lewej) oraz wysokoenergetyczne (po prawej). Rysunek opublikowano w J Mol Graph Model. w 2024.

Rysunek 30. CGR dla podzbioru wszystkich miRNA o największej wiarygodności (ang. confident) oraz konserwatywnych (ang. conserved). Tego typu sekwencje są dostępne poprzez miRBase oraz miRTarBase. Rysunek opublikowano w J Mol Graph Model. w 2024.

Rysunek 31. Częstość występowania dinukleotydów pośród ludzkich miRNA. Te, po analizie energetycznej podzielono na dwa podzbiory – niskoenergetyczne (po lewej) oraz wysokoenergetyczne (po prawej). Wyróżniono przy tym podział miRNA na tzw. region seed (pierwsze 7 nt., oznaczono ciemniejszym kolorem) oraz pozostałą część (oznaczono jaśniejszym kolorem). Rysunek opublikowano w J Mol Graph Model. w 2024.

Rysunek 32. Stosunek częstości występowania dinukleotydów pośród miRNA wysokoenergetycznych oraz niskoenergetycznych. Dodatkowo, od wyznaczonego stosunku odjęto 1. Rysunek opublikowano w J Mol Graph Model. w 2024.

118:4110811372

Rysunek 33. Zależność pomiędzy zawartością A i C oraz G i C dla genów niezbędnych i opcjonalnych. Jedna kropka symbolizuje jeden gen. Kolorem czerwonym oznaczone są geny niezbędne, a niebieskim geny opcjonalne. Rysunek opublikowano w J Theor Biol. w 2021.

Rysunek 34. Zależność pomiędzy A i C oraz G i C dla pseudogenomów złożonych tylko z genów niezbędnych lub opcjonalnych. Jedna kropka symbolizuje jedną bakterię. Rysunek opublikowano w J Theor Biol. w 2021.

Rysunek 35. Rozkład entropii wyznaczonej dla każdego genu. Wyróżniono geny niezbędne i opcjonalne. Wykres częściowo przycięto z lewej strony. Rysunek opublikowano w J Theor Biol. w 2021.

Rysunek 36. Dywergencja Kullbacka-Leiblera da pseudogenomów złożonych z genów niezbędnych (E) oraz całych genomów (ALL). Porównywano korespondujące pseudogenomy i genomy oraz te przypadające dla różnych organizmów. Wykres częściowo przycięto na osi OY. Rysunek opublikowano w J Theor Biol. w 2021.

Rysunek 37. Spacery DNA dla 45 bakterii, dla których dostępne były zarówno zbiory genów niezbędnych oraz opcjonalnych. Każda składowa związana jest z jednym genem. Geny niezbędne oznaczono kolorem czerwonym, opcjonalne niebieskim. Rysunek opublikowano w J Theor Biol. w 2021.

Rysunek 38. Oryginalne sekwencje jednej z bakterii (a); geny niezbędne zastąpione losowymi o tym samym rozkładzie długości (b); geny opcjonalne zostały zastąpione sekwencjami losowymi o tym samym rozkładzie długości (c); zarówno sekwencje niezbędne jak i optymalne zostały zastąpione sekwencjami losowymi (d). Rysunek opublikowano w J Theor Biol. w 2021.

Rysunek 39. Zależność pomiędzy długością genu o odległością do końca reprezentacji. Każdy punkt związany jest z jedną sekwencją. Generalnie, wraz ze wzrostem długości genu obserwujemy wzrost odległości od końca. Dane przedstawiono w skali logarytmicznej. Rysunek opublikowano w J Theor Biol. w 2021.

Rysunek 40. Średni poziom R, czyli odległości do końca reprezentacji, pośród genów niezbędnych i opcjonalnych dla kolejnych bakterii. Dodatkowo, wygenerowano losowe sekwencje z tego samego rozkładu długości. Rysunek opublikowano w J Theor Biol. w 2021.

Rysunek 41. Zależności pomiędzy średnim poziomem R, czyli odległości do końca reprezentacji, dla losowych sekwencji (a); oryginalnych sekwencji (b); stosunków oryginalnych sekwencji do losowych (c). Rysunek opublikowano w J Theor Biol. w 2021.

Rysunek 42. Zależność pomiędzy r1 a r2. Zastosowano dodatkowe transformacje dla wygodniejszej reprezentacji punktów. Czerwonym kolorem oznaczono geny niezbędne, niebieskim opcjonalne. Rysunek opublikowano w J Theor Biol. w 2021.

