Uniwersytet Wroclawski
Wydzial Matematyki i Informatyki
Instytut Matematyczny
specjalność: Analiza danych

Bartosz Chmiela

Statystyczna analiza częstości występowania kodonów w kodzie genetycznym

Praca licencjacka
napisana pod kierunkiem
dr. hab. Krzysztofa Topolskiego

Wroclaw 2020

Spis treści

1 Wstęp do tematyki pracy

1.1 Budowa kodu genetycznego

Genom jest nośnikiem informacji w organizmamach, będącym odpowiedzialnym za odpowiedni wzrost, funkcjonowanie oraz reprodukcję organizmu. Przenoszony jest przez kwas deoksyrybonukleinowy, w skrócie *DNA* (z ang. *deoxyribonucleic acid*) znajdujący się głównie w jądrach komórkowych. *DNA* jest cząsteczką zbudowaną z dwóch nici połączonych ze sobą nukleotydami: adeniny (A), guaniny (G), cytozyny (C), tyminy (T) (lub uracylu (U) w przypadku *RNA*). Nukleotydy są parami komplementarne, tzn. że łączą się tylko w odpowiednie pary: A-T(U), G-C. Dzięki temu mając tylko jedną nić *DNA* organizm jest w stanie odtworzyć drugą – co wskazuje na prostotę duplikacji informacji genetycznej. Tak zbudowana kombinacja nukletydów tworzy genom, za pomocą którego organizm potrafi generować potrzebne mu proteiny (białka, omawiane w sekcji 1.3).

1.2 Kodony

Trójkę nukleotydów nazywa się *kodonem*, ponieważ koduje on powstanie jednego aminokwasu. Nie oznacza to jednak, że na jeden kodon przypada na jeden aminokwas – jest wręcz przeciwnie, 20 aminokwasów (oraz koniec kodowania) są kodowane przez kilka kodonów, których jest aż $4^3 = 64$. Istnieją pewnie kombinacje nukleotydów oznaczające początek i koniec, które oznaczają gdzie zaczyna i kończy się translacja DNA.

TTT	Fenyloalanina (F)	TCT	Seryna (S)	TAT	Tyrozyna (Y)	TGT	Cysteina (C)
TTC		TCC		TAC		TGC	
TTA	Leucyna (L)	TCA		TAA	STOP	TGA	STOP
TTG		TCG		TAG		TGG	Tryptofan (W)
CTT		CCT	Prolina (P)	CAT	Histydyna (H)	CGT	Arginina (R)
CTC		CCC		CAC		CGC	
CTA		CCA		CAA	Glutamina (Q)	CGA	
CTG		CCG		CAG		CGG	
ATT	Izoleucyna (I)	ACT	Treonina (T)	AAT	Aspargina (N)	AGT	Seryna (S)
ATC		ACC		AAC		AGC	
ATA		ACA		AAA	Lizyna (K)	AGA	Arginina (R)
ATG	Metionina(M)/START	ACG		AAG		AGG	
GTT	Walina (V)	GCT	Alanina (A)	GAT	Kwas asparaginowy (D)	GGT	Glicyna (G)
GTC		GCC		GAC		GGC	
GTA		GCA		GAA	Kwas glutaminowy (E)	GGA	
GTG		GCG		GAG		GGG	

Tablica 1.2.1: Tabela przedstawiająca kodony i odpowiadające im aminokwasy.

1.3 Proteiny

Proteiny (białka) powstają z połączenia aminokwasów, występują we wszystkich organizmach żywych oraz wirusach. Mają wiele funkcji – są budulcami wielu struktur występujących w komórkach oraz biorą udział w regulowaniu procesów życiowych.

1.4 Geny

Gen jest to fragment DNA, zawierający informację o białku. W genie nie tylko zapisane jest jak stworzyć białko (za pomocą kodonów), jest w nim również zawarta informacja o tym kiedy to białko potrzeba wyprodukować, w jakich ilościach (ekspresja genu) oraz do jakiego miejsca ma dotrzeć.

1.4.1 Fragmenty genu

Ze względu na to, że w genie istnieje więcej informacji niż kodowanie samego białka, istnieje problem ze zrozumieniem kodu genetycznego (w przypadku organizmów niebateryjnych). Okazuje się, że sekwencja kodująca tworzenie białka nie jest zawarta w jednym kawałku kodu genetycznego, lecz jest podzielona na kawałki, pomiędzy którymi zawarte są inne informacje. Kawałki sekwencji kodującej nazywamy *eksonami*, a inforomację zapełniającą luki pomiędzy *intronami*. W eksonach znajdziemy również kawałki kodu genetycznego nazywane *UTR* (z ang. *untranslated regions*). Suma kawałków kodujących budowę białka nazywa sIę sewkencją kodującą (z ang. *coding sequence*, w skrócie *CDS*) i własnie te będą przedmiotem badań tej pracy.

1.5 Chromosomy

Chromosom to cząsteczka przechowywująca cały materiał genetyczny, znajdująca się w jądrze komórkowym. Położenie genów w chromosomie jest opisywane za pomocą loci (l. poj. locus). Liczba chromosomów w organizmach różni się, od jednego do setek.

2 Wybór organizmów i danych

W tej pracy zbadano dwa organizmy, aby porównać czy istnieją różnice w wynikach w zależności od ilości chromosomów.

2.1 Organizm haploidalny

Organizm haploidalny, to taki który zawiera jeden zestaw chromosomów. W tej pracy jako przykład organizmu haploidalnego została wybrana bakteria *E. Coli*, która jest znana z wielu prac biologicznych. Jej popularność w badaniach genetycznych jest związana z szybkim rozmnażaniem i wysoką ekspresją protein. Bakteria ta posiada tylko jeden chromosom, który cały został przebadany w tej pracy. Użyty genom pochodzi ze strony międzynarodowego banku genów GenBank.

2.2 Organizm diploidalny

Organizm diploidalny, to taki który zawiera podwójny zestaw chromosmów. W tej pracy jako przykład organizmu diploidalnego została wybrana muszka owocowa (z łac. *Drosophila melanogaster*). Jest ona również znana z wielu badań genetycznych, ze względu szybkie rozmnażanie (cykl życia trwa zaledwie 12 dni oraz składa dużą liczbę potomstwa) oraz prostotę w utrzymaniu. Muszka owocowa posiada kilka chromosomów, jednak do badań w tej pracy został wybrany chomosom X. Użyty geneom pochodzi ze strony europejskiego banku genów Ensembl.

3 Część programistyczna

Wszystkie obliczenia prowadzone w tej pracy zostały wkonane przy użyciu języka *R*. Język ten został wybrany, ze względu na dużą ilość bilbiotek statystycznych oraz bioinformatycznych.

3.1 Rozszerzenia plików

Istnieje wiele rozszerzeń plików, w których przechowywane są sekwencje DNA: .gff (z ang. *General feature format*), .gff2, .gff3, .gtf , .fasta, .fna. Zawierają one wiele innych informacji poza samą sekwencją DNA, takich jak: miejsce na chromosomie sekwencji, rodzaj sekwencji (CDS, exon, UTR, itp.), nazwa proteiny wyprodukowanej z sekwencji, nazwa genu, identyfikatory oraz wiele innych. Do badań w tej pracy został użyty format .fna ze względu na prostotę odczytu pliku.

3.2 Biblioteka *Biostrings*

Biblioteka *Biostrings* pochodząca z Bioconductor zawiera wiele funkcji do przetwarzania łancuchów znaków, w tym specjalnie funkcje związane z sekwencjami DNA. Funkcje oraz klasy używane przy pracy nad badaniem genomów organizmów:

- Obiekty *AA_ALPHABET, DNA_ALPHABET* zawierają znaki używane do opisu odpowiednio aminokwasów oraz DNA.
- Obiekt *GENETIC_CODE* reprezentuje tabelę kodu genetycznego.
- Klasa *DNAString* jest rozszerzeniem klasy *XString* i służy do efektywnego przechowywania długich sekwencji znaków alfabetu DNA.
- Klasa *DNAStringSet* służy do przechowywania wielu obiektów *DNAString*.
- Funkcja *readDNAStringSet* wczytuje wiele sekwencji DNA z pliku (domyślnie w rozszerzeniu *fasta*).
- Funkcja *trinucleotideFrequency* zlicza występowanie znaków w sekwencji DNA (wraz z parametrem *step=3* zlicza ilości kodonów).
- Funkcja *translate* tłumaczy sekwencję kodonów na sekwencję aminokwasów.
- Funkcja *codons* znajduje początki oraz końce kodonów w sekwencji DNA.

3.3 Biblioteka *biomaRt*

Biblioteka *biomaRt* (również pochodząca z pakietu *Bioconductor*), służy do pobierania sekwencji kodujących z banku genów *Ensembl*.

3.4 Autorski kod

Kod napisany przez twórcę pracy w dużej mierze opiera się na funkcjach z biblioteki *Biostrings* (3.2), która w efektywny sposób przetwarza sekwencje kodujące. Ciała funkcji zawarte są w dodatku do pracy D.

- Funkcja *divideDNA3* dzieli i zwraca *DNAString* na 3 części: początek, środek i koniec.
- Funkcja *whichCodon* znajduje kodony odpowiadające ustalonemu aminokwasowi.
- Funkcja *codonFreq* zwraca częstość kodonów kodujących ustalony aminokwas.

- Funkcja *deleteSeq* usuwa krótkie sekwencje kodujące.
- Funkcja *testCodonInd* przeprowadza test niezależności χ^2 sprawdzający niezależność zmiennej ozaczającej rodzaj kodonu i zmiennej położenia kodonu w sekwencji kodującej.
- Funkcja *testCodonFreq* przeprowadza test zgodności χ^2 z jednostajnym rozkładem występowania kodonów.
- Funkcja *AAPos* zwraca wektor pozycji, na których występuje kodon kodujący ustalony aminokwas w sekwencji kodującej.
- Funkcja *codonPos* zwraca listę wektorów pozycji, na których występują kodony kodujące ustalony aminokwas w sekwencji kodującej.
- Funkcja *breaks* zmienia wektor pozycji na wektor długości przerw pomiędzy wystąpieniami.
- Funkcja *AABreaks* zwraca wektor długości przerw w występowaniu ustalonego aminowkwasu w sekwencji kodującej.
- Funkcja *codonBreaks* zwraca listę wektorów długości przerw w występowaniu kodonów kodujących ustalony aminokwas w sekwencji kodującej.
- Funkcja *genomeAABreaks* zwraca wektor długości przerw w występowaniu ustalonego aminowkwasu we wszystkich sekwencjach kodujących.
- Funkcja *genomeCodonBreaks* zwraca listę wektorów długości przerw w występowaniu kodonów kodujących ustalony aminokwas we wszystkich sekwencjach kodujących.
- Funkcja *quantIntervals* funkcja zwracająca wektor początków przedziałów, dzielących półprostą dodatnią na przedziały których prawdopodobieństwo jest równe.
- Funkcja *testCodonDist* przeprowadza test zgodności χ^2 z zadanym rozkładem długości przerw pomiędzy kodonami kodującymi ustalony aminokwas i również długości przerw pomiędzy ustalonym aminokwasem.
- Funkcja *transitionMatrix* zwraca macierz przejść pomiędzy kodonami kodującymi ustalony aminokwas w sekwencji kodującej.
- Funkcja *transitionGenome* zwraca macierz przejść pomiędzy kodonami kodującymi ustalony aminokwas we wszystkich sekwencjach kodujących.
- Funkcja *discrepancyCoeff* oblicza współczynnik rozbieżności testu χ^2.

