Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Tomasz Gajewski
Nr albumu: 167269

Tablice substytucyjne dla modelu przyrównania kontekstowego

Praca magisterska
na kierunku **INFORMATYKA**

Praca wykonana pod kierunkiem
dr Anny Gambin
Instytut Informatyki

sierpień 2001

Pracę przedkładam do oceny

Data Podpis autora pracy

Praca jest gotowa do oceny przez recenzenta

Data Podpis kierującego pracą

Streszczenie

W pracy tej rozważane są metody oceniania prawdopodobieństw zajść zdarzeń ewolucyjnych takich jak substytucja, insercja, czy delecja w modelu, w którym zależą one od aminokwasów otaczających mutujące pozycje. W oparciu o wcześniej istniejące algorytmy dla tablic substytucyjnych w wersji bezkontekstowej został zaproponowany algorytm dla przypadku kontekstowego. Poczynione zostały pewne obserwacje, z których wnioski zostały przedstawione w niniejszej pracy.

Klasyfikacja według *ACM Computing Classification System*:
I. Computing Methodologies
I.5 Pattern Recognition
I.5.3 Clustering—Similarity measures

Słowa kluczowe:
przyrównanie kontekstowe, uliniowienie kontekstowe, podobieństwo sekwencji, tablice substytucyjne, PAM, BLOSUM

Spis treści

1 Wstęp **5**
1.1 Przyrównanie (uliniowienie) 5
1.2 Tablice substytucyjne 6
1.2.1 PAM 7
1.2.2 BLOSUM 7

2 Przyrównanie kontekstowe **9**

3 Tablice do przyrównania kontekstowego **11**
3.1 Zarys metody 11
3.2 Szczegółowy opis 12
3.3 Częstości wystąpień aminokwasów w kontekstach 13
3.4 Obserwowane częstości wystąpień substytucji 14
3.5 Oczekiwane częstości wystąpień substytucji 14
3.6 Wartości w tablicach substytucyjnych 15

4 Program **17**
4.1 Parametry 17
4.2 Opis plików wejściowych 18
4.2.1 Plik konwersji aminokwasów na konteksty 18
4.2.2 Pliki bloków 18
4.3 Opis plików wynikowych 19
4.3.1 Plik dla programu przyrównującego 19
4.3.2 Plik czytelny dla człowieka 19
4.4 Złożoność 20

5 Otrzymane tablice **23**
5.1 Bazy bloków 23
5.2 Funkcje definiujące kontekst 24

6 Obserwacje **25**
6.1 Porównanie wygenerowanej programem tablicy z tablicą BLOSUM62 . . . 25
6.2 Wpływ zmiany parametrów programu na otrzymane tablice i obliczone przyrównania . . . 26
6.3 Rozkład różnic wartości przyrównań bez kontekstu i w kontekście ‘hp’ . . . 29
6.4 Własność trójkąta w tablicach . . . 29

7 Otwarte problemy **31**
7.1 Interpretacja zygzaka . . . 31
7.2 Wyeliminowanie niesymetryczności tablic . . . 32
7.3 Dobór funkcji definiującej kontekst . . . 32
7.4 Metodologia oceniania osiągniętych wyników . . . 33

Rozdział 1

Wstęp

Biologia jest nauką o życiu i tym samym ma ogromne znaczenie dla człowieka. Przez lata ewoluowała wraz z rosnącymi możliwościami technicznymi, dającymi naukowcom — biologom, coraz to inne narzędzia i metody do zgłębiania tajemnic działania organizmów. W ostatnich latach możemy zaobserwować bardzo dynamiczny rozwój dziedzin biologii związanych z funkcjonowaniem organizmów na poziomie molekularnym. Biochemia oraz biologia molekularna, bo to o nich tutaj mowa, dopracowały się ostatnio metod i narzędzi umożliwiających zbieranie ogromnych ilości informacji. Ciągi nukleotydów tworzących kwasy nukleinowe oraz ciągi aminokwasów tworzące białka, mogą być w dzisiejszych czasach stosunkowo łatwo i tanio odczytane. Niestety informacja na poziomie molekuł nie przekłada się w łatwy sposób na wiedzę o funkcjach poszczególnych struktur, które tworzą. Tu powstaje miejsce na współpracę biologów z informatykami, aby móc wykorzystać moc dzisiejszych komputerów w celu zgłębienia funkcji i struktur białek, korzystając z informacji o składzie, bez konieczności wykonywania drogich, żmudnych i czasochłonnych doświadczeń.

1.1 Przyrównanie (uliniowienie)

Jednym z tematów, w których współpraca informatyków z biologami przyniosła pozytywne efekty jest problem stwierdzenia, czy para sekwencji jest ewolucyjnie związana ze sobą — czyli innymi słowy, czy jest prawdopodobne, że wywodzą się ze wspólnego przodka.