Rysunek 43. Średni poziom r1 dla kolejnych bakterii. Czerwonym kolorem oznaczono geny niezbędne, niebieskim opcjonalne.

Rysunek 44. Średni poziom r2 dla kolejnych bakterii. Czerwonym kolorem oznaczono geny niezbędne, niebieskim opcjonalne.

Rysunek 45. Schemat działania prezentowanych metod. Sekwencje najpierw przetwarzane są za pomocą CGR. Następnie, te reprezentacje traktujemy jako obrazy (w przypadku dalszego porównania za pomocą SSIM) lub ciągi liczbowe (DFT i wykładnik Hursta).

119:9293052767

Rysunek 46. Zestaw sekwencji powstałych z jednej, początkowej sekwencji wskutek mutacji. Numeracja wskazuje na kierunek powstania. W formacie Newick można je przedstawić jako (((seq1111,seq1112),(seq1121,seq1122)),((seq1211,seq1212),(seq1221,seq1222)),((seq2111,seq2112),(seq2121,seq2122)),((seq2211,seq2212),(seq2221,seq2222))) i stanowi ono znaną referencję.

Rysunek 47. Macierz BLOSUM62 oraz macierz powstała wskutek transformacji SoftMax, przekształcając te wartości na prawdopodobieństwa.

Rysunek 48. Nieznormalizowane odległości JRF dla losowych topologii drzew.

Rysunek 49. Zestaw sekwencji powstałych z jednej, początkowej sekwencji wskutek mutacji. Numeracja wskazuje na kierunek powstania. Część sekwencji jest nieznana (nieobserwowana). W formacie Newick można je przedstawić jako (((seq1222)),((seq2111,seq2112),(seq2121,seq2122)),((seq2211,seq2212),(seq2221,seq2222))) i stanowi ono referencję.

Rysunek 50. Nieznormalizowane odległości JRF dla losowych topologii drzew.

Rysunek 51. Względne czasy (w sekundach) porównania dwóch sekwencji w zależności od ich długości i rozważanej metody.

Rysunek 52. CGR dla pierwszej sekwencji z każdego z zestawów. W zależności od długości sekwencji dobrano różny rozmiar kropek przy wizualizacji.

Rysunek 53. Wpływ reszt w sekwencji (niebieski) oraz punktu startowego (pomarańczowy). Widzimy, że dla kolejnych pozycji wpływ punktu startowego maleje bardzo szybko. Rysunek opublikowano w Genomics w 2021.

Rysunek 54. Waga obecnej i poprzednich pozycji dla różnych funkcji ważących. Przykładowo, 1 na osi OX oznacza tutaj poprzednią resztę względem aktualnej. Rysunek opublikowano w Genomics w 2021.

Rysunek 55. Jednowymiarowe wykresy ukazujące rozkład punktów dla rozważanych funkcji ważących. u samej góry znajduje się oryginalna funkcja $\frac{1}{2^n}$. Rysunek opublikowano w Genomics w 2021.

Rysunek 56. Przedstawienie losowej sekwencji o długości 2000 w reprezentacji CGR przy różnych funkcji ważących. Wygenerowano jednostronne (lewe) reprezentacje oraz te uwzględniające zarówno poprzednie jak i kolejne reszty (obustronne). Rysunek opublikowano w Genomics w 2021.

Rysunek 57. Dla sekwencji AUGCUCAUGCAAAA przedstawiono jej strukturę drugorzędową oraz reprezentacje CGR (czerwony) oraz CG-RNA (niebieski). W formacie Vienna [140] strukturę drugorzędową można przedstawić tutaj jako .(((....)))..., gdzie nawias oznacza sparowany nukleotyd. Sparowane nukleotydy przesunięte są bliżej centrum w przypadku CG-RNA, co zostało zaznaczone przerywanymi liniami. Rysunek opublikowano w Comput Biol Med. w 2022.

Rysunek 58. Zależność pomiędzy E1 i E2 dla 100 losowo wygenerowanych sekwencji. Struktury drugorzędowe zostały przewidziane z użyciem algorytmu MFE. Rysunek opublikowano w Comput Biol Med. w 2022.