3.4.1 Opis ważniejszych funkcji

- Funkcja *testCodonFreq*

  Do wykonania testu zgodności opisanego w sekcji 4.1 potrzebny jest wektor z częstością występowania każdego kodonu, który uzyskany jest przy zsumowaniu kolumn macierzy otrzymanej za pomocą funkcji *codonFreq*. Taki wektor może posłużyć jako argument funkcji *chi.test*, która dokonuje odpowiedniego testu – w tym przypadku

testu zgodności χ^2 z rozkładem jednostajnym. Domyślnie prawdopodobieństwo w tej funkcji jest obliczane jako odwrotność liczby kodonów (kategorii). Stąd można wyznaczyć liczbę oczekiwaną jako sumę wszystkich wystąpień przez liczbę kodonów.

- Funkcja *testCodonDist*

  Do wykonania testu zgodności opisanego w sekcji 4.2 potrzebny jest wektor z sumowaną ilością wystąpień długości przerw mieszczących się w odpowiednich przedziałach. Przedziały wyznaczone są za pomocą funkcji *quantIntervals*, tak aby prawdopodobieństwa przedziałów według zadanego rozkładu były w przybliżeniu równe sobie. Do przetestowania zgodności z zadanym rozkładem jest również potrzebny wektor prawdopodobieństw tych przedziałów z teoretycznego zadanego rozkładu, który jest uzyskany za pomocą odpowiedniej dystrybuanty. Tak stworzone wektory mogą posłużyć jako argumenty funkcji *chi.test*, która dokonuje odpowiedniego testu – w tym przypadku testu zgodności χ^2 z zadanym rozkładem. Za pomocą wektora prawdopodobieństw można również obliczyć liczbę oczekiwaną wystąpień w przedziale, mnożąc go przez ilość wszystkich przerw.

- Funkcja *testCodonInd*

  Do wykonania testu niezależności opisanego w sekcji 4.3 potrzebna jest tabela kontyngencji, której jeden wiersz zostaje uzyskany po zsumowaniu kolumn macierzy otrzymanej za pomocą *codonFreq*. Taką operację należy powtórzyć dla każdego *data.frame'u*, który ma w sobie początki/środki/końce wszystkich sekwencji kodujących. Tak przygotowana tablica może posłużyć jako argument do funkcji *chi.test*, która dokonuje odpowiedniego testu – w tym przypadku testu niezależności χ^2.

4 Część statystyczna

4.1 Testowanie równości prawdopodobieństw w występowaniu kodonów

Pierwszym pytaniem, które możemy postawić przy analizie częstości kodonów jest, czy występują one po równo? Aminokwasy są kodowane przez różne ilości kodonów oraz niektóre z nich wydają się być bardziej odporne na mutacje (podmianę jednego nukleotydu). Stosunek par nukleotydów w DNA również nie musi być sobie równy, co może sugerować różną częstość w występowaniu kodonów.

4.1.1 Test zgodności χ^2

Do przetestowania hipotezy o równości prawdopodobieństw może posłużyć test zgodności χ^2. Niech zmienna X ma k możliwych kategorii x_i, których prawdopodobieństwo wystąpienia wynosi $P(X = x_i) = p_i$, gdzie p_i jest nieznane oraz $\sum_{i=1}^{k} p_i = 1$. Rozważając problem testowania zgodności rozkładu $\{p_i\}_{i=1}^{k}$ z ustalonym $\{p_i^0\}_{i=1}^{k}$ można sformułować hipotezy:

$$H_0 : p_i = p_i^0,\ i = 1, \ldots, k, \quad H_1 : \exists_i\, p_i \neq p_i^0. \tag{4.1.1}$$

Statystyka testowa w tym problemie ma postać:

$$Q = \sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0}. \tag{4.1.2}$$

Gdzie n_i to liczba obserwacji i-tej kategorii, n to suma wszystkich obserwacji, a np_i^0 to oczekiwana ilość wystąpień tego kodonu. Statystyka ta przy założeniu hipotezy zerowej (4.1.1) ma w przybliżeniu rozkład χ^2 z $k-1$ stopniami swobody

W przypadku występowania kodonów zmienna X ma k możliwych wartości, gdzie k oznacza liczbę kodonów za pomocą których jest reprezentowany jeden aminokwas oraz prawdopodobieństwo wystąpienia i-tego kodonu jest nieznane i wyniosi $P(X = cod_i) = p_i$. Rozważając problem testowania zgodności rozkładu z rozkładem jednostajnym można sformułować hipotezy:

$$H_0 : p_i = \frac{1}{k},\ i = 1, \ldots, k, \quad H_1 : \exists_i\ p_i \neq \frac{1}{k}. \tag{4.1.3}$$

Statystyka testowa w tym problemie ma postać:

$$Q = \sum_{i=1}^{k} \frac{(n_i - \frac{n}{k})^2}{\frac{n}{k}}. \tag{4.1.4}$$

Gdzie n_i to liczba wystąpień i-tego kodonu, n to liczba wystąpień wszystkich kodonów kodujących ustalony aminokwas, a $\frac{n}{k}$ to oczekiwana ilość wystąpień tego kodonu. Statystyka ta przy założeniu hipotezy zerowej (4.1.3) ma w przybliżeniu rozkład χ^2 z $k-1$ stopniami swobody.

4.1.2 Współczynnik rozbieżności

Test χ^2 jest matematycznie poprawny, ale podobnie jak inne testy statystyczne, odrzuca hipotezę zerową, jeśli wielkość próbki jest wystarczająco duża. Taka sytuacja ma często miejsce przy analizie danych genetycznych, gdzie próbki liczą dziesiątki tysięcy nie są wyjątkowe. Zatem test w jego oryginalnej formie jest praktycznie bezużyteczny dla próbek o takich rozmiarach. Dlatego proponuje się brać pod uwagę nie tylko poziom istotności, ale także wielkość opisującą tzw. *test resistance*. Biorąc pod uwagę, że statystyki chi-kwadrat rosną liniowo wraz z wielkość próby, jeśli różnice między teoretyczną a empiryczncznną częstotliwości są ustalone, można za *test resistane* przyjąć współczynnik rozbieżności *discrepancy coefficient*

$$C = \frac{\chi^2}{N},$$

Powszechnie uważa się dopasowanie za zadowalające, jeśli $C \leq 0.02$ Badania symulacyjne wskazują, że w niektórych sytuacjach można zaakceptować model gdy $C \leq 0.05$

4.1.3 Wyniki

Wyniki testowania są przedstawione w tabelach 4.1.1 i 4.1.2. W tabelach znajdują się częstości wystąpień odpowiednich kodonów, oczekiwana liczba kodonów, wartość statystyki, współczynnik rozbieżności oraz p-wartość (liczba stopni swobody statystyki to ilość kodonów w tabeli -1). Część z tabel znajduje się w dodatku A.

GCA	GCC	GCG	GCT	L.oczek.	Statystyka	C	Pval
33847	41359	52805	25139	38288	10781	0.07	0.00

(a) Test dla aminokwasu A

TGC	TGT	L.oczek.	Statystyka	C	Pval
10553	8623	9588	194	0.01	0.00

(b) Test dla aminokwasu C

GAC	GAT	L.oczek.	Statystyka	C	Pval
31358	53487	42422	5772	0.07	0.00

(c) Test dla aminokwasu D

GAA	GAG	L.oczek.	Statystyka	C	Pval
64132	30848	47490	11664	0.12	0.00

(d) Test dla aminokwasu E

TTC	TTT	L.oczek.	Statystyka	C	Pval
25675	35909	30792	1701	0.03	0.00

(e) Test dla aminokwasu F

GGA	GGC	GGG	GGT	L.oczek.	Statystyka	C	Pval
14539	45651	19443	39419	29763	22980	0.19	0.00

(f) Test dla aminokwasu G

CAC	CAT	L.oczek.	Statystyka	C	Pval
15209	20839	18024	879	0.02	0.00

(g) Test dla aminokwasu H

ATA	ATC	ATT	L.oczek.	Statystyka	C	Pval
9058	38856	48152	32022	26052	0.27	0.00

(h) Test dla aminokwasu I

AAA	AAG	L.oczek.	Statystyka	C	Pval
55421	18221	36821	18791	0.26	0.00

(i) Test dla aminokwasu K

CTA	CTC	CTG	CTT	TTA	TTG	L.oczek.	Statystyka	C	Pval
6323	17093	83341	18640	22428	21021	28141	135700	0.80	0.00

(j) Test dla aminokwasu L

AAC	AAT	L.oczek.	Statystyka	C	Pval
35229	30956	33093	276	0.00	0.00

(k) Test dla aminokwasu N

CCA	CCC	CCG	CCT	L.oczek.	Statystyka	C	Pval
13707	9150	36865	11873	17899	27384	0.38	0.00

(l) Test dla aminokwasu P

CAA	CAG	L.oczek.	Statystyka	C	Pval
23882	48281	36082	8250	0.11	0.00

(m) Test dla aminokwasu Q

AGA	AGG	CGA	CGC	CGG	CGT	L.oczek.	Statystyka	C	Pval
4771	3057	6414	34270	10431	33200	15357	67960	0.74	0.00

(n) Test dla aminokwasu R

Tablica 4.1.1: Wyniki testu zgodności dla E.Coli.