Problem ten nie jest trywialny ze względu na procesy zachodzące w organizmach na przestrzeni tysiącleci, które wprowadzają i utrwalają zmiany w materiale genetycznym. Chaos w sekwencjach powodowany przez losowe mutacje, krzyżowanie i naturalną selekcję sprawia, że jedynymi znanymi dziś metodami badania związków pomiędzy sekwencjami są metody probabilistyczne i staty-

```
HBA_HUMAN GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
          GS+ + G +   +D L  ++ H+ D+  A +AL D    ++AH+
F11G11.2  GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Rysunek 1.1: Przykładowe przyrównanie dwóch sekwencji.

styczne. Więcem można znaleźć w [1].

Zwykle proces stwierdzenia czy sekwencje są ze sobą związane ewolucyjnie składa się z dwóch faz.

Pierwszą jest znalezienie przyrównania dwóch sekwencji. Jest to takie dopasowanie do siebie sekwencji, aby możliwie największa liczba aminokwasów odpowiadających sobie była taka sama (lub podobna). Przykładowe przyrównanie jest przedstawione na rysunku 1.1. Optymalne przyrównania znajduje się poprzez dodanie spacji wewnątrz sekwencji i zwykle służą do tego algorytmy oparte na programowaniu dynamicznym.

Drugą fazę stanowi ocena otrzymanego wcześniej przyrównania i zdecydowanie czy oznacza ono, że sekwencje są spokrewnione, czy też powstało zupełnie przypadkowo.

1.2 Tablice substytucyjne

Porównując sekwencje zwykle szukamy dowodu na to, że pochodzą one od tego samego przodka poprzez serie mutacji oraz selekcji. W sensie podstawowym pojęcie mutacji jest tożsame z podstawieniami, które zamieniają aminokwasy w sekwencji, oraz z insercjami i delecjami, które dodają bądź usuwają ciąg aminokwasów z sekwencji. Insercje i delecje są nazywane przerwami. Naturalna selekcja powoduje, że niektóre zmiany (mutacje) są możliwe do zaobserwowania częściej niż inne.

Taka charakterystyka zmian sekwencji doprowadziła do opracowania koncepcji oceniania podobieństwa sekwencji z użyciem *tablic substytucyjnych*. Wartość przyrównania w tej metodzie jest sumą wartości dla każdej uliniowionej pary aminokwasów plus wartości za przerwy. Tablice substytucyjne zawierają dla każdej pary aminokwasów wartość, która ocenia jak bardzo *"prawdopodobne"* jest wystąpienie danej mutacji.

Pozostaje jednak problem jak wypełnić tablicę substytucyjną. Powstało wiele metod wyliczenia wartości do tablic, opierających się na różnych założeniach i w związku z tym dających różne wyniki. Można o nich przeczytać np. w [2] i [3].

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
	2	-2	0	0	-3	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3	A
		12	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0	C
A	4		4	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-7	-4	D
C	0	9		4	-5	0	1	-2	0	-3	-2	1	-1	2	-1	0	0	-2	-7	-4	E
D	-2	-3	6		9	-5	-2	1	-5	2	0	-3	-5	-5	-4	-3	-3	-1	0	7	F
E	-1	-4	2	5		5	-2	-3	-2	-4	-3	0	0	-1	-3	1	0	-1	-7	-5	G
F	-2	-2	-3	-3	6		6	-2	0	-2	-2	2	0	3	2	-1	-1	-2	-3	0	H
G	0	-3	-1	-2	-3	6		5	-2	2	2	-2	-2	-2	-2	-1	0	4	-5	-1	I
H	-2	-3	-1	0	-1	-2	8		5	-3	0	1	-1	1	3	0	0	-2	-3	-4	K
I	-1	-1	-3	-3	0	-4	-3	4		6	4	-3	-3	-2	-3	-3	-2	2	-2	-1	L
K	-1	-3	-1	1	-3	-2	-1	-3	5		6	-2	-2	-1	0	-2	-1	2	-4	-2	M
L	-1	-1	-4	-3	0	-4	-3	2	-2	4		2	0	1	0	1	0	-2	-4	-2	N
M	-1	-1	-3	-2	0	-3	-2	1	-1	2	5		6	0	0	1	0	-1	-6	-5	P
N	-2	-3	1	0	-3	0	1	-3	0	-3	-2	6		4	1	-1	-1	-2	-5	-4	Q
P	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7		6	0	-1	-2	2	-4	R
Q	-1	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5		2	1	-1	-2	-3	S
R	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5		3	0	-5	-3	T
S	1	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4		4	-6	-2	V
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5		17	0	W
V	0	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4		10	Y
W	-3	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11		
Y	-2	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	7	
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	

Rysunek 1.2: Tablice substytucyjne: BLOSUM62 (dolna część) oraz PAM250 (górna część).

1.2.1 PAM

PAM, który jest skrótem od *Percent Accepted Mutations*, został wymyślony przez Margaret Dayhoff w 1970 roku. Przeprowadziła ona rozległe badania częstości substytucji aminokwasów między sobą podczas ewolucji. Badania obejmowały przyrównanie sekwencji aminokwasów tworzących białka należące do kilku rodzin białek oraz konstrukcję drzew filogenetycznych dla każdej rodziny. To umożliwiło stworzenie tablic PAM.

Tablice PAM są oparte na częstościach mutacji zaobserwowanych w rodzinach bardzo spokrewnionych białek (podobieństwo między sekwencjami wynosiło co najmniej 85%), więc większość mutacji stanowiły pojedyńcze substytucje. Wyliczona przez M. Dayhoff tablica PAM1 odpowiada jednemu procentowi pojedynczych mutacji, czyli jednej mutacji na każde 100 aminokwasów. Tablice PAM dla większych odległości ewolucyjnych otrzymywane są poprzez podniesienie tablicy PAM1 do odpowiedniej potęgi. I tak np. $PAM100 = (PAM1)^{100}$, co odpowiada 100 pojedynczym mutacjom na każde 100 aminokwasów, co daje około 50% identycznych pozycji (aminokwasy na konkretnych pozycjach mogą mutować wielokrotnie).

1.2.2 BLOSUM

BLOSUM jest metodą otrzymania tablic substytucyjnych wymyśloną przez Stevena i Jorje G. Henikoffów w 1992 roku mającą w założeniach poprawić miarę odległości między dwoma sekwencjami szczególnie dla sekwencji bardziej od siebie oddalonych ewolucyjnie. Tablice PAM obliczone przez M. Dayhoff oparte zostały na sekwencjach identycznych w conajmniej 85%, a w większości przypadków szukane są znacznie odleglejsze homologie.

Henikoffowie do obliczenia tablic BLOSUM użyli bazy danych bloków — BLOCKS (BLOSUM pochodzi od BLOcks SUbstitution Matrix). Bloki są to silnie konserwowane fragmenty sekwencji uliniowione ze sobą bez spacji.

Główną ideą jest zmierzenie stosunku obserwowanego prawdopodobieństwa wystąpienia pary aminokwasów względem prawdopodobieństwa oczekiwanego, które jest liczone z prawdopodobieństw wystąpień pojedynczych aminokwasów przy założeniu, że są to zdarzenia niezależne. Po zlogarytmowaniu tych ilorazów otrzymujemy tablicę wynikową zwaną tablicą logarytmów ilorazów (ang. *logarithm of odds matrix*).

Główna różnica pomiędzy tablicami PAM i BLOSUM polega na innym sposobie zróżnicowania dystansu ewolucyjnego. W tablicach BLOSUM jest on osiągany poprzez klasteryzację. Polega ona na sklejaniu ze sobą sekwencji, które są do siebie podobne w conajmniej pewnym stopniu. Zmiana liczby procent, przy której zachodzi klasteryzacja prowadzi do innej tablicy wynikowej. Tak więc otrzymujemy serię tablic BLOSUM, np. BLOSUM62 jest wyliczony z bloków sekwencji sklejanych gdy są w conajmniej 62% identyczne. Mniejszy procent podobieństwa powoduje wyliczenie tablicy rozpoznającej lepiej odleglejsze homologie.

Rozdział 2

Przyrównanie kontekstowe

Stosowane dotychczas metodologie wykrywanie homologii pomiędzy sekwencjami opierają się mocno na jednym założeniu:

> Prawdopodobieństwo zajścia mutacji aminokwasu A w aminokwas B
> na danej pozycji zależy tylko od tego czym są A i B.

Jest to uproszczenie, które jest w oczywisty sposób nieprawdziwe, ale zostało poczynione z następujących powodów:

- moc obliczeniowa komputerów nie była wystarczająca, aby zajmować się czymś więcej niż jedną pozycją w sekwencji jednocześnie.

- algorytmowi biorącemu pod uwagę oprócz informacji o parze aminokwasów mutujących z jednego w drugi inne dane, siłą rzeczy nie wystarczają dane znajdujące się w tablicy substytucyjnej. Wymaga on do działania dodatkowych parametrów, które podobnie jak w przypadku tablic substytucyjnych muszą być otrzymane metodami statystycznymi ze znanych zsekwencjonowanych białek. Liczba zsekwencjonowanych organizmów, była za mała, aby można było próbować wyliczać z posiadanych danych nowe informacje, ale rośnie z roku na rok coraz szybciej i można już próbować stosować nowe metodologie.

Można rozważać wiele płaszczyzn w jakich kontekst w jakim znajduje się aminokwas ma wpływ na prawdopodobieństwo zajścia mutacji. Kontekst może być rozumiany w dwojaki sposób:

- otoczenie aminokwasu w przestrzeni — jest to bardzo intuicyjne podejście, ale niestety silnie związane z problemem zwijania się białek, który nie jest jak dotąd rozwiązany, więc nie jest przedmiotem niniejszej pracy

- otoczenie aminokwasu w sekwencji — jest to jakby uboższa wersja poprzedniego podejścia. Aminokwasy znajdujące się w sekwencji w bezpośrednim otoczeniu obserwowanego aminokwasu są na pewno także blisko w przestrzeni, ale nie muszą być one wszystkimi takimi aminokwasami.

Innym aspektem, który można rozważać zwracając uwagę na kontekst w jakim wystąpił dany aminokwas i wpływ kontekstu na jego mutowalność są cechy kontekstu. Zostało napisanych wiele prac opisujących różnorodne cechy fizyczne i chemiczne aminokwasów, które mogą mieć wpływ na procesy zachodzące w ich otoczeniu ([4], [5]). Przykładami cech fizycznych mogą być masa, wielkość czy kąt pomiędzy przyłączonymi aminokwasami, a przykładami cech chemicznych mogą być polarność, hydrofobowość, ładunek czy aromat.