120:6102240129

Rysunek 59. Zależność pomiędzy E1, E2 a ilością sparowanych nukleotydów dla kolejnych zagnieżdżonych struktur sekwencji zawierającej ciąg powtórzeń A oraz T. Rysunek opublikowano w Comput Biol Med. w 2022.

Rysunek 60. Rozkłady energii E1 oraz E2 dla różnych struktur. Struktury podzielone zostały na te z dużą przerwą (>7) oraz małą (<4). Rysunek opublikowano w Comput Biol Med. w 2022.

Rysunek 61. Hydrofobowość i hydrofilowość aminokwasów. Współrzędne punktu startowego określają odpowiednie średnie danej własności. Rysunek opublikowano w Comput Biol Med. w 2022.

Rysunek 62. Zależność pomiędzy wariancją a średnią dla kolejnych własności dla rozważanych rodzin białkowych. Rysunek opublikowano w Comput Biol Med. w 2022.

Rysunek 63. Zależność pomiędzy objaśnianą wariancją a pozostawioną liczbą cech. Rysunek opublikowano w Comput Biol Med. w 2022.

Rysunek 64. Sieć neuronowa do predykcji środkowego aminokwasu a także do nauki reprezentacji słów. Podana sieć składa się z jednej warstwy ukrytej (ang. hidden layer). Jako funkcje aktywacji wybrano ReLU oraz LogSoftmax. Rysunek opublikowano w Comput Biol Med. w 2022.

Rysunek 65. Wpływ różnych mutacji na reprezentację CGR. Czarne kropki wskazują na oryginalną sekwencję. w ostatnim przypadku zastosowano wszystkie rodzaje mutacji na raz.

Rysunek 66. Wartość rankingu po zastosowaniu różnych funkcji wagowych dla przykładowej sekwencji o długości 100 oraz zmienionej. Wprowadzono 5 substytucji na pozycjach 9, 35, 56, 57 oraz 94. Rysunek opublikowano w Genomics w 2021.

Rysunek 67. Drzewa filogenetyczne w oparciu o CGR dla 8 sekwencji mRNA genu SPARC. Rozważano 4 funkcje wagowe (a-d), klasyczną $1/2^x$, $1/x^2$, $1/(1+e^x)$ oraz $(1+1/x)^{-x^{\wedge}1.5}$ odpowiednio. Rysunek opublikowano w Genomics w 2021.

Rysunek 68. Drzewo filogenetyczne skonstruowane w oparciu o CGR oraz DFT za pomocą metody UPGMA. Wybrano klasyczną funkcję wagową $f(x)=1/2^x$. Rysunek opublikowano w Genomics w 2021.

Rysunek 69. Drzewo filogenetyczne skonstruowane w oparciu o CGR oraz DFT za pomocą metody UPGMA. Jako funkcje wagowo wybrano $f(x)=1/x^2$($f(x) = 1/(1+e^x)$). Rysunek opublikowano w Genomics w 2021.

Rysunek 70. Drzewo filogenetyczne skonstruowane w oparciu o CGR oraz DFT za pomocą metody UPGMA. Jako funkcję wagową wybrano $f(x) = (1+1/x)^{-x^{\wedge}1.5}$. Rysunek opublikowano w Genomics w 2021.

Rysunek 71. Drzewo filogenetyczne dla sekwencji ND5 w oparciu o CGR i DFT. Aminokwasy sparametryzowano w sposób losowy. Rysunek opublikowano w Comput Biol Med. w 2022.

Rysunek 72. Drzewo filogenetyczne dla sekwencji ND5 w oparciu o CGR i DFT. Aminokwasy sparametryzowano za pomocą ich hydrofobowości i hydrofilności. Rysunek opublikowano w Comput Biol Med. w 2022.

121:4615190413

Rysunek 73. Drzewo filogenetyczne dla sekwencji ND5 w oparciu o CGR i DFT. Aminokwasy sparametryzowano za pomocą techniki wektorów osadzenia. Rysunek opublikowano w Comput Biol Med. w 2022.

Rysunek 74. Drzewo filogenetyczne dla sekwencji SPARC w oparciu o CGR i DFT. Aminokwasy sparametryzowano za pomocą ich hydrofobowości i hydrofilności. Rysunek opublikowano w Comput Biol Med. w 2022.

Rysunek 75. Drzewo filogenetyczne dla sekwencji SPARC w oparciu o CGR i DFT. Aminokwasy sparametryzowano za pomocą techniki wektorów osadzenia. Rysunek opublikowano w Comput Biol Med. w 2022.