GCA	GCC	GCG	GCT	L.oczek.	Statystyka	C	Pval
51872	134719	64746	47498	74709	66424	0.22	0.00

(a) Test dla aminokwasu A

TGC	TGT	L.oczek.	Statystyka	C	Pval
49482	18351	33917	14287	0.21	0.00

(b) Test dla aminokwasu C

GAC	GAT	L.oczek.	Statystyka	C	Pval
91489	105796	98643	1038	0.01	0.00

(c) Test dla aminokwasu D

GAA	GAG	L.oczek.	Statystyka	C	Pval
71820	170095	120958	39923	0.17	0.00

(d) Test dla aminokwasu E

TTC	TTT	L.oczek.	Statystyka	C	Pval
78122	43607	60865	9786	0.08	0.00

(e) Test dla aminokwasu F

GGA	GGC	GGG	GGT	L.oczek.	Statystyka	C	Pval
64856	121726	17863	56212	65164	84661	0.32	0.00

(f) Test dla aminokwasu G

CAC	CAT	L.oczek.	Statystyka	C	Pval
62728	43592	53160	3444	0.03	0.00

(g) Test dla aminokwasu H

ATA	ATC	ATT	L.oczek.	Statystyka	C	Pval
32541	87366	55222	58376	26001	0.15	0.00

(h) Test dla aminokwasu I

AAA	AAG	L.oczek.	Statystyka	C	Pval
51408	141463	96436	42048	0.22	0.00

(i) Test dla aminokwasu K

CTA	CTC	CTG	CTT	TTA	TTG	L.oczek.	Statystyka	C	Pval
29260	53673	150561	26647	14303	57223	55278	221804	0.67	0.00

(j) Test dla aminokwasu L

AAC	AAT	L.oczek.	Statystyka	C	Pval
95438	84558	89998	658	0.00	0.00

(k) Test dla aminokwasu N

CCA	CCC	CCG	CCT	L.oczek.	Statystyka	C	Pval
54750	67090	73358	21215	54103	29970	0.14	0.00

(l) Test dla aminokwasu P

CAA	CAG	L.oczek.	Statystyka	C	Pval
62388	156284	109336	40318	0.18	0.00

(m) Test dla aminokwasu Q

AGA	AGG	CGA	CGC	CGG	CGT	L.oczek.	Statystyka	C	Pval
16811	21865	32345	76431	32782	35949	36031	61792	0.29	0.00

(n) Test dla aminokwasu R

Tablica 4.1.2: Wyniki testu zgodności dla muszki owocowej.

Rozważając tylko p-wartości z tabel 4.1.1 i 4.1.2 można wywnioskować, że kodony nie występują równomiernie w sekwencjach kodujących bakterii i muszki owocowej. Jednakże współczynnik rozbieżności sugeruje, że większość testów nie jest znacząca, te testy które są znaczące odrzucają hipotezę o równym prawdopodobieństwie. Znaczące testy najczęściej są przy aminokwasie, który jest kodowany za pomocą dwoch kodonów.

4.2 Testowanie zgodności z rozkładami

Wyniki poprzedniej sekcji 4.1 sugerują, że występowanie kodonów pochodzi z innego rozkładu niż jednostajny. W tej pracy spróbowano odnaleźć rozkład długości przerw pomiędzy wystąpieniami aminokwasu oraz kodonów kodujących ten aminokwas. Zmienna X_i będzie oznaczała długość przerwy pomiędzy ponownym wystąpieniem kodonu, a n liczbę wszystkich zmiennych X_i.

4.2.1 Rozkład geometryczny

Rozkład geometryczny jest rozkładem dyskretnym, który mierzy prawdopodobieństwo pojawienia się wygranej po serii porażek. Odpowiada to sytuacji, w której wygraną jest pojawienie się ustalonego kodonu, a przegraną wystąpienie jakiegokolwiek innego kodonu. Typowym przykładem zastosowania rozkładu geometrycznego jest modelowanie rzutu monetą. Jest zatem naturalnym pomysłem przybliżenie tym rozkładem przerw w występowaniu kodonów i aminokwasów. Funkcja masy rozkładu $Geom(\theta)$ ma postać:

$$P(X = k|\theta) = (1-\theta)^k\theta, \quad \theta \in (0,1). \tag{4.2.1}$$

4.2.2 Rozkład logarytmiczny

Rozkład logarytmiczny jest również rozkładem dyskretnym, wyprowadzonym z rozwinięcia logarytmu w szereg Taylor'a:

$$-\log(1-p) = \sum_{k=1}^{\infty} \frac{p^k}{k}, \quad p \in (0,1).$$

Gdzie log oznacza logarytm naturalny. R. Fisher wykorzystał ten rozkład do modelowania względnej liczebności gatunków. Funkcja masy rozkładu $Log(\theta)$ ma postać:

$$P(X = k|\theta) = \frac{-1}{\log(1-\theta)}\frac{p^k}{k}, \quad \theta \in (0,1). \tag{4.2.2}$$

4.2.3 Estymator największej wiarogodności

Do znalezienia parametru rozkładu do którego chcemy dopasować dane, można posłużyć się estymatorem największej wiarogodności (ENW), który jest znajdywany za pomocą maksymalizacji logarytmu funkcji wiarogodności, która ma postać:

$$L(\theta) = \Pi_{i=1}^{n} f(x_i|\theta). \tag{4.2.3}$$

W przypadku rozkładu geometrycznego ENW można wyznaczyć i ma postać:

$$\hat{\theta} = \frac{n}{\sum_i^n X_i}. \tag{4.2.4}$$

Natomiast w przypadku rozkładu logarytmicznego ENW jest rozwiązaniem równania:

$$\bar{X} = \frac{\hat{\theta}}{-(1-\hat{\theta})\log(1-\hat{\theta})}. \tag{4.2.5}$$

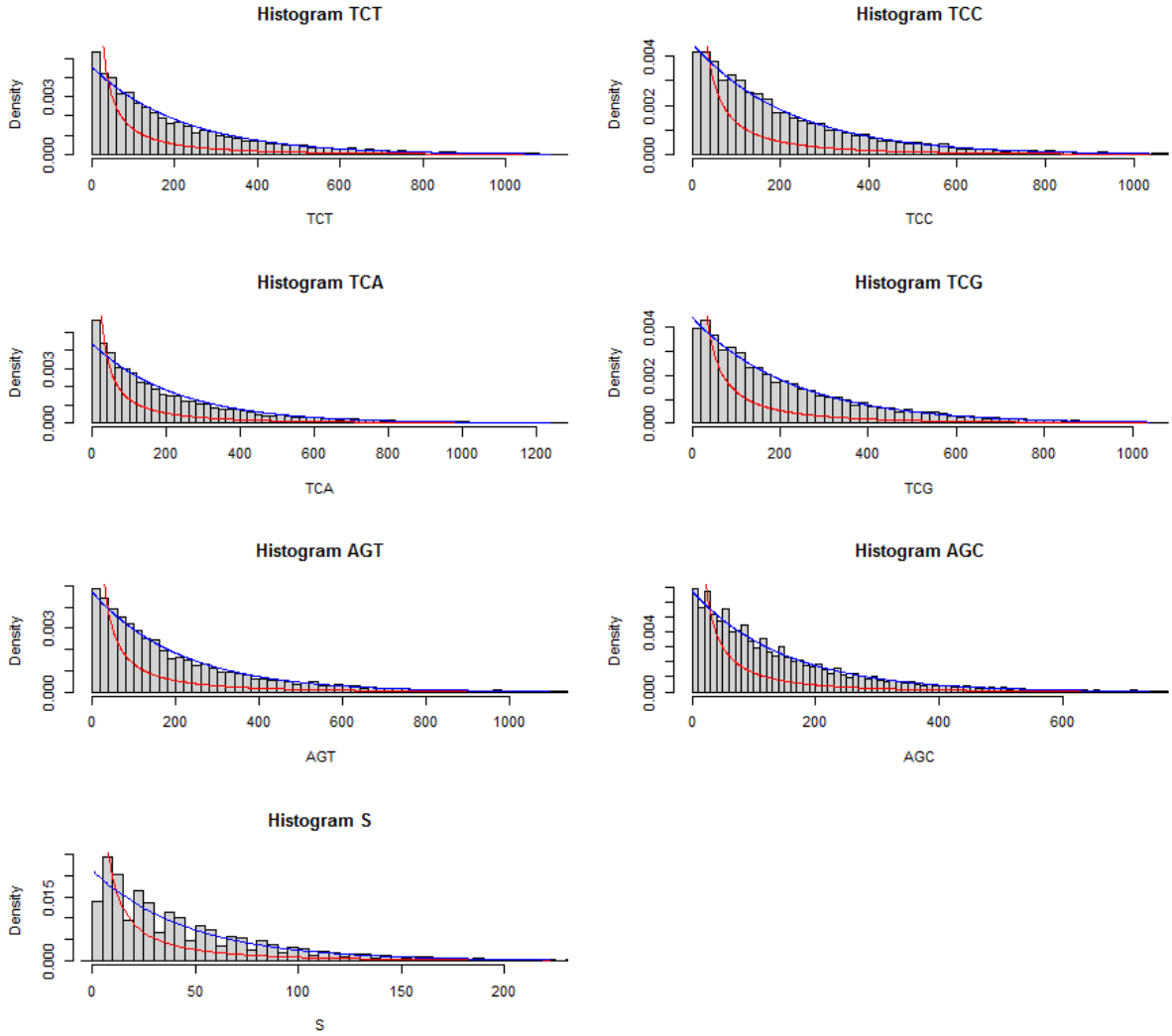
Lub można wykorzystać prostszą konstrukcję:

$$\hat{\theta} = 1 - \sum_{j\geq 1} \frac{j f_j}{\sum_{j\geq 1} j^2 f_j}. \tag{4.2.6}$$

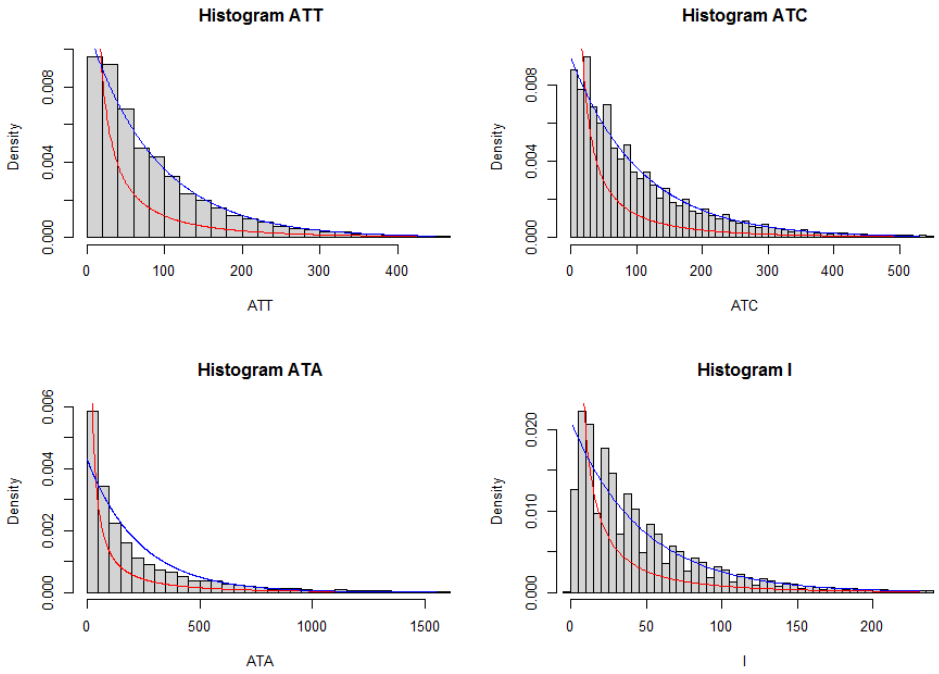
Gdzie f_j jest proporcją obserwacji które są równe j oraz $\bar{X}$ jest średnią próbkową.