Więcej motywacji przemawiających za rozpatrywaniem przyrównania kontekstowego można znaleźć w pracach [6] i [7].

Rozdział 3

Tablice do przyrównania kontekstowego

W pracy [8] zaproponowano zestaw algorytmów służących do przyrównywania kontekstowego. Przyjęto założenie, że kontekstem będą dwa aminokwasy sąsiadujące z rozważanym. Algorytmy przyrównania kontekstowego wymagają do pracy zestawu tablic substytucyjnych (jednej dla każdego możliwego kontekstu), które charakteryzowałyby możliwość zajścia mutacji w danym kontekście. W dalszej części niniejszej pracy przedstawiony jest algorytm obliczający tablice, które mogą być wykorzystane przez algorytmy z [8].

3.1 Zarys metody

Idea algorytmu oparta jest na pomyśle Stevena i Jorii G. Henikoffów wykorzystanym przy tworzeniu tablic substytucyjnych BLOSUM ([9]). Tablice substytucyjne wyliczane są z bazy bloków utworzonych z uliniowionych (bez spacji) fragmentów sekwencji aminokwasów zakonserwowanych podczas ewolucji.

Podobnie jak w przypadku tablic BLOSUM wartości w tablicach są logarytmami ilorazów obserwowanej częstości wystąpień zamiany aminokwasu w inny do oczekiwanej częstości zajścia takiej zamiany. Wprowadzone zostały dwie duże zmiany.

Pierwszą jest to, że wszystkie częstości są zliczane oddzielnie dla każdego kontekstu. Jest to oczywista zmiana jeśli się weźmie pod uwagę to, że usiłujemy znaleźć wartości charakteryzujące wpływ kontekstu na zachodzące mutacje. Wymusiła ona zmiany w sposobie tworzenia klastrów, gdyż zachowanie metody używanej przez Henikoffów spowodowałoby powstanie niejednoznacznych kontekstów (tj. takich w których kontekstem nie byłaby para aminokwasów, a para ‘profili’ powstałych ze sklejenia aminokwasów z klastra). Sklejanie sekwencji w ramach klastra zachodzi dla każdej kolumny oddzielnie w taki sposób by łączyć tylko sekwencje mające taki sam kontekst.

Drugą zmianą jest rezygnacja z założenia, że tablice są symetryczne, czyli że substytucja aminokwasu A w B jest tak samo prawdopodobna jak aminokwasu B w A. Asymetria została osiągnięta poprzez założenie, że na danej pozycji jeden aminokwas był pierwotny i z niego zachodziła zamiana na inne aminokwasy. Przyjęto, że pierwotnym aminokwasem jest ten, który ma w kolumnie największą wartość względnej częstości występowania (jest to iloraz częstości występowania aminokwasu w kolumnie przez częstość występowania w całej bazie bloków).

3.2 Szczegółowy opis

Aby móc precyzyjnie opisać sposób obliczania tablic potrzebnych jest kilka definicji. Niech:

$$\mathbf{A} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$$

będzie zbiorem aminokwasów i niech $\mathbf{C}$ będzie zbiorem kontekstów definiowanym przez funkcję:

$$f : \mathbf{A} \to \mathbf{C}$$

Oznaczmy zbiór wszystkich bloków przez $\mathcal{B}$, blok przez B oraz sekwencje go tworzące przez $s_1, \dots, s_j$ gdzie j jest liczbą sekwencji tworzących blok.

Niech $s(k)$ oznacza k-ty aminokwas sekwencji s.

Przypomnijmy konwencję Iversona

$$[\varphi] = \begin{cases} 1 & \mathrm{wttw, gdy}\ \varphi \\ 0 & \mathrm{wttw, gdy}\ \neg\varphi \end{cases}$$

Przez $p \in [0, 1]$ oznaczmy stałą będącą współczynnikiem podobieństwa sekwencji. Wtedy powiemy, że:

$$s_x \sim s_y \text{ (jest podobna) } \Leftrightarrow \sum_{i \in [1 \dots n]} [s_x(i) = s_y(i)] \geq pn$$

gdzie n jest długością sekwencji.

Domknięcie przechodnie relacji $\sim$ oznaczymy $\approx$ i nazwiemy *relacją klastrującą*. W oczywisty sposób widać, że relacja $\approx$ jest relacją równoważności.

Zdefiniujmy następującą rodzinę relacji $\{\bowtie_{\bullet i}\}_{i=2,\dots,n-1}$ zdefiniowanych dla konkretnego bloku. Dwie sekwencje w bloku są w relacji $\bowtie_{\bullet i}$ jeśli mają równe konteksty w otoczeniu aminokwasu i:

$$\bowtie_{\bullet i} = \{(s_x, s_y) : f(s_x(i-1)) = f(s_y(i-1)) \wedge f(s_x(i+1)) = f(s_y(i+1))\}$$

Każda z relacji $\bowtie_{\bullet i}$ jest relacją równoważności.

Przyjrzyjmy się teraz następującej rodzinie relacji:

$$\bowtie_i = \bowtie_{\bullet i} \cap \approx$$

gdzie $i \in [2 \dots n-1]$. Klasy abstrakcji tych relacji odpowiadają sekwencjom znajdującym się w jednym klastrze oraz mającym taki sam kontekst w otoczeniu pozycji i.

Dla klasy abstrakcji S relacji $\bowtie_i$ definiujemy

$$s(i,b) = \frac{\sum_{s_x \in S}[s_x(i) = b]}{|S|}$$

Wtedy oznaczmy:

$$\begin{array}{rcl} s^{\bowtie_i}(i-1) & := & s_x(i-1) \\ s^{\bowtie_i}(i,b) & := & s(i,b) \\ s^{\bowtie_i}(i+1) & := & s_x(i+1) \end{array}$$

oraz niech $B^{\bowtie_i}$ będzie zbiorem wszystkich $s^{\bowtie_i}$ zdefiniowanych dla $S \in [B/\bowtie_i]$.

Niech

$$B^{\bowtie_i}(a,c) = \{s_j^{\bowtie_i} \in B^{\bowtie_i} : s_j^{\bowtie_i}(i-1) = a \wedge s_j^{\bowtie_i}(i+1) = c\}$$

oraz

$$\Psi(B,a,c) = \{B^{\bowtie_i}(a,c) : i = 2,\dots,n-1\}$$

gdzie n jest szerokością bloku B.

3.3 Częstości wystąpień aminokwasów w kontekstach

Oznaczmy przez

$$F_p(a,b,c) : \mathbf{C} \times \mathbf{A} \times \mathbf{C} \to [\mathbf{0},\mathbf{1}]$$

funkcję częstości wystąpień aminokwasów w kontekstach. Wyraża się ona następującym wzorem:

$$F_p(a,b,c) = \frac{\sum_{B \in \mathcal{B}} \sum_{B^{\bowtie_i} \in \Psi(B,a,c)} \sum_{s^{\bowtie_i} \in B^{\bowtie_i}} s^{\bowtie_i}(i,b)}{\sum_{B \in \mathcal{B}} \sum_{B^{\bowtie_i} \in \Psi(B,a,c)} \sum_{s^{\bowtie_i} \in B^{\bowtie_i}} \sum_{x \in \mathbf{A}} s^{\bowtie_i}(i,x)}$$

który jest równoważny wzorowi:

$$F_p(a,b,c) = \frac{\sum_{B \in \mathcal{B}} \sum_{B^{\bowtie_i} \in \Psi(B,a,c)} \sum_{s^{\bowtie_i} \in B^{\bowtie_i}} s^{\bowtie_i}(i,b)}{\sum_{B \in \mathcal{B}} \sum_{B^{\bowtie_i} \in \Psi(B,a,c)} \sum_{s^{\bowtie_i} \in B^{\bowtie_i}} 1}$$

3.4 Obserwowane częstości wystąpień substytucji

Niech $f_p(B^{\bowtie_i}(a,c),b)$ będzie częstością występowania aminokwasu b w zbiorze $B^{\bowtie_i}(a,c)$. Wyraża się ona następującym wzorem:

$$f_p(B^{\bowtie_i}(a,c),b) = \frac{\sum_{s^{\bowtie_i} \in B^{\bowtie_i}(a,c)} s^{\bowtie_i}(i,b)}{|B^{\bowtie_i}(a,c)|}$$

Wtedy niech

$$\overline{b}(B^{\bowtie_i}(a,c))$$

oznacza taki aminokwas b, dla którego wyrażenie

$$f_p(B^{\bowtie_i}(a,c),b)/F_p(a,b,c)$$

będzie miało maksymalną wartość.

Niech m będzie minimalną mocą zbioru $B^{\bowtie_i}(a,c)$ dla której warto rozpatrywać substytucje w nim zachodzące. Wtedy niech

$$\Psi'(B,a,c,b) = \{B^{\bowtie_i} \in \Psi(B,a,c) : \overline{b}(B^{\bowtie_i}) = b \wedge |B^{\bowtie_i} \geq m\}$$

Wtedy przez:

$$M_o : \mathbf{C} \times \mathbf{A} \times \mathbf{C} \times \mathbf{A} \to [\mathbf{0},\mathbf{1}]$$

oznaczymy funkcję częstości substytucji jednego aminokwasu w drugi w ustalonym kontekście. Przy założeniu niezależności zdarzeń funkcja M_o wyraża się następującym wzorem:

$$M_o(a,b,c,b') = \frac{\sum_{B \in \mathcal{B}} \sum_{B^{\bowtie_i} \in \Psi'(B,a,c,b)} \sum_{s^{\bowtie_i} \in B^{\bowtie_i}} s^{\bowtie_i}(i,b')}{\sum_{B \in \mathcal{B}} \sum_{B^{\bowtie_i} \in \Psi'(B,a,c,b)} \sum_{s^{\bowtie_i} \in B^{\bowtie_i}} \sum_{x \in \mathbf{A}} s^{\bowtie_i}(i,x)}$$

3.5 Oczekiwane częstości wystąpień substytucji

Przez:

$$M_e : \mathbf{C} \times \mathbf{A} \times \mathbf{C} \times \mathbf{A} \to [\mathbf{0},\mathbf{1}]$$

oznaczmy funkcję oczekiwanej częstości substytucji jednego aminokwasu w drugi w ustalonym kontekście. Przy założeniu niezależności zdarzeń wystąpień aminokwasów funkcja M_e wyraża się następującym wzorem:

$$M_e(a,b,c,b') = F_p(a,b,c) \cdot F_p(a,b',c)$$

3.6 Wartości w tablicach substytucyjnych

Ostateczne wartości w tablicach substytucyjnych podobnie jak w pracach opisujących tworzenie tablic PAM oraz BLOSUM ([10] i [9]) wyrażają się wzorem:

$$M(a, b, c, b') = \log_2 \frac{M_o(a, b, c, b')}{M_e(a, b, c, b')}$$

czyli są logarytmem ilorazu częstości obserwowanej do częstości oczekiwanej.

Rozdział 4

Program

Program został napisany w języku C++. Do poprawnej pracy potrzebna jest baza bloków, z których obliczane wartości w tablicach substytucyjnych. Po uruchomieniu wypisuje wszystkie ustawienia i po odczekaniu kilku sekund rozpoczyna obliczenia.

4.1 Parametry

Program przyjmuje następujące parametry:

-s -similarity-for-clustering <float> — ustawia współczynnik podobieństwa sekwencji używany przy tworzeniu klastrów w bloku;

-a -expected-average <float> — powoduje zwiększenie lub zmniejszenie otrzymanych wartości w tablicach, aby średnia była równa tej zadanej parametrem;

-v -expected-variance <float> — powoduje przeskalowanie otrzymanych wartości w tablicach, aby wariancja była równa tej zadanej parametrem;

-m -min-S0-size <int> — ustawia minimalną moc zbioru sekwencji z takim samym kontekstem, dla której dane z tego zbioru będą zliczane;

-d -input-files-dir <string> — ustawia katalog, w którym poszukiwane będą pliki z definicjami bloków (domyślną wartością jest '.');

-e -input-files-ext <string> — ustawia rozszerzenie dla plików zawierających pliki z definicjami bloków (domyślną wartością jest '.block');