122:3164387518

Spis tabel

Tabela 1. Porównanie częstości występowania podsekwencji i do nich komplementarnych oraz odwrotnie komplementarnych dla genomu E. coli (NCBI ID: HG738867.1). Podano średnie wartości stosunków częstości oraz odchylenia.

Tabela 2. Własności fizykochemiczne oraz cechy zasad azotowych występujących w DNA wyznaczają naturalne ich podziały. Te stanowią potencjalne parametryzacje CGR – ułożenie nukleotydów w rogach kwadratu.

Tabela 3. Reprezentacja PseAAC dla ludzkiego białka SPARC. Wyniki wygenerowano za pomocą narzędzia dostępnego online (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/).

Tabela 4. Rozważany zbiór bakterii. Oprócz nazwy organizmu podano również liczbę genów niezbędnych i opcjonalnych. W przypadku trzech bakterii zbiór genów opcjonalnych nie był dostępny. Tabelę opublikowano w J Theor Biol. w 2021.

Tabela 5. Współczynnik korelacji pomiędzy grupą genów niezbędnych i opcjonalnych dla kolejnych bakterii, gdzie rozważano kolejne kwartyle oraz średnią ich długość. Niskie współczynniki oraz wysokie wartości prawdopodobieństwa testowego (p-value) potwierdzają brak liniowej zależności pomiędzy tymi grupami genów. Tabelę opublikowano w J Theor Biol. w 2021.

Tabela 6. Wartości oszacowanych współczynników a i b dla modeli regresji liniowej $y = ax + b$. Podano również prawdopodobieństwo testowe (p-value) oraz ocenę modelu (R^2). W nawiasach znajdują się błędy oszacowań. Gwiazdką oznaczono parametry oszacowane dla kontroli. Tabelę opublikowano w J Theor Biol. w 2021.

Tabela 7. Wartości oczekiwane i odchylenia standardowe dla każdego z nukleotydów pośród genów niezbędnych i opcjonalnych. Tabelę opublikowano w J Theor Biol. w 2021.

Tabela 8. Zależności liniowe dla średnich i odchyleń standardowych rozkładu przerw dla kolejnych nukleotydów. Wartości te liczone były dla każdej bakterii – pomiędzy zbiorem genów niezbędnych i opcjonalnych. Gwiazdką oznaczono parametry oszacowane dla kontroli. Tabelę opublikowano w J Theor Biol. w 2021.

Tabela 9. Wartości współczynników modelu $y = ax + b$, gdzie badano odległość od końca reprezentacji między pulą genów opcjonalnych i niezbędnych. Dodatkowo umieszczono informację o prawdopodobieństwie testowym (p-value) i jakości modelu (R^2). Gwiazdką oznaczono parametry oszacowane dla kontroli. Tabelę opublikowano w J Theor Biol. w 2021.

Tabela 10. Zestaw różnych metod używanych do generowania drzew filogenetycznych. W tabeli przedstawiono skróty nazw rozważanych metod. Wyróżniono w jaki sposób były analizowane sekwencje oraz jak wygenerowano ostateczne drzewo.

Tabela 11. Rozważane zbiory sekwencji w zależności od typu sekwencji, długości sekwencji pierwotnej oraz stopnia podobieństwa kolejnych sekwencji potomnych.

Tabela 12. Nieunormowane odległości JRF dla różnych zestawów sekwencji oraz rozważanych metod. W przypadku sekwencji białkowych rozważano dwa typy CGR – oparty o hydrofobowość/hydrofilowość lub redukcja do 4 liter (oznaczone gwiazdką).

123:7737615185

Tabela 13. Nieznormalizowane odległości JRF dla różnych zestawów sekwencji oraz rozważanych metod. W przypadku sekwencji białkowych rozważano dwa typy CGR – oparty o hydrofobowość/hydrofilowość lub redukcja do 4 liter (oznaczone gwiazdką).

Tabela 14. Sekwencje neuraminidazy różnych wirusów grypy. Prócz nazwy wirusa podano również skróconą nazwę oraz identyfikator z NCBI.

Tabela 15. Klasyfikacja wirusów grypy w oparciu o metody CDU i CDW odpowiednio.