4.2.4 Historgramy

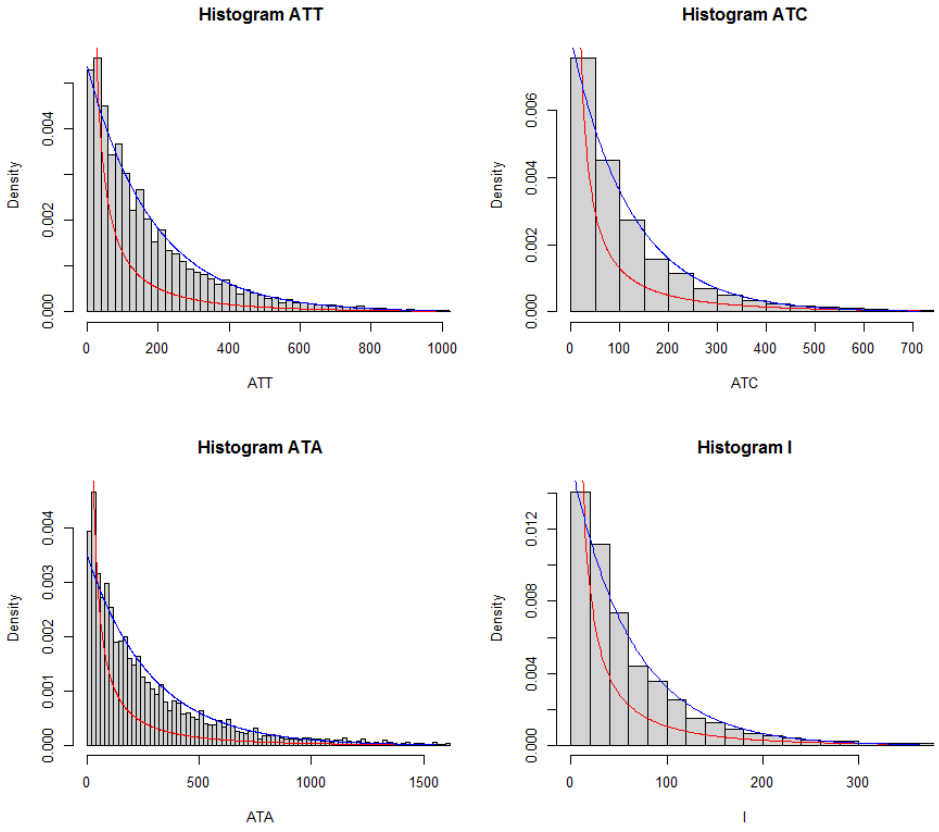
Pierwszym krokiem badania zgodności rozkładu zmiennej z założonym jest graficzne porównanie histogramu gęstościowego z funkcją masy zakładanego rozkładu. Zamieszczone zostały tutaj przykładowe histogramy dla obu organizmów wraz z wykresami funkcji masy. Czerwona linia to wykres funkcji masy rozkładu logarytmicznego, a niebieska linia to wykres funkcji masy rozkładu geometrycznego.



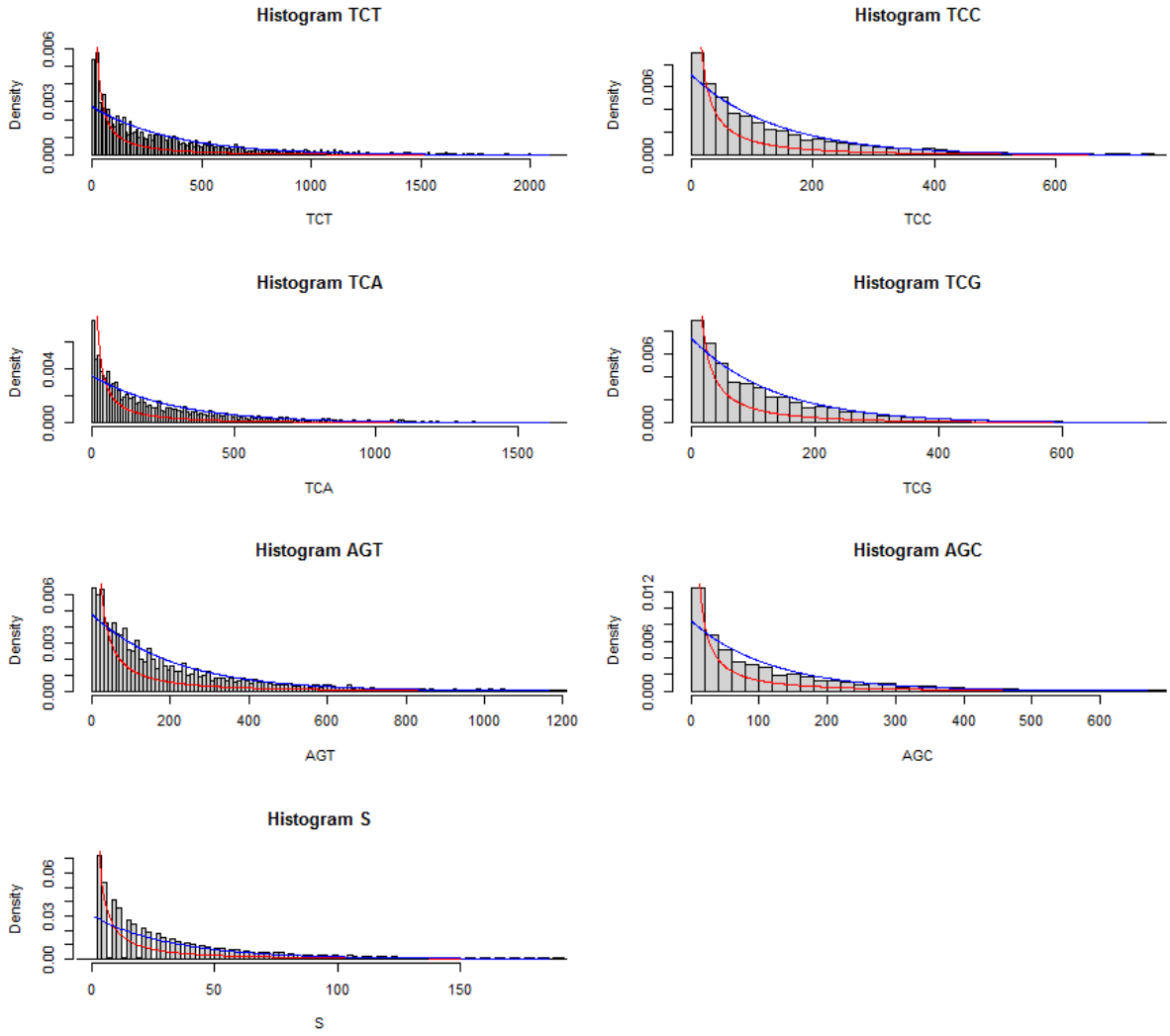
Rysunek 4.2.1: Histogramy długości przerw pomiędzy kodonami oraz aminokwasem S dla bakterii E.Coli.



Rysunek 4.2.2: Histogramy długości przerw pomiędzy kodonami oraz aminokwasem I dla bakterii E.Coli.



Rysunek 4.2.3: Histogramy długości przerw pomiędzy kodonami oraz aminokwasem I dla muszki owocowej.

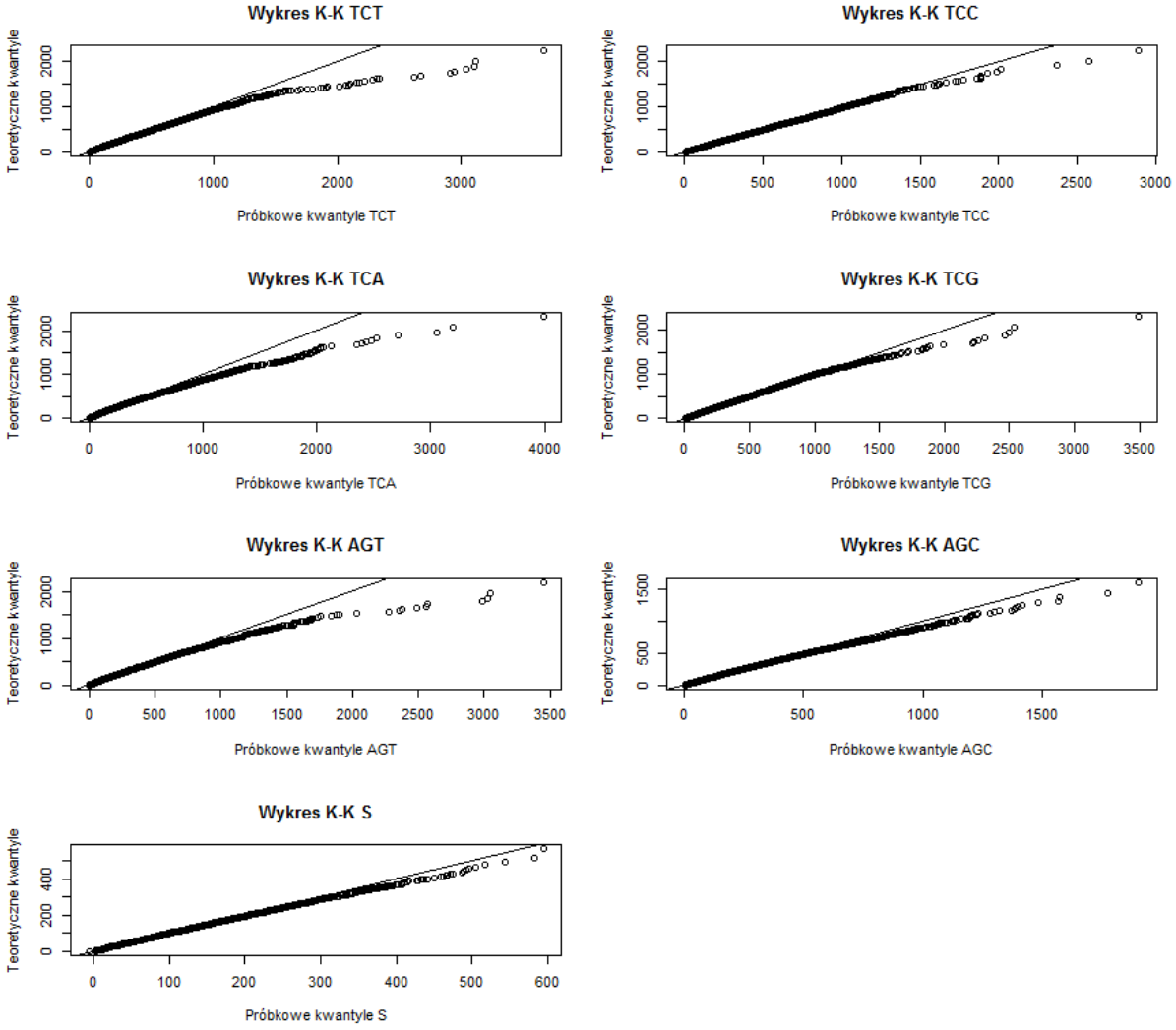


Rysunek 4.2.4: Histogramy długości przerw pomiędzy kodonami oraz aminokwasem S dla muszki owocowej.

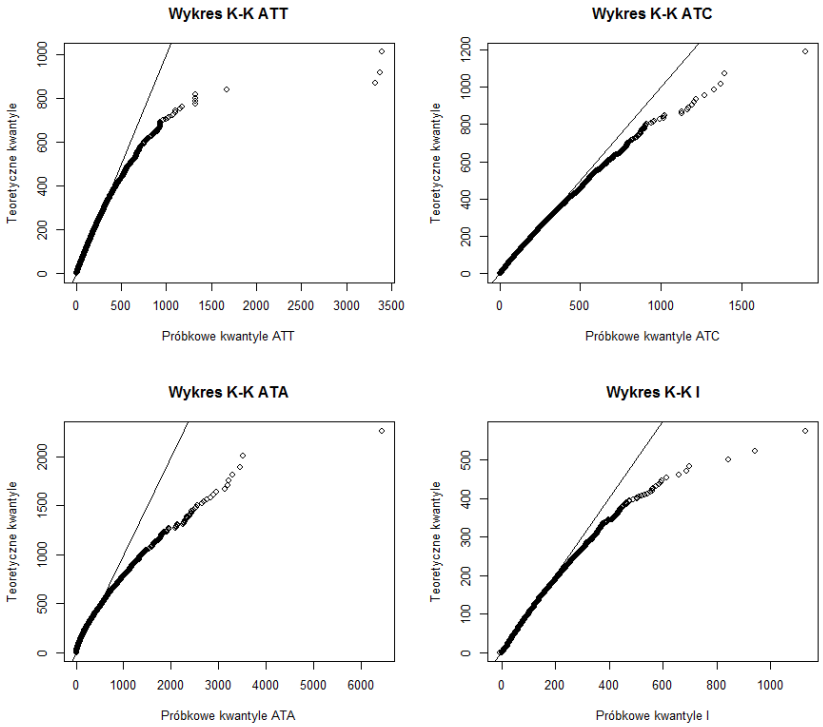
Funkcja masy rozkładu geometrycznego jest zbliżona w kształcie do histogramów, jedyne jej niedopasowanie pojawia się na początku dziedziny, w tym miejscu natomiast przybliżenie funkcją masy rozkładu logarytmicznego wydaje się sensowniejsze. Ten drugi jednak nie jest dopasowany na reszcie dziedziny. Z tego też powodu w dalszej części pracy zajemiemy się badaniem zgodności tylko rozkładu geometrycznego.

4.2.5 Wykresy kwantylowo-kwantylowe

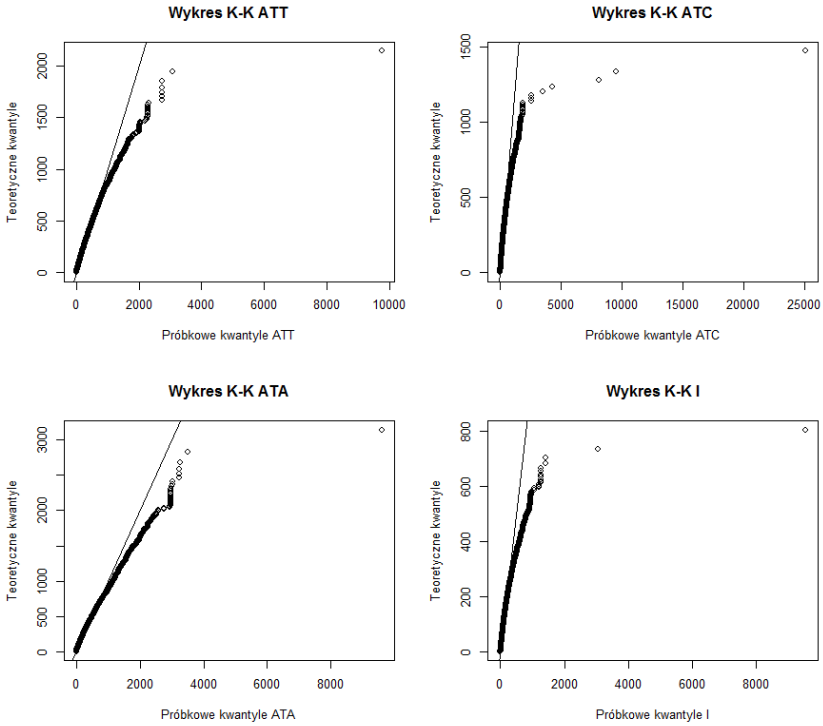
Dla potwierdzenia intuicji dobrze jest zobaczyć wykresy kwantylowo-kwantylowe.



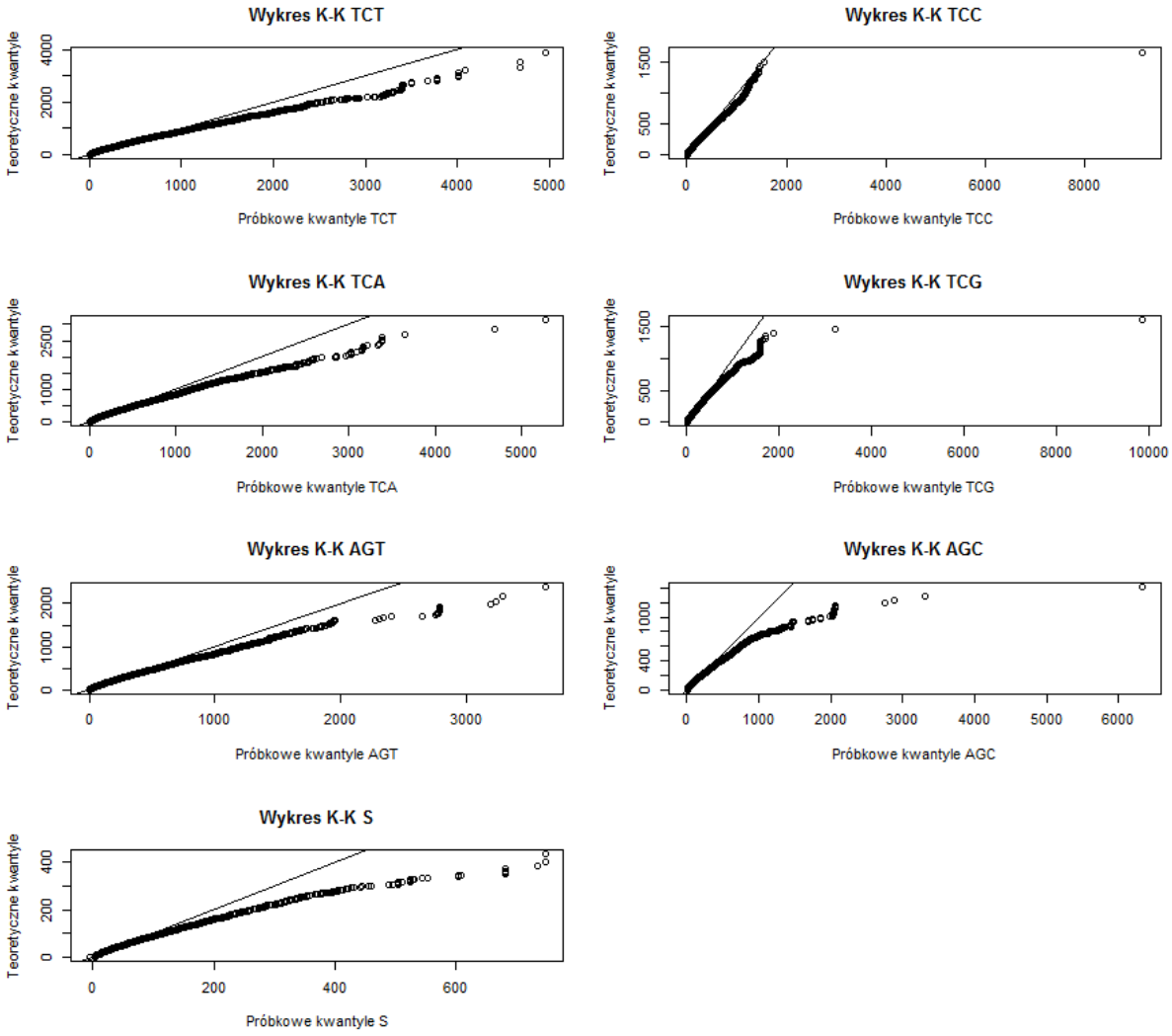
Rysunek 4.2.5: Wykresy kwantylowo-kwantylowe długości przerw pomiędzy kodonami oraz aminokwasami S dla bakterii E.Coli.



Rysunek 4.2.6: Wykresy kwantylowo-kwantylowe długości przerw pomiędzy kodonami oraz aminokwasami I dla bakterii E.Coli.



Rysunek 4.2.7: Wykresy kwantylowo-kwantylowe długości przerw pomiędzy kodonami oraz aminokwasami I dla muszki owocowej



Rysunek 4.2.8: Wykresy kwantylowo-kwantylowe długości przerw pomiędzy kodonami oraz aminokwasami S dla muszki owocowej

Na ogół widać w przybliżeniu dobre dopasowanie rozkładem geometrycznym, odstęp od prostej pojawia się dla większych wartości zmiennych.

4.2.6 Test zgodności χ^2

Do przetestowania hipotezy o zgodności rozkładu danych z rozkładem geometrycznym może posłużyć test zgodności χ^2. Ze względu na dużą ilość pojedynczych wartości zmiennej X_i, warunek $np_i^0 \geq 5$, gdzie p_i^0 oznacza zakładne w hipotezie zerowej prawdopodobieństwo wystąpienia długości i-tej przerwy, może nie zachodzić, co prowadzi do niepoprawnej aproksymacji rozkładem χ^2. Z tego powodu stworzone zostały nowe zmienne, zliczające wystąpienia długości przerw w ustalonych przedziałach, tak aby zakładane prawdopodobieństwa w tych przedziałach były sobie w przybliżeniu równe. Takie przedziały można uzyskać za pomocą funkcji kwantylowej. Hipotezy w tym problemie są zatem sformułowane tak samo jak w problemie (4.1.1).

4.2.7 Wyniki

Wyniki zostaly przedstawione w tabelach 4.2.1 i 4.2.2. W tabelach znajdują się liczby wystąpień przerw kodonu w odpowiednim przedziale (numerowanym od 1 do 5), oczekiwana liczba wystąpień w przedziale, wynik statystyki testowej (liczba stopni swobody to ilość przedziałów -1) oraz p-wartość. Część z tabel znajduje się w dodatku B.

	1	2	3	4	5	Stat	C	Pval
Liczba GCT	5368	5420	4758	4625	4968			
L.Oczek. GCT	4898	5080	5071	5059	5028	125	4.98×10^{-3}	4.29×10^{-26}
Liczba GCC	8292	8433	8710	8012	7912			
L.Oczek. GCC	8200	8129	8449	8289	8291	47	1.14×10^{-3}	1.47×10^{-9}
Liczba GCA	7022	7265	6492	6299	6769			
L.Oczek. GCA	6742	6782	6743	6776	6802	89	2.63×10^{-3}	2.15×10^{-18}
Liczba GCG	10020	11314	11016	10115	10340			
L.Oczek. GCG	10064	11030	10412	10685	10612	79	1.51×10^{-3}	1.83×10^{-16}
Liczba A	16173	39823	32697	32099	32358			
L.Oczek. A	27619	32377	30569	31383	31203	6663	4.35×10^{-2}	0

(a) Test dla aminokwasu A

	1	2	3	4	5	Stat	C	Pval
Liczba TGT	2006	1741	1564	1594	1718			
L.Oczek. TGT	1703	1729	1739	1723	1727	81	9.44×10^{-3}	8.96×10^{-17}
Liczba TGC	2399	2134	1987	2008	2025			
L.Oczek. TGC	2104	2101	2114	2120	2112	59	5.60×10^{-3}	4.57×10^{-12}
Liczba C	4431	3825	3607	3417	3896			
L.Oczek. C	3914	4053	4007	3992	4013	182	9.12×10^{-3}	2.56×10^{-38}

(b) Test dla aminokwasu C

	1	2	3	4	5	Stat	C	Pval
Liczba GAT	8887	11961	11413	11034	10192			
L.Oczek. GAT	10673	10598	10794	10703	10716	545	1.02×10^{-2}	8.99×10^{-117}
Liczba GAC	5986	6463	6734	6060	6115			
L.Oczek. GAC	6212	6214	6367	6286	6277	51	1.64×10^{-3}	1.69×10^{-10}
Liczba D	13332	18095	18427	18737	16254			
L.Oczek. D	15901	17967	16682	17245	17088	768	9.06×10^{-3}	4.50×10^{-165}

(c) Test dla aminokwasu D

	1	2	3	4	5	Stat	C	Pval
Liczba GAA	10966	15070	12826	12886	12384			
L.Oczek. GAA	12517	12975	12876	12903	12858	548	8.55×10^{-3}	2.25×10^{-117}
Liczba GAG	6032	6430	6459	5976	5951			
L.Oczek. GAG	6132	6088	6204	6226	6196	51	1.66×10^{-3}	2.17×10^{-10}
Liczba E	17359	18914	22062	18452	18193			
L.Oczek. E	17859	19509	19635	18893	19127	388	4.08×10^{-3}	1.01×10^{-82}

(d) Test dla aminokwasu E

	1	2	3	4	5	Stat	C	Pval
Liczba TTT	7399	7419	7610	6734	6747			
L.Oczek. TTT	7151	7115	7259	7189	7193	95	2.65×10^{-3}	1.11×10^{-19}
Liczba TTC	5152	5388	5391	4673	5071			
L.Oczek. TTC	5112	5062	5218	5140	5141	70	2.75×10^{-3}	1.76×10^{-14}
Liczba F	10355	14136	13250	11658	12185			
L.Oczek. F	11752	12790	12459	12309	12377	395	6.42×10^{-3}	2.27×10^{-84}

(e) Test dla aminokwasu F

Tablica 4.2.1: Wyniki testu zgodności dla bakterii Ecoli.

	1	2	3	4	5	Stat	C	Pval
Liczba GCT	13518	8256	7874	8011	9839			
L.Oczek. GCT	9427	9543	9481	9517	9527	2469	5.20×10^{-2}	0
Liczba GCC	28453	29069	26046	24748	26403			
L.Oczek. GCC	26837	26661	27172	26970	27078	561	4.17×10^{-3}	3.58×10^{-120}
Liczba GCA	16632	8210	8095	7985	10950			
L.Oczek. GCA	10314	10306	10437	10389	10424	5405	1.04×10^{-1}	0
Liczba GCG	17080	12609	11050	11111	12896			
L.Oczek. GCG	12662	13198	12927	12939	13016	2099	3.24×10^{-2}	0
Liczba A	68250	69051	51397	47842	62295			
L.Oczek. A	57787	61666	59490	58638	61257	5885	1.97×10^{-2}	0

(a) Test dla aminokwasu A

	1	2	3	4	5	Stat	C	Pval
Liczba TGT	5551	3303	2897	2930	3670			
L.Oczek. TGT	3659	3665	3683	3664	3678	1328	7.24×10^{-2}	2.36×10^{-286}
Liczba TGC	16072	9709	7904	6692	9105			
L.Oczek. TGC	9866	9808	9997	9903	9905	5447	1.10×10^{-1}	0
Liczba C	22237	13224	10572	9462	12338			
L.Oczek. C	13468	13719	13655	13557	13672	7817	1.15×10^{-1}	0

(b) Test dla aminokwasu C

	1	2	3	4	5	Stat	C	Pval
Liczba GAT	25629	20375	20402	18607	20783			
L.Oczek. GAT	20716	21243	21307	21269	21258	1582	1.50×10^{-2}	0
Liczba GAC	22741	18152	17021	15452	18123			
L.Oczek. GAC	17694	18704	18312	18448	18329	2035	2.23×10^{-2}	0
Liczba D	41724	46971	34268	35270	39052			
L.Oczek. D	38173	39664	40350	39070	40047	2987	1.51×10^{-2}	0

(c) Test dla aminokwasu D

	1	2	3	4	5	Stat	C	Pval
Liczba GAA	21084	13327	12441	10522	14446			
L.Oczek. GAA	14092	14483	14506	14330	14406	4867	6.78×10^{-2}	0
Liczba GAG	42585	35990	31356	27786	32378			
L.Oczek. GAG	33293	33415	34397	34734	34253	4553	2.68×10^{-2}	0
Liczba E	61670	49690	44055	41183	45317			
L.Oczek. E	47625	48183	47642	49649	48829	6155	2.54×10^{-2}	0

(d) Test dla aminokwasu E

	1	2	3	4	5	Stat	C	Pval
Liczba TTT	10317	9660	8211	7489	7930			
L.Oczek. TTT	8579	8834	8677	8756	8758	715	1.64×10^{-2}	1.29×10^{-153}
Liczba TTC	19151	17548	14453	12919	14051			
L.Oczek. TTC	15179	15870	15590	15852	15628	2001	2.56×10^{-2}	0
Liczba F	28945	26589	23787	20484	21924			
L.Oczek. F	23444	24944	24444	24301	24619	2311	1.90×10^{-2}	0

(e) Test dla aminokwasu F

Tablica 4.2.2: Wyniki testu zgodności dla muszki owocowej.

Z tabel jednoznacznie wynika, że rozkłady tych zmiennych nie pochodzą z rozkładu geometrycznego, zwłaszcza w przypadku wystąpień samego aminokwasu. W przypadku bakterii E. Coli wszystkie testy są istotne, natomiast w przypadku muszki owocowej kilka z testów może być nie istotna – współczynnik rozbieżności jest w przybliżeniu 0.05.

4.2.8 Liczba przedziałów

Wybór 5 przedziałów jest spowodowany granicznymi wartościami współczynnika rozbieżności (ok. $0.02 \leq C \leq 0.05$, część testów jest istotna, a część nie). Dla większej ilości przedziałów, statystyka χ^2 często zwiększa się, a ponieważ liczba obserwacji zostaje taka sama, współczynnik zwiększa się, przez co wnioskujemy o braku istotności testu. Przykładowe wyniki statystyki testowej, współczynnika zbieżności oraz p-wartości dla różnej ilosci przedziałów są podane w tabelach 4.2.3, 4.2.4, 4.2.5 i 4.2.6.

	Stat	C	Pval
GCT	120	4.78×10^{-3}	8.26×10^{-27}
GCC	44	1.07×10^{-3}	2.66×10^{-10}
GCA	89	2.65×10^{-3}	3.55×10^{-20}
GCG	56	1.07×10^{-3}	5.87×10^{-13}
A	1401	9.15×10^{-3}	5.43×10^{-305}

(a) Test dla aminokwasu A

	Stat	C	Pval
GAT	109	2.05×10^{-3}	1.58×10^{-24}
GAC	73	2.34×10^{-3}	1.19×10^{-16}
D	964	1.14×10^{-2}	4.45×10^{-210}

(b) Test dla aminokwasu D

Tablica 4.2.3: Wyniki testu zgodności dla bakterii Ecoli dla 3 przedziałów.

	Stat	C	Pval
GCT	206	8.23×10^{-3}	1.21×10^{-39}
GCC	94	2.29×10^{-3}	1.80×10^{-16}
GCA	254	7.52×10^{-3}	1.09×10^{-49}
GCG	251	4.77×10^{-3}	4.21×10^{-49}
A	26285	1.72×10^{-1}	0

(a) Test dla aminokwasu A

	Stat	C	Pval
GAT	811	1.52×10^{-2}	5.75×10^{-169}
GAC	161	5.16×10^{-3}	2.96×10^{-30}
D	1482	1.75×10^{-2}	9.49×10^{-314}

(b) Test dla aminokwasu D

Tablica 4.2.4: Wyniki testu zgodności dla bakterii Ecoli dla 10 przedziałów.

	Stat	C	Pval
GCT	1041	2.19×10^{-2}	8.84×10^{-227}
GCC	663	4.93×10^{-3}	6.83×10^{-145}
GCA	2696	5.20×10^{-2}	0
GCG	1175	1.82×10^{-2}	4.57×10^{-256}
A	1462	4.89×10^{-3}	2.68×10^{-318}

(a) Test dla aminokwasu A

	Stat	C	Pval
GAT	1315	1.24×10^{-2}	2.66×10^{-286}
GAC	1161	1.27×10^{-2}	7.38×10^{-253}
D	2883	1.46×10^{-2}	0

(b) Test dla aminokwasu D

Tablica 4.2.5: Wyniki testu zgodności dla muszki owocowej dla 3 przedziałów.

	Stat	C	Pval
GCT	3183	6.70×10^{-2}	0
GCC	2083	1.55×10^{-2}	0
GCA	10256	1.98×10^{-1}	0
GCG	3317	5.12×10^{-2}	0
A	63033	2.11×10^{-1}	0

(a) Test dla aminokwasu A

	Stat	C	Pval
GAT	1765	1.67×10^{-2}	0
GAC	2295	2.51×10^{-2}	0
D	5844	2.96×10^{-2}	0

(b) Test dla aminokwasu D

Tablica 4.2.6: Wyniki testu zgodności dla muszki owocowej dla 10 przedziałów.

4.3 Testowanie niezależności

Brak dobrego dopasowania rozkładu geometrycznego do rozkładu zminnej opisującej długości przerw w występowaniu odpowiedniego kodonu może być spowodowany złamaniem założenia o niezależności. Kodony opisują informację genetyczną, zatem niekoniecznie powinny być traktowane jako zmienne losowe. Warto jest zatem sprawdzić niezależność zmiennej X opisującej wystąpienie kodonu ze zmienną Y opisującą położenie kodonu w sekwencji kodującej (początek/środek/koniec). Do przetestowania tego problemu można wykorzystać test niezależności, korzystający ze statystyki χ^2.

4.3.1 Test niezależności χ^2

Niech zmienna X ma k kategorii, równe liczbie kodonów opisujących ustalony aminokwas, a zmienna Y ma $l = 3$ kategorii odpowiadających położeniu kodonu w sekwencji kodującej (początek/środek/koniec). Oznaczmy przez p_{ij} prawdopodobieństwo zaobserwowania i-tego kodonu w j-tej lokacji, $p_{i.}$ rozkład brzegowy X i $p_{j.}$ rozkład brzegowy Y. Sformułowanie hipotez ma postać:

$$H_0 : \text{X i Y są niezależne}, \quad H_1 : \text{X i Y są zależne}. \tag{4.3.1}$$

Co sprowadza się do matematycznej postaci:

$$H_0 : p_{ij} = p_{i.}p_{j.},\ i = 1, \ldots, k,\ j = 1, \ldots, l \quad H_1 : \exists_{i,j}\ p_{ij} \neq p_{i.}p_{j.}. \tag{4.3.2}$$

Statystyka testowa w tym problemie ma postać:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}. \tag{4.3.3}$$

Gdzie n_{ij} to liczba zaobserwowanych realizacji zmiennych należących do i-tej i j-tej kategorii, $n_{i.} = \sum_{j=1}^{l} n_{ij}$ oraz $n_{.j} = \sum_{i=1}^{k} n_{ij}$. Statystyka ta przy założeniu hipotezy zerowej (4.3.2) ma w przybliżeniu rozkład χ^2 z $(k-1)(l-1)$ stopniami swobody.

4.3.2 Wyniki

Wyniki zostały przedstawione w tabelach 4.3.1 i 4.3.2. W tabelach znajdują się ilości wystąpień kodonu w odpowiedniej części sekwencji kodującej, wartość statystyki testowej, współczynnik rozbieżności oraz p-wartość. Część z tabel znajduje się w dodatku C.

	beg	mid	end	Stat	C	Pval
GCT	11170	11238	11231	1360	9.54×10^{-3}	7.39×10^{-291}
GCC	14501	10072	9356			
GCA	16645	14942	13926			
GCG	8707	10509	10332			

(a) Test dla aminokwasu A

	beg	mid	end	Stat	C	Pval
TGT	3404	11598	11408	191	4.47×10^{-3}	3.06×10^{-42}
TGC	2904	6731	6713			

(b) Test dla aminokwasu C

	beg	mid	end	Stat	C	Pval
GAT	10149	6909	6788	7	1.18×10^{-4}	2.35×10^{-2}
GAC	17266	11147	11192			

(c) Test dla aminokwasu D

	beg	mid	end	Stat	C	Pval
GAA	20608	11933	12464	32	5.06×10^{-4}	7.55×10^{-8}
GAG	9502	4901	5356			

(d) Test dla aminokwasu E

	beg	mid	end	Stat	C	Pval
TTT	7829	7687	7426	301	5.50×10^{-3}	3.47×10^{-66}
TTC	13213	9633	9058			

(e) Test dla aminokwasu F

	beg	mid	end	Stat	C	Pval
GGT	4808	8042	8558	1899	1.66×10^{-2}	0
GGC	14762	12766	12627			
GGA	6474	6897	7138			
GGG	12663	9773	9707			

(f) Test dla aminokwasu G

	beg	mid	end	Stat	C	Pval
CAT	4908	6617	6383	120	3.08×10^{-3}	6.90×10^{-27}
CAC	6864	7102	7200			

(g) Test dla aminokwasu H

	beg	mid	end	Stat	C	Pval
ATT	3325	5559	6155	2690	3.28×10^{-2}	0
ATC	13089	9254	8927			
ATA	16521	9523	9575			

(h) Test dla aminokwasu I

	beg	mid	end	Stat	C	Pval
AAA	18310	13832	14133	1407	2.03×10^{-2}	2.85×10^{-306}
AAG	5830	8404	8900			

(i) Test dla aminokwasu K

	beg	mid	end	Stat	C	Pval
TTA	2288	3384	3535	3900	2.63×10^{-2}	0
TTG	6384	4764	4484			
CTT	27061	15243	14597			
CTC	7088	5730	5396			
CTA	8280	7700	8156			
CTG	7022	8794	8641			

(j) Test dla aminokwasu L

	beg	mid	end	Stat	C	Pval
AAT	11093	10031	10118	54	9.36×10^{-4}	1.34×10^{-12}
AAC	10428	8166	8599			

(k) Test dla aminokwasu N

Tablica 4.3.1: Tablice kontyngencji dla bakterii E.Coli.

	beg	mid	end	Stat	C	Pval
GCT	18809	33296	33962	10828	3.31×10^{-2}	0
GCC	44724	30225	31589			
GCA	21696	25074	27192			
GCG	14804	23022	22958			

(a) Test dla aminokwasu A

	beg	mid	end	Stat	C	Pval
TGT	17408	26697	27316	568	5.38×10^{-3}	4.25×10^{-124}
TGC	6122	14268	13715			

(b) Test dla aminokwasu C

	beg	mid	end	Stat	C	Pval
GAT	30293	20138	19802	1	1.27×10^{-5}	3.84×10^{-1}
GAC	34742	23127	22365			

(c) Test dla aminokwasu D

	beg	mid	end	Stat	C	Pval
GAA	24125	24342	21594	4418	2.45×10^{-2}	0.00×10^{0}
GAG	55389	27529	27363			

(d) Test dla aminokwasu E

	beg	mid	end	Stat	C	Pval
TTT	25355	17819	17024	270	2.72×10^{-3}	2.31×10^{-59}
TTC	14350	12633	11920			

(e) Test dla aminokwasu F

	beg	mid	end	Stat	C	Pval
GGT	22156	30330	29035	7849	2.97×10^{-2}	0
GGC	41104	28519	29512			
GGA	6276	13446	14043			
GGG	17845	16260	15967			

(f) Test dla aminokwasu G

	beg	mid	end	Stat	C	Pval
CAT	20032	23590	23229	313	2.55×10^{-3}	8.73×10^{-69}
CAC	14250	20994	20814			

(g) Test dla aminokwasu H

	beg	mid	end	Stat	C	Pval
ATT	11109	10213	10435	77	4.95×10^{-4}	5.69×10^{-16}
ATC	29132	25098	24448			
ATA	17003	15305	14059			

(h) Test dla aminokwasu I

	beg	mid	end	Stat	C	Pval
AAA	18037	17073	16238	2686	1.87×10^{-2}	0
AAG	45352	23630	23009			

(i) Test dla aminokwasu K

	beg	mid	end	Stat	C	Pval
TTA	9479	11278	11428	10893	3.74×10^{-2}	0
TTG	17687	17173	16463			
CTT	50950	26441	26527			
CTC	8502	13587	12485			
CTA	4956	7675	7564			
CTG	18854	15023	14911			

(j) Test dla aminokwasu L

	beg	mid	end	Stat	C	Pval
AAT	32386	21172	20991	83	6.07×10^{-4}	7.66×10^{-19}
AAC	28856	17221	16812			

(k) Test dla aminokwasu N

Tablica 4.3.2: Tablice kontyngencji dla bakterii muszki owocowej.

Z tabel 4.3.1 i 4.3.2 jednoznacznie wynika istnienie zależności między położeniem, a wystąpieniem kodonu. Współczynniki rozbieżności prawie zawsze są poniżej poziomu 0.02 (w kilku przypadkach poniżej 0.05) oraz wszystkie p-wartości są bliskie zera.

4.4 Macierze przejść

Poza zależnością wystąpień kodonów od położenia w sekwencji kodującej, sensownym jest pomyśleć o zależności wystąpienia kodonu od tego jaki poprzednio wystąpił kodon kodujący ten sam aminokwas. Na podstawie zliczeń można stworzyć estymatory częstościowe, za pomocą których otrzymane zostaną macierze przejść. W ten sposób można by modelować występowanie kodonów za pomocą łańcuchów Markova, których pewnym przybliżeniem były by macierze przejść otrzymane w sposób wyżej wymienony. Takie macierze mogłyby posłużyć do wyznaczenia stanów stacjonaranych łańcuchów, które nie będą tematem tej pracy.

4.4.1 Wyniki

Macierze przejść zostały przedstawione w tabelach 4.4.1 i 4.4.2. Z tabel tych wynika, że przejścia pomiędzy kodonami nie są równomiernie rozłożone, pewnie kodony przechodzą cześciej w inne. Niektóre przejścia są mało prawdopodobne – zwłaszcza w przypadkach gdzie aminowkas jest kodowany przez wiele kodonów. Wyniki te zgadzają z wynkiami z sekcji 4.1, kodony które dominują w pewnym aminokwasie mają większe prawdopodobieństwo przejścia do nich.

	GCT	GCC	GCA	GCG
GCT	0.188	0.269	0.230	0.313
GCC	0.156	0.302	0.207	0.335
GCA	0.176	0.254	0.242	0.328
GCG	0.146	0.257	0.210	0.387

(a) Macierz dla aminokwasu A

	TGT	TGC
TGT	0.475	0.525
TGC	0.420	0.580

(b) Macierz dla aminokwasu C

	GAT	GAC
GAT	0.643	0.357
GAC	0.606	0.394

(c) Macierz dla aminokwasu D

	GAA	GAG
GAA	0.679	0.321
GAG	0.660	0.340

(d) Macierz dla aminokwasu E

	TTT	TTC
TTT	0.607	0.393
TTC	0.530	0.470

(e) Macierz dla aminokwasu F

	GGT	GGC	GGA	GGG
GGT	0.359	0.378	0.114	0.150
GGC	0.330	0.413	0.104	0.152
GGA	0.311	0.335	0.166	0.188
GGG	0.297	0.371	0.135	0.197

(f) Macierz dla aminokwasu G

	CAT	CAC
CAT	0.604	0.396
CAC	0.537	0.463

(g) Macierz dla aminokwasu H

	ATT	ATC	ATA
ATT	0.523	0.391	0.086
ATC	0.479	0.453	0.068
ATA	0.482	0.298	0.220

(h) Macierz dla aminokwasu I

	AAA	AAG
AAA	0.759	0.241
AAG	0.731	0.269

(i) Macierz dla aminokwasu K

	TTA	TTG	CTT	CTC	CTA	CTG
TTA	0.181	0.139	0.129	0.099	0.047	0.405
TTG	0.143	0.145	0.110	0.096	0.040	0.466
CTT	0.148	0.129	0.131	0.107	0.043	0.441
CTC	0.127	0.117	0.112	0.108	0.035	0.500
CTA	0.177	0.140	0.126	0.100	0.051	0.405
CTG	0.106	0.114	0.097	0.100	0.031	0.551

(j) Macierz dla aminokwasu L

	AAT	AAC
AAT	0.523	0.477
AAC	0.413	0.587

(k) Macierz dla aminokwasu N

	CCT	CCC	CCA	CCG
CCT	0.191	0.142	0.208	0.459
CCC	0.190	0.169	0.187	0.454
CCA	0.175	0.126	0.216	0.483
CCG	0.141	0.107	0.173	0.579

(l) Macierz dla aminokwasu P

	CAA	CAG
CAA	0.377	0.623
CAG	0.302	0.698

(m) Macierz dla aminokwasu Q

	CGT	CGC	CGA	CGG	AGA	AGG
CGT	0.408	0.383	0.059	0.099	0.031	0.020
CGC	0.367	0.413	0.058	0.106	0.033	0.023
CGA	0.313	0.329	0.103	0.130	0.074	0.052
CGG	0.311	0.344	0.077	0.165	0.056	0.047
AGA	0.257	0.254	0.110	0.118	0.171	0.090
AGG	0.234	0.240	0.118	0.164	0.154	0.090

(n) Macierz dla aminokwasu R

	TCT	TCC	TCA	TCG	AGT	AGC
TCT	0.188	0.165	0.145	0.135	0.140	0.226
TCC	0.147	0.169	0.117	0.134	0.151	0.281
TCA	0.151	0.138	0.172	0.145	0.163	0.231
TCG	0.133	0.150	0.134	0.176	0.147	0.260
AGT	0.134	0.134	0.144	0.147	0.174	0.266
AGC	0.127	0.146	0.115	0.147	0.156	0.310

(o) Macierz dla aminokwasu S

	ACT	ACC	ACA	ACG
ACT	0.194	0.403	0.155	0.249
ACC	0.156	0.458	0.116	0.270
ACA	0.176	0.338	0.211	0.274
ACG	0.148	0.407	0.143	0.302

(p) Macierz dla aminokwasu T

	GTT	GTC	GTA	GTG
GTT	0.289	0.211	0.168	0.331
GTC	0.254	0.222	0.152	0.371
GTA	0.279	0.201	0.170	0.349
GTG	0.227	0.211	0.141	0.421

(q) Macierz dla aminokwasu V

	TAT	TAC
TAT	0.596	0.404
TAC	0.535	0.465

(r) Macierz dla aminokwasu Y

Tablica 4.4.1: Tablice przejść dla bakterii E.Coli.

	GCT	GCC	GCA	GCG
GCT	0.199	0.430	0.191	0.181
GCC	0.147	0.504	0.142	0.207
GCA	0.171	0.362	0.251	0.216
GCG	0.146	0.428	0.166	0.260

(a) Macierz dla aminokwasu A

	TGT	TGC
TGT	0.313	0.687
TGC	0.257	0.743

(b) Macierz dla aminokwasu C

	GAT	GAC
GAT	0.563	0.437
GAC	0.507	0.493

(c) Macierz dla aminokwasu D

	GAA	GAG
GAA	0.369	0.631
GAG	0.265	0.735

(d) Macierz dla aminokwasu E

	TTT	TTC
TTT	0.391	0.609
TTC	0.340	0.660

(e) Macierz dla aminokwasu F

	GGT	GGC	GGA	GGG
GGT	0.246	0.449	0.238	0.066
GGC	0.208	0.501	0.225	0.066
GGA	0.212	0.417	0.306	0.065
GGG	0.198	0.463	0.241	0.098

(f) Macierz dla aminokwasu G

	CAT	CAC
CAT	0.445	0.555
CAC	0.388	0.612

(g) Macierz dla aminokwasu H

	ATT	ATC	ATA
ATT	0.325	0.478	0.196
ATC	0.303	0.531	0.166
ATA	0.333	0.449	0.217

(h) Macierz dla aminokwasu I

	AAA	AAG
AAA	0.334	0.666
AAG	0.239	0.761

(i) Macierz dla aminokwasu K

	TTA	TTG	CTT	CTC	CTA	CTG
TTA	0.158	0.195	0.116	0.111	0.115	0.304
TTG	0.048	0.203	0.084	0.152	0.092	0.421
CTT	0.066	0.194	0.107	0.138	0.091	0.404
CTC	0.030	0.159	0.075	0.163	0.090	0.483
CTA	0.053	0.189	0.097	0.151	0.106	0.404
CTG	0.028	0.156	0.070	0.177	0.079	0.490

(j) Macierz dla aminokwasu L

	AAT	AAC
AAT	0.518	0.482
AAC	0.428	0.572

(k) Macierz dla aminokwasu N

	CCT	CCC	CCA	CCG
CCT	0.142	0.282	0.281	0.295
CCC	0.088	0.342	0.230	0.340
CCA	0.116	0.277	0.294	0.313
CCG	0.082	0.312	0.237	0.369

(l) Macierz dla aminokwasu P

	CAA	CAG
CAA	0.348	0.652
CAG	0.259	0.741

(m) Macierz dla aminokwasu Q

	CGT	CGC	CGA	CGG	AGA	AGG
CGT	0.193	0.370	0.151	0.136	0.065	0.085
CGC	0.161	0.391	0.136	0.156	0.064	0.091
CGA	0.177	0.328	0.163	0.146	0.085	0.100
CGG	0.150	0.351	0.156	0.175	0.069	0.099
AGA	0.158	0.263	0.166	0.132	0.146	0.135
AGG	0.155	0.309	0.153	0.150	0.092	0.141

(n) Macierz dla aminokwasu R

	TCT	TCC	TCA	TCG	AGT	AGC
TCT	0.112	0.212	0.131	0.215	0.128	0.202
TCC	0.063	0.241	0.087	0.241	0.132	0.237
TCA	0.097	0.218	0.132	0.232	0.120	0.201
TCG	0.058	0.220	0.084	0.262	0.126	0.249
AGT	0.074	0.213	0.099	0.208	0.157	0.249
AGC	0.049	0.202	0.072	0.198	0.141	0.337

(o) Macierz dla aminokwasu S

	ACT	ACC	ACA	ACG
ACT	0.175	0.355	0.216	0.255
ACC	0.125	0.428	0.170	0.277
ACA	0.151	0.318	0.274	0.256
ACG	0.113	0.371	0.179	0.337

(p) Macierz dla aminokwasu T

	GTT	GTC	GTA	GTG
GTT	0.209	0.230	0.120	0.442
GTC	0.166	0.262	0.092	0.480
GTA	0.203	0.237	0.138	0.422
GTG	0.151	0.241	0.089	0.519

(q) Macierz dla aminokwasu V

	TAT	TAC
TAT	0.421	0.579
TAC	0.373	0.627

(r) Macierz dla aminokwasu Y

Tablica 4.4.2: Tablice przejść dla bakterii muszki owocowej.

5 Wnioski

5.1 Czułość testu χ^2

Czułośc testu χ^2 jest największym problemem w testowaniu tego typu danych, gdzie obserwacji jest ok. 50-100 tysięcy. Powoduje to częste odrzucanie hipotez zerowych z niską p-wartością. Kontrolowanie go za pomocą prostego współczynnika rozbieżności, może wprowadzać w błąd. Różnice w tym współczynniku widać pomiedzy organizmami, ponieważ muszka owocowa ma znacznie dłuższe sekwencje kodujące. Do dokładniejszego kontrolowania tego testu można posłużyć się lepszymi, bardziej skomplikowanymi współczynnikami.

5.2 Równość prawdopodobieństw w występowaniu kodonów

Z samych liczności kodonów z tabel opisanych w sekcji 4.1.3 możemy już zauważyć, że większość kodonów nie występuje równomiernie. Wszystkie p-wartości są bliskie zeru, jednakże jest to spowodowane ogromnymi liczbami wystąpień i dużymi różnicami pomiędzy przewidywaną wartością, co daje duże wartości statystyki. Współczynnik rozbieżności w przypadku mniejszej ilości kodonów opisujących aminokwas jest zawsze mała, ze względu na duże liczby wystąpień (często rzędu 10^5). Gdy kodonów jest więcej, jak w przypadku aminokwasu R oraz nierównomiernie rozkładają się, różnice pomiędzy przewidywaną wartością są tak duże, że współczynnik rozbieżności staje się stosunkowo duży, co sugeruje nam nieistotność tego testu, gdzie tabela wskazuje na znaczącą nierówność prawdopodobieństw.

5.3 Rozkład geometryczny jako opis długości przerw między kodonami

Rozkład geometryczny wydaje się być naturalnym wyborem do opisu długości przerw między aminokwasami i kodonami kodującymi ten aminokwas, lecz być może przez zależności w położeniach kodonów test zgodności χ^2 zawsze go odrzuca. Niezależność kodonów od położenia jest zdecydowanie odrzucana przez test niezależności χ^2 zawsze z niskim współczynnikiem rozbieżności. Wykresy kwantylowe z sekcji 4.2.5 sugerują niedopasowanie tego rozkładu na końcach dziedziny, gdzie rozkład daje bardzo niskie prawdopodobieństwa, a występują tam wciąż obserwacje. Być może powinno się zastosować inne podziały półprostej dodaniej, z większą ilością przedziałów na końcach. Nie jest to możliwe w przypadku testu χ^2, który wymaga przybliżonych podobieństw na każdym przedziale. W tabelach z sekcji 4.2.7 widać pewne znaczące niedopasowania na niektórych przedziałach. Większość testów jest przeprowadzona na niskim poziomie współczynnika rozbieżności, co sugeruje poprawność odrzucania tego rozkładu.

5.4 Różnice w organizmach

Główną różnicą w testowaniu różnych organizmów jest długość sekwencji kodujących. Ze tego względu wyniki testów w przypadku muszki owocowej, zawsze dają p-wartości bliskie zeru, lecz niekoniecznie mniejsze współczynniki rozbieżności, co powoduje wątpliwość w wynikach testu tego organizmu.

A Testowanie równości prawdopodobieństw w występowaniu kodonów

A.1 E. Coli

AGC	AGT	TCA	TCC	TCG	TCT	L.oczek.	Statystyka	C	Pval
26113	15324	13344	14678	14306	14217	16330	7160	0.07	0.00

(a) Test dla aminokwasu S

ACA	ACC	ACG	ACT	L.oczek.	Statystyka	C	Pval
13316	37013	24680	14768	22444	16017	0.18	0.00

(b) Test dla aminokwasu T

GTA	GTC	GTG	GTT	L.oczek.	Statystyka	C	Pval
17837	24063	42594	29480	28494	11686	0.10	0.00

(c) Test dla aminokwasu V

TAC	TAT	L.oczek.	Statystyka	C	Pval
20007	26874	23441	1006	0.02	0.00

(d) Test dla aminokwasu Y

Tablica A.1.1: Wyniki testu zgodności dla E.Coli.

A.2 Muszka owocowa

AGC	AGT	TCA	TCC	TCG	TCT	L.oczek.	Statystyka	C	Pval
89704	46160	31285	74869	77947	22655	57103	66299	0.19	0.00

(a) Test dla aminokwasu S

ACA	ACC	ACG	ACT	L.oczekiwanych	Statystyka	C	Pval
45040	85263	64634	30002	56235	30705	0.14	0.00

(b) Test dla aminokwasu T

GTA	GTC