```
Ah
Cp
Dp
...
```

Rysunek 4.1: Fragment przykładowego pliku konwersji.

`-c -conversion-file <string>` — wskazuje plik zawierający definicję funkcji konwertującej aminokwasy na konteksty (domyślnie każdy aminokwas jest konwertowany na kontekst równy małej literze odpowiadającej aminokwasowi);

`-o -output-file <string>` — wskazuje plik, do którego mają być zapisane wynikowe tablice w formacie akceptowanym przez program R. Szklarczyka;

`-t -tab-output-file <string>` — wskazuje plik, do którego mają być zapisane wynikowe tablice w formacie czytelnym dla człowieka;

`-h -help` — powoduje wypisanie listy akceptowanych parametrów i kończy działanie.

4.2 Opis plików wejściowych

Program czyta dwa rodzaje plików. Pierwszym z nich jest plik z definicją funkcji konwertującej aminokwasy na konteksty. Drugim rodzajem plików wejściowych są definicje bloków.

4.2.1 Plik konwersji aminokwasów na konteksty

Format tego pliku jest bardzo prosty. W każdej linii znajduje się definicja kontekstu dla jednego aminokwasu. Pierwszym znakiem w linii ma być kod aminokwasu, a drugim kod odpowiedniego kontekstu. Reszta linii jest ignorowana.

Kody aminokwasów są dużymi literami ('ACDEFGHIKLMNPQRSTVWY'). Kody kontekstów mają być małymi literami.

Fragment przykładowego pliku konwersji znajduje się na rysunku 4.1.

4.2.2 Pliki bloków

Każde kolejne dwie linie pliku definiującego blok zawierają definicję jednej sekwencji wchodzącej w skład bloku. Pierwsza linia zawiera nazwę, a druga ciąg

```
PA5523
CPLFVARAHGARFHDVDGHAYVDFCLGDTGSMFGHSPAP
MJ0603
YPFFVEKAKDCYLFDVDGNCYIDYCLAYGPMVLGHANDA
HP0306
TPPFILKGKGAYLYDVDNNHYIDFVQSWGPLIFGHADEE
PA4088
CPLLVKEAQGARVTDIDGQQYVDFALGDSGAMFGHAQPA
MTH228
HPFFAAGGSGCLLESVDGESYIDYCLAYGPLILGHAHPR
DR0555
VPRFIASAQGAYLTDADGARYLDYIGSWGPMILGHNHPA
```

Rysunek 4.2: Przykładowy plik definiujący blok.

aminokwasów tworzących sekwencję. Wszystkie sekwencje w danym bloku muszą być równej długości i składać się wyłącznie z kodów aminokwasów.

Przykładowy plik definiujący blok znajduje się na rysunku 4.2.

4.3 Opis plików wynikowych

Program generuje dwa pliki wynikowe zawierające te same wartości, ale prezentowane w różnej formie. Oba są plikami tekstowymi, ale format pierwszego jest zgodny z przyjmowanym przez program R. Szklarczyka, a format drugiego jest bardziej przejrzysty dla człowieka.

4.3.1 Plik dla programu przyrównującego

Format jest bardzo prosty. Każda linia zawiera jedną wartość tabeli. Pierwszym znakiem w linii jest kod lewego kontekstu, drugim źródłowy aminokwas, trzecim kod prawego kontekstu i czwartym docelowy aminokwas. Potem następuje wartość jaka jest związana z tą pozycją w tablicy.

Fragment przykładowego pliku wynikowego w formacie dla programu przyrównującego znajduje się na rysunku 4.3.

4.3.2 Plik czytelny dla człowieka

Wartości są w nim wypisane w postaci serii tabel — po jednej dla każdego możliwego kontekstu. Każda tabela zawiera $20 * 20$ wartości i jest po prostu tablicą substytucyjną dla konkretnego kontekstu.

```
nAnA2.438438
nAnC-1.295498
nAnD-2.790822
nAnE-2.288589
nAnF-2.601891
nAnG-0.948917
nAnH-2.844210
nAnI-1.882654
...
```

Rysunek 4.3: Fragment przykładowego pliku wynikowego w formacie dla programu przyrównującego.

```
Array for context: nn
b\b2    A       C       D       E       F       G
A       2.43844 -1.2955 -2.7908 -2.2885 -2.6018 -0.9489
C       1.11637 5.15998 -1.2539 -1.3183 -0.3663 -0.3512
D       -1.2573 -2.8139 3.52258 0.56958 -2.7812 -1.2827
E       -0.5233 -2.7867 0.69198 3.23031 -2.5054 -1.6476
F       -1.3772 -1.5176 -2.6869 -2.5378 3.62105 -2.2947 ...
G       -0.8129 -2.4286 -1.5879 -2.0756 -3.0287 3.15731
H       -0.3215 -1.1494 0.00020 0.02861 -0.2278 -0.8189
I       -1.5427 -1.9075 -3.3922 -2.9632 -0.7938 -3.1445
K       -0.8037 -2.5744 -1.0053 -0.0570 -2.3489 -1.6730
                            ...
```

Rysunek 4.4: Wycinek przykładowego pliku wynikowego w formacie przyjaznym dla człowieka.

Wycinek przykładowego pliku wynikowego w formacie przyjaznym dla człowieka znajduje się na rysunku 4.4.

4.4 Złożoność

Wykonanie programu podzielone jest na dwie fazy:

- policzenie częstości wystąpień aminokwasów w kontekstach w całej bazie bloków;
- policzenie tablic substytucyjnych poprzedzone zliczeniem liczb substytucji z jednego aminokwasu w inny w określonym kontekście.

W każdej z tych faz przeglądane są wszystkie bloki i wykonywane są na nich pewne operacje. Przyjmijmy, że m będzie liczbą sekwencji w bloku, a n długością sekwencji (czyli szerokością bloku). Wtedy w pierwszej fazie algorytmu koszt dla jednego bloku jest równy $O(m^2n + mn)$. Pierwszy składnik sumy pochodzi od procesu przydzielania sekwencji do klastrów (trzeba policzyć podobieństwo każdej pary sekwencji, a policzenie podobieństwa wykonywane jest w koszcie liniowym ze względu na długość sekwencji). Drugi składnik sumy pochodzi zliczania częstości wystąpień aminokwasów.

W drugiej fazie koszt także wynosi $O(m^2n + mn)$ i pierwszy składnik sumy także pochodzi od procesu przydzielania sekwencji do klastrów. Pozostała część kosztu pochodzi od zliczania zamian aminokwasów.

Całkowity koszt wykonania obliczeń dla jednego bloku wynosi $O(m^2n)$, a co za tym idzie cały program wykonuje się w czasie

$$O(km_{\max}^2 n_{\max})$$

gdzie k jest liczbą bloków, a $m_{\max}$ i $n_{\max}$ są maksymalną wysokością i szerokością spośród wszystkich bloków.

Rozdział 5

Otrzymane tablice

Za pomocą opisanego wcześniej programu zostało wygenerowanych kilkadziesiąt różnych zestawów tablic substytucyjnych. Każdy zestaw tablic powstał poprzez uruchomienie programu z innymi parametrami, dzięki czemu można wpływ poszczególnych parametrów na osiągane wyniki.

Spośród wszystkich parametrów wejściowych najszerszego omówienia wymaga pochodzenie źródłowej bazy bloków oraz ustalenie funkcji definiującej kontekst.

5.1 Bazy bloków

- BLOCKS Database ([11]) — jest bazą bloków generowaną co pewien czas i publicznie dostępną do wykorzystania. Na użytek tej pracy oparłem się na bazie BLOCKS+ datowanej na 15 listopada 2000 roku. Jest to rozszerzona baza BLOCKS 12.0. Ogółem zawiera 11117 bloków z 2433 grup silnie spokrewnionych sekwencji[1].

- COGs Database ([12]) — jest bazą grup spokrewnionych ze sobą sekwencji. Nazwa takiej grupy — COG — pochodzi od angielskiego terminu *Clusters of Orthologous Groups*. W tej pracy wykorzystana została baza oparta na 34 kompletnych genomach organizmów reprezentujących 26 głównych gałęzi filogenetycznych[2].

Każdy blok w bazie BLOCKS podzielony jest na podbloki zawierające sekwencje bardzo do siebie podobne. Bloki z tej bazy zostały użyte do wygenero-

[1] obecnie dostępna jest nowsza wersja BLOCKS+ z dnia 1 sierpnia 2001 roku oparta na bazie BLOCKS 13.0

[2] dostępna jest nowa wersja oparta na 44 genomach organizmów reprezentujących 30 gałęzi filogenetycznych

wania tablic na dwa sposoby. Pierwszym było użycie bloków takich jakie są zapisane w bazie BLOCKS. W drugim przypadku jako bloki potraktowane zostały wspomniane wcześniej podbloki złożone z sekwencji blisko ze sobą spokrewnionych, co jest bardziej podobne do metodologii zastosowanej przy tworzeniu tablic PAM (brano tylko grupy sekwencji podobne do siebie w conajmniej 85%). W dalszej części pracy bloki otrzymane na dwa powyższe sposoby są oznaczane odpowiednio jako BLOCKS i BLOCKS2.

Baza COG nie zawiera gotowych bloków, które można by było użyć do wygenerowania tablic, zawiera natomiast przyrównania sekwencji wykonane algorytmem ClustalW (są to przyrównania ze spacjami). Bloki zostały wyekstrahowane z przyrównań poprzez wycinanie jak największych prostokątów (bloków) nie zawierających spacji. Otrzymano w ten sposób 35272 bloki z 3415 uliniowionych grup sekwencji.

5.2 Funkcje definiujące kontekst

Funkcja definiująca kontekst może być ustalona w dowolny sposób, ale najrozsądniejszym wydaje się oparcie się o pewne właściwości aminokwasów. W tej pracy skupiłem się na trzech różnych funkcjach definiujących kontekst:

- brak kontekstu — w porównaniu do istniejących tablic wprowadza tylko brak symetryczności tablicy, ale nadaje się dobrze do pewnych obserwacji dokonywanych bezpośrednio na wartościach w tablicach;
- kontekst hp — kontekst definiujący polarność i hydrofobowość aminokwasów;
- pełny kontekst — każdy aminokwas definiuje odpowiadający sobie kontekst.

Innymi możliwymi kontekstami są wspomniane w rozdziale 2 cechy fizyczne i chemiczne aminokwasów jak masa, wielkość, polarność, ładunek czy aromat.

Rozdział 6

Obserwacje

Badanie właściwości otrzymanych tablic i ich wpływu na otrzymywane przyrównania nie jest tematem tej pracy, niemniej jednak poczyniłem pewne obserwacje. Opracowanie metody wiarygodnego przetestowania otrzymanych wyników może być tematem oddzielnej pracy.

6.1 Porównanie wygenerowanej programem tablicy z tablicą BLOSUM62

Tablice wygenerowane przez program są dużym zbiorem liczb, których poprawność bardzo trudno jest oceniać człowiekowi. Są to dane statystyczne i nie istnieje definicja ich poprawności. Nasuwa się więc od razu pytanie jak sprawdzić czy wartości w wygenerowanej tablicy "mają jakiś sens". Dosyć prostym sposobem jest porównanie jednej z istniejących i powszechnie stosowanych tablic substytucyjnych z wygenerowaną przez program tablicą, w której kontekst jest zawsze taki sam (zbiór wartości funkcji definiującej kontekst ma 1 element).

Wynik takiego porównania pokazany jest na rysunku 6.1. Zawiera on tablicę różnic między wartościami z tablicy BLOSUM62 i tablicą wygenerowaną przez program przy następujących parametrach: stała klastrująca – 0.8, minimalna wielkość zbioru – 3, baza bloków – BLOCKS. Wartości otrzymanej tablicy zostały przeskalowane i przesunięte, aby otrzymać takie same wartości średniej i wariancji jakie ma tablica tablica BLOSUM62 (średnia – -1.06, wariancja – 4.7814).

Różnice nie okazały się być bardzo znaczące (wariancja jest równa 1.76). Jest tylko 18 wartości w tablicy, które są większe od 2 lub mniejsze od -2. Co więcej 13 spośród nich dotyczy wartości dla mutacji tryptofanu (W) w inne (takie różnice nie występują dla mutacji w tryptofan).

Małe wartości w tablicy różnic pozwalają sądzić, że algorytm wyliczający tablice nie zawiera w sobie jakichś znaczących błędów koncepcyjnych. Niepokój

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
A	0	1	2	2	1	1	2	1	2	1	1	1	2	2	2	1	1	1	2	2	A
C	-2	2	-2	-2	-2	**-3**	-2	-1	-2	-1	-1	**-3**	-2	-2	-2	-2	-2	-2	0	-1	C
D	-1	1	1	1	1	0	0	0	0	-1	0	0	1	1	0	0	0	0	0	0	D
E	-1	0	1	0	0	0	1	-1	1	-1	0	1	1	1	1	0	0	0	1	0	E
F	0	0	0	0	1	0	1	0	0	-1	0	0	-1	0	0	0	0	-1	2	2	F
G	1	0	1	1	1	1	1	0	0	0	0	1	1	0	1	1	1	1	**3**	1	G
H	-2	-2	-1	0	-1	-1	1	-2	-2	**-3**	-1	0	-1	-1	-1	-2	-2	-2	-1	1	H
I	1	1	1	1	1	0	1	0	0	1	1	1	1	0	0	1	1	1	1	1	I
K	0	0	0	1	0	0	0	-1	0	0	1	0	1	0	0	0	0	0	1	0	K
L	2	2	1	2	1	1	1	2	2	1	2	1	1	2	2	2	2	2	2	2	L
M	-2	-1	-1	-2	-1	-2	-1	-1	-1	-1	-2	-1	-1	-1	-1	-1	-2	-1	0	-1	M
N	-1	-1	0	0	-1	0	1	-1	0	-1	0	1	0	0	1	0	0	-1	0	0	N
P	-1	-1	-1	-1	-2	-1	-1	-2	-1	-2	0	-1	1	-1	-1	-1	-1	-1	-1	-1	P
Q	-1	-1	0	0	-1	-1	0	-2	0	-1	0	-1	0	-1	0	0	-1	-1	1	0	Q
R	0	0	0	1	0	0	1	-1	1	0	1	1	0	1	0	0	0	-1	0	0	R
S	1	1	2	2	1	1	2	1	2	1	2	2	1	2	1	0	1	1	2	1	S
T	1	1	1	1	1	0	0	0	0	1	1	1	1	0	1	0	1	0	**3**	1	T
V	1	1	2	1	1	1	1	2	1	1	2	1	2	2	1	1	1	0	2	2	V
W	**-4**	-2	**-4**	**-3**	-2	-2	**-3**	**-4**	**-4**	**-4**	-2	**-4**	**-4**	**-3**	**-4**	**-4**	-2	**-4**	2	-1	W
Y	-1	-1	-1	0	1	-1	1	0	-1	-1	0	-1	-1	0	-1	-1	-1	0	2	1	Y
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	

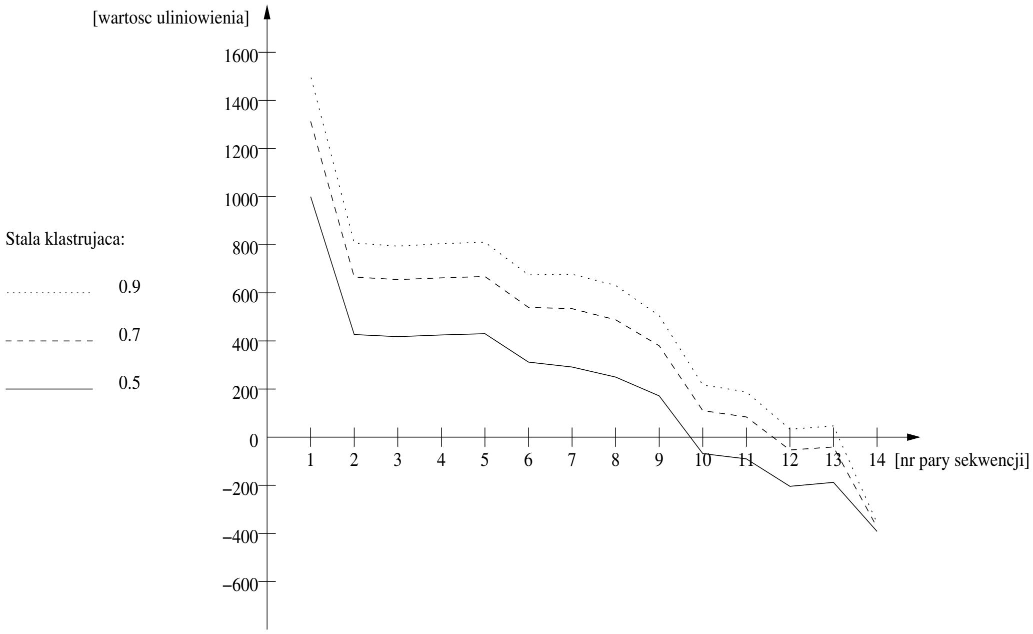
Rysunek 6.1: Tablica różnic wartości z tablic BLOSUM62 i tablicy wyliczonej z bloków bazy BLOCKS przy stałej klastrującej 0.8 i braku kontekstu. Przed odjęciem wyliczona tablica została przeskalowana i przesunięta aby miała taką samą wartość średnią i wariancję.

może wzbudzać jedynie anomalia dla wspomnianego wcześniej tryptofanu. Jeżeli wybór wspólnego przodka taką metodą jak w algorytmie jest dobry to wyniki w tablicy mogą oznaczać wykrycie nowej właściwości tego aminokwasu, ale może też się okazać, że poczynione założenia nie są prawdziwe i wprowadzone zostały do tablicy sztuczne zaburzenia.

6.2 Wpływ zmiany parametrów programu na otrzymane tablice i obliczone przyrównania

Jak opisałem w rozdziale 4.1 działanie programu jest zdeterminowane kilkoma parametrami. Naturalnym jest pytanie jaki wpływ mają one na otrzymane tablice oraz na przyrównania wykonywane z użyciem tych tablic. Ze względu na dużą liczbę kombinacji parametrów oraz dosyć długi czas wykonania programu przyrównującego (sześcienna złożoność czasowa względem długości przyrównywanych sekwencji) nie jest możliwe obliczenie przyrównań z wykorzystaniem wszystkich możliwych tablic. Można natomiast spróbować zaobserwować wpływ poszczególnych parametrów na otrzymane wartości przyrównań.

Wykresy 6.2, 6.3 oraz 6.4 pokazują zmiany wartości otrzymanych przyrównań dla czternastu różnych par sekwencji. Pierwsze trzynaście par stanowią sekwencje spokrewnione ze sobą (należą do jednej grupy COG). Dla kontrastu czternastą parę stanowi sekwencja z tej samej grupy COG wraz z sekwencją o podobnej długości, ale wygenerowaną losowo. Można dzięki temu zauważyć, że rzeczywiście algorytm przyrównujący znajduje homologie sekwencji w przypadku poprzednich



Rysunek 6.2: Wykres obrazujący wpływ zmiany wartości stałej klasteryzacji na otrzymane wartości przyrównań.

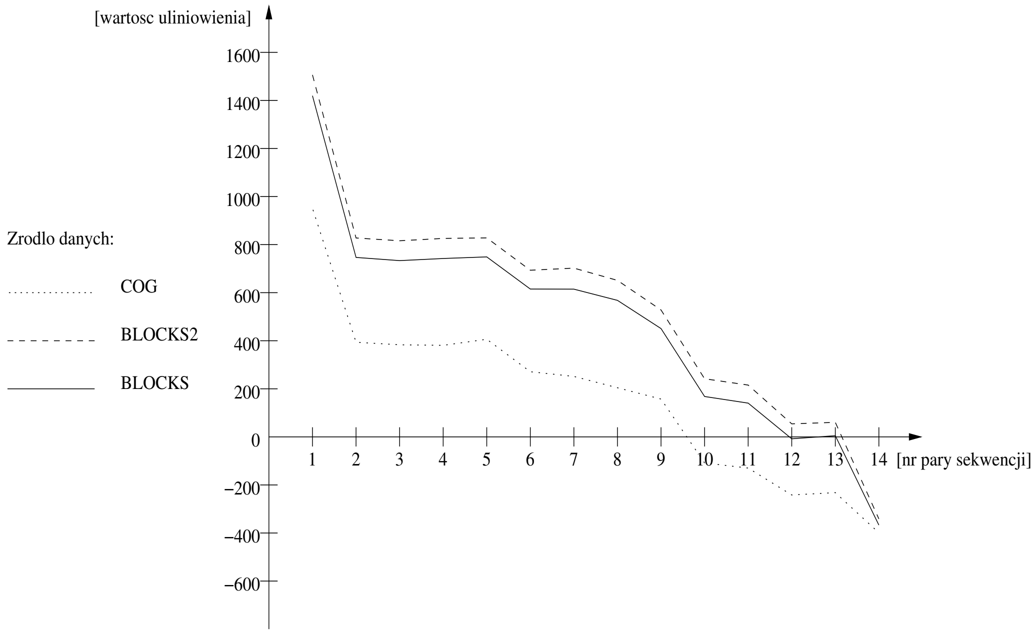
par.

Wszystkie tablice, na których wykonane zostały przyrównania zostały przekształcone tak, by miały taką samą wartość średnią i wariancję.

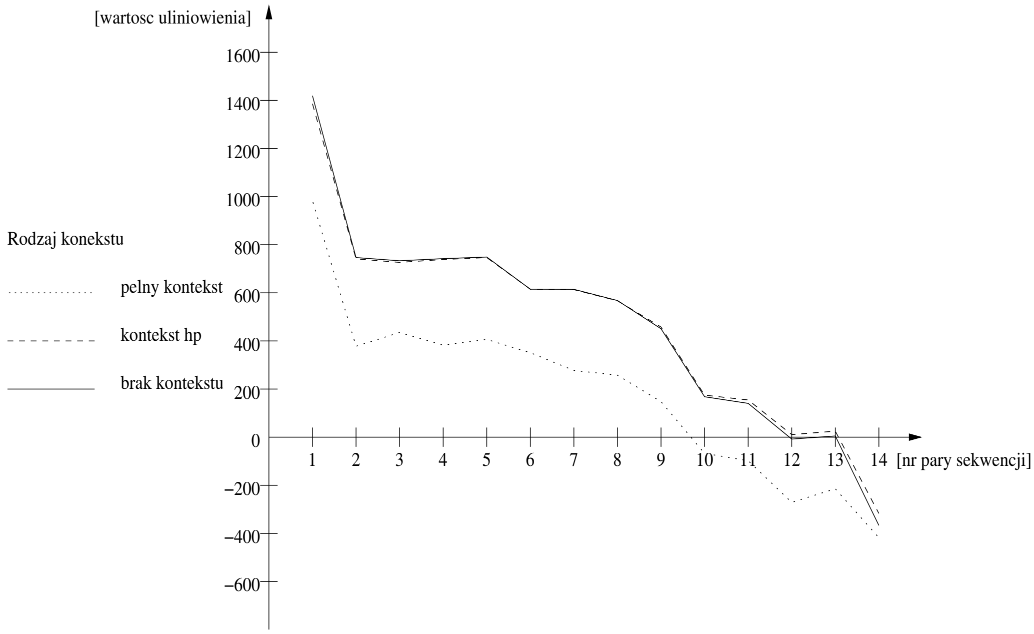
Na wykresie 6.2 zaobserwować możemy wpływ wartości stałej klastrującej na otrzymywane wartości przyrównań. W przeciwieństwie do oczekiwań — niezgodnie z intuicją pochodzącą od metodologii wykorzystanej w tablicach BLOSUM — zmniejszenie stałej klastrującej nie powoduje uzyskiwania lepszych wyników dla par sekwencji mniej ze sobą. Najlepsze wyniki osiągnięte zostały dla największej wartości stałej klastrującej, co daje podstawy by sądzić, że pomysł tworzenia klastrów jest chybiony.

Niewiele jestem w stanie wywnioskować z wykresu 6.3, który przedstawia zależność wartości przyrównań od źródłowej bazy bloków. W zasadzie jedyną obserwacją jest prawidłowość, że dla tablic wygenerowanych z bazy BLOCKS2 otrzymaliśmy lepszy wynik od BLOCKS tylko dla najbardziej spokrewnionych ze sobą sekwencji. To odzwierciedla różnice pomiędzy bazami bloków BLOCKS i BLOCKS2 (BLOCKS2 zawiera bloki z sekwencji bardzo mocno ze sobą spokrewnionych).

Wnioski narzucające się po krótkim spojrzeniu na wykres zależności od przyjętego kontekstu (wykres 6.4) nie są obiecujące. Praktyczny brak różnic pomiędzy wynikami osiągniętymi z kontekstem pustym i z kontekstem ‘hp’ pozwala



Rysunek 6.3: Wykres obrazujący wpływ zmiany źródła danych bloków na otrzymane wartości przyrównań.



Rysunek 6.4: Wykres obrazujący wpływ kontekstu na otrzymane wartości przyrównań.

wnioskować jedną z dwóch możliwości. Pierwszą jest to, że hydrofobowość i polarność aminokwasów nie ma dużego wpływu na zachodzące na sąsiednich pozycjach mutacje i należy próbować znaleźć inny podział (być może na więcej niż dwie grupy). Druga możliwość jest bardziej pesymistyczna. Może się okazać, że rozpatrywanie kontekstu złożonego z pary sąsiednich aminokwasów nie wystarczy i wtedy trzeba byłoby zmienić metodologię.

Na pierwszy rzut oka niepokoją też bardzo znacznie gorsze wyniki otrzymane z pełnym kontekstem. Tu jednak stosunkowo łatwo znaleźć wytłumaczenie dla takiego stanu rzeczy. Posiadane bazy bloków są za małe by otrzymać wszystkie 160000 wartości z tablic dla pełnego kontekstu ($20 \cdot 20$ tablic po $20 \cdot 20$ wartości). Otrzymane tablice zawierają 114581 wartości co oznacza, że 45419 przypadków substytucji aminokwasu w inny w określonym kontekście nie miało miejsca. Prostym wnioskiem z tego jest także to, że statystyczna jakość pozostałych wartości pozostawia wiele do życzenia. Dodatkowo algorytm przyrównujący wykorzystuje w przyrównaniach tylko takie substytucje, dla których istnieją wartości w tablicach. Brak wszystkich wartości powoduje niemożność wykonania pewnych substytucji, co może powodować wymuszone insercje i delecje, które mogą być droższe od zamiany.

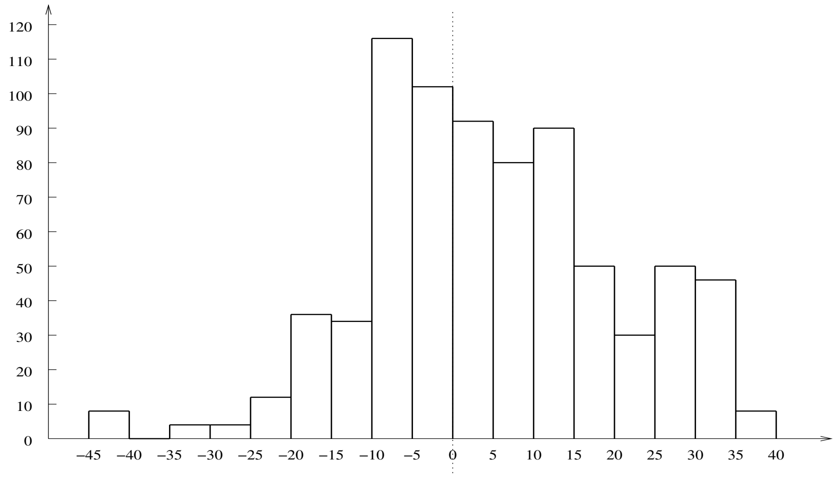
6.3 Rozkład różnic wartości przyrównań bez kontekstu i w kontekście ‘hp’

Na wykresie 6.4 widać różnice pomiędzy przyrównaniami wykonanymi na parach sekwencji bez kontekstu oraz z kontekstem ‘hp’. Różnice były niewielkie. W tym rozdziale przedstawiam wyniki dokładniejszego doświadczenia, w którym dokonanych zostało po 756 przyrównań. Pary sekwencji powstały z sekwencji należących do jednej grupy COG (zostały przyrównane do siebie wszystkie możliwe kombinacje).

Wykres 6.5 przedstawia liczby par sekwencji, dla których różnice wartości przyrównań wyniosły określone wartości. Średnia ze wszystkich różnic wyniosła 5 co informuje o tym, że wartości przyrównań bez kontekstu były przeciętnie większe. Na wykresie można zaobserwować pewne zaburzenia od rozkładu normalnego, którego można się było spodziewać, co może być tematem dalszych badań.

6.4 Własność trójkąta w tablicach

Wydawać by się mogło, że naturalną własnością tablic substytucyjnych powinna być własność trójkąta. O dziwo jednak nie zachodzi i nie jest to spowodowane



Rysunek 6.5: Wykres obrazujący różnice w wynikach przyrównań wykonanych w kontekście ‘hp’ i bez kontekstu. Pozioma oś zawiera różnice pomiędzy wartościami przyrównań bez kontekstu, a przyrównaniami w kontekście ‘hp’. Na pionowej osi jest liczba przyrównań dla których różnica była z odpowiedniego zakresu.

pojedynczymi anomaliami (własność trójkąta nie zachodzi dla ponad $1/4$ przypadków).

Rozdział 7

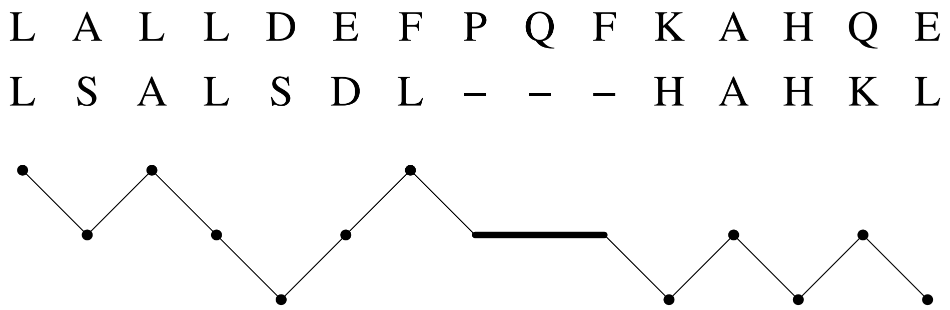
Otwarte problemy

Praca ta nie porusza wielu tematów, które są bądź mogą być istotne z punktu widzenia przydatności otrzymanych tablic. W tym rozdziale sygnalizuję pewne możliwe kierunki badań, których zgłębienie ma szansę wnieść nową wiedzę do dziedziny kontekstowego przyrównywania sekwencji.

7.1 Interpretacja zygzaka

W przeciwieństwie do przyrównania bez kontekstu mutacje na sąsiednich pozycjach w sekwencji nie są niezależne (substytucja na pozycji i zmienia kontekst mutacji na pozycjach $i - 1$ oraz $i + 1$). W związku z tym program przyrównujący oprócz wygenerowania przyrównania i obliczenia jego wartości generuje także częściowy porządek na operacjach (pojedyncze mutacje) jakie zostały wykonane na pierwszej sekwencji, aby zmutowała do drugiej. Ten częściowy porządek może być przedstawiony w postaci *zygzaka*, w którym operacje zachodzące przed innymi są umieszczone wyżej, a zachodzące później są umieszczone niżej. Przykładowe przyrównanie sekwencji wraz z wygenerowanym zygzakiem znajduje się na rysunku 7.1.

Interpretacja takiego zygzaka może przynieść ciekawe rezultaty, niemniej jednak nie jestem przynajmniej na razie wielkim optymistą. Doświadczenia pokazały, że jeśli moc zbioru kontekstów jest mała (np. 2 w przypadku kontekstu ‘hp’) to wygenerowane zygzaki nie są zbyt wyrafinowane, gdyż mutacja aminokwasu w inny należący do tej samej grupy konktekstu nie wpływa na mutacje na sąsiednich pozycjach. Z kolei jak napisałem w rozdziale 6.2 tablice dla pełnego kontekstu nie są najwyższej jakości i jeszcze przez dłuższy czas prawdopodobnie nie będzie wystarczająco dużo danych (bloków), aby wygenerować lepsze.



Rysunek 7.1: Przykładowe przyrównanie dwóch sekwencji wraz z wygenerowanym zygzakiem obrazującym kolejność zmian, które się dokonały przy mutacji z jednej sekwencji do drugiej.

7.2 Wyeliminowanie niesymetryczności tablic

Możliwym jest, że założenie o tym, że wszystkie mutacje w kolumnie bloku zachodzą z jednego aminokwasu w pozostałe jest błędnym założeniem. Jeżeli rzeczywiście tak jest to usunięcie go może wyeliminować wprowadzone w ten sposób zaburzenia do tablic wynikowych.

Inną możliwą modyfikacją mogłaby być próba wybierania aminokwasu będącego wspólnym przodkiem w jakiś inny sposób (nie ten, który ma największą względną częstość), chociaż nie przychodzi mi do głowy żadne konkretne rozwiązanie.

7.3 Dobór funkcji definiującej kontekst

Niemożność wygenerowania tablic pełnego kontekstu powoduje, że należy szukać takiego przybliżenia, które da najlepsze wyniki. Problem znalezienia najlepszej funkcji definiującej kontekst jest otwarty i prawdopodobnie nie ma jednoznacznego rozwiązania — różne zastosowania będą wymagały innych funkcji kontekstu.

Na pierwszy rzut oka widać, że jest wiele możliwości definiowania kontekstu. Można oprzeć się na jakiejś właściwości fizycznej lub chemicznej, lub na kilku jednocześnie. Można także próbować znaleźć całkowicie nowy podział, aminokwasów na konteksty, który może stać się nową właściwością aminokwasów inną niż dotychczas znane.

7.4 Metodologia oceniania osiągniętych wyników

Duża liczba możliwych sposobów obliczania tablic oraz obliczania przyrównań wymaga wypracowania metody oceniania otrzymanych wyników. Znalezienie takiej metody utrudnia fakt, że złożoność wykorzystywanych algorytmów uniemożliwia wykonywanie bardzo dużej liczby przyrównań. Opracowanie dobrej metody oceniania otrzymanych tablic jest w związku z tym chyba najtrudniejszym i najpilniejszym zagadnieniem wymagającym uwagi.

Literatura

[1] Michael S. Waterman *Introduction to Computational Biology*, Chapman & Hall (1996)

[2] Boris Steipe http://www.lmb.uni-muenchen.de/groups/bioinformatics/ *Bioinformatics Lecture 2000*

[3] Michael Zuker http://bioinfo.math.rpi.edu/ zukerm/MATH-4960/ *Algorithms in Computational Molecular Biology*

[4] Thomas R. Ioerger *The Context-Dependence of Amino Acid Properties* Intelligent Systems in Molecular Biology (1997)

[5] Liping Wei, Russ B. Altman, Jefrey T. Chang *Using the radial distributions of physical features to compare amino acid environments and align amino acid sequences* Section on Medical Informatics, Stanford University

[6] Jens Ledet Jensen, Anne-Mette Krabbe Pedersen *Probabilistic model of DNA sequence evolution with context dependent rates of substitution*, Adv. Appl. Prob. 32, 499-517 (2000)

[7] Michael Schoeniger and Arndt von Haeseler *A stochastic Model for the Evolution of Autocorrelated DNA Sequences*, Molecular Phylogenetics and Evolution, Volume 3, No. 3, September, str. 240-247 (1994)

[8] Anna Gambin, Radosław Szklarczyk, Jerzy Tiuryn, Jerzy Tyszkiewicz *Algorithms for Contextual Alignment*, Submitted to publication

[9] Steven Henikoff and Jorja G. Henikoff *Amino acid substitution matrices from protein blocks*, Proc. Natl. Acad. Sci. USA, Volume 89, pp. 10915-10919, (November 1992)

[10] Margareth O. Dayhoff *Atlas of Protein Sequence and Structure*, vol. 5, Natl. Biomed. Res. Found., Washington, D.C., 1978, ch. Supplement 3

[11] *BLOCKS Database Homepage* http://blocks.fhcrc.org/

[12] *COGs Homepage* http://www.ncbi.nlm.nih.gov/COG/