Tabela 16. Klasyfikacja wirusów grypy w oparciu o metody CHU i CHW odpowiednio.

Tabela 17. Klasyfikacja wirusów grypy w oparciu o metody CSU i CSW odpowiednio.

Tabela 18. Klasyfikacja wirusów grypy w oparciu o metody LZU i LZW odpowiednio.

Tabela 19. Rozważane organizmy, dla których rozważano sekwencje nukleotydowe oraz aminokwasowe SPARC z odniesieniami ID do NCBI. Pozostawiono angielskie nazwy.

Tabela 20. Porównanie topologii drzew skonstruowanych w oparciu o sekwencje nukleotydowe i aminokwasowe. W formacie Newick przedstawiono topologie obydwu z drzew.

Tabela 21. Rozważane zestawy referencyjne. W przypadkach 3., 4. oraz 5. rozważany był horyzontalny transfer genów. Ostatni z zestawów wygenerowano za pomocą EvolSimulator i składa się z 5 podzestawów.

Tabela 22. Zestawienie podstawowych informacji podsumowujących zestawy takich jak liczba sekwencji czy średnia długość sekwencji w zestawie. Ostatnie trzy kolumny zawierają wartość unormowanej odległości Robinsona-Fouldsa (nRF). Dla ostatniego zestawu podano wynik średni z podzestawów.

Tabela 23. Wpływ wszystkich dotychczasowych nukleotydów oraz punktu startowego na wartość w danym punkcie. Tabelę opublikowano w Genomics w 2021.

Tabela 24. Struktury dla ALMV-3 wygenerowane za pomocą algorytmu MFE oraz odpowiednie oceny E1 i E2 tych struktur. Narzucony został warunek parowania 12 lub 9 par nukleotydów. W pierwszym przypadku, średnie energie wynoszą 0.24 ($\pm$0.01) i -0.38 ($\pm$0.01) odpowiednio. W drugim przypadku, są to już 0.33 ($\pm$0.03) oraz -0.29 ($\pm$0.03). Tabelę opublikowano w Comput Biol Med. w 2022.

Tabela 25. Wartości E_1 oraz E_2 dla struktur ALMV-3. Porównano strukturę biologiczną (referencja) oraz przewidywania za pomocą MFE oraz metody centroidów. Zauważmy, że MFE i referencją są porównywalne. Co ciekawe, jeżeli znormalizujemy wartości E2 (tzn. kolejne wartości zostaną jeszcze podzielone przez współrzędne CGR, oznaczone gwiazdką), to taka wielkość ma najniższą wartość dla referencji. Tabelę opublikowano w Comput Biol Med. w 2022.

Tabela 26. Porównanie techniki emb-CGR z klasycznym CGR oraz PseAAC. Tabelę opublikowano w Comput Biol Med. w 2022.

Tabela 27. Porównanie klasycznego CGR z trzema funkcjami ważącymi przy pomocy S_c. Średnie zostały policzone tylko dla wartości niezerowych. Dodatkowo, w nawiasach podano średnie dla dodatnich i ujemnych wartości. Tabelę opublikowano w Genomics w 2021.

124:1041777115

Tabela 28. Porównanie klasycznego CGR z obustronnym CGR z czterema funkcjami ważącymi przy pomocy S_c. Średnie zostały policzone tylko dla wartości niezerowych. Dodatkowo, w nawiasach podano średnie dla dodatnich i ujemnych wartości. Tabelę opublikowano w Genomics w 2021.

Tabela 29. Rozważane organizmy wraz z NCBI ID sekwencji SPARC. Tabelę opublikowano w Genomics w 2021.

Tabela 30. Procent identyczności sekwencji białka neuraminidazy pomiędzy A/turkey/VA/505477–18/2007(H5N1) a innymi wybranymi wirusami H5N1/H1N1. Tabelę opublikowano w Genomics w 2021

Tabela 31. Procent identyczności sekwencji białka neuraminidazy pomiędzy A/duck/Guangxi/030D/2009(H1N1) a innymi wybranymi wirusami H5N1/H1N1. Tabelę opublikowano w Genomics w 2021.

Tabela 32. Rozważane organizmy. Podano również ID z NCBI. Tabelę opublikowano w Comput Biol Med. w 2022.

Tabela 33. Rozważane organizmy. Podano również ID z NCBI. Tabelę opublikowano w Comput Biol Med. w 2022.

125:9204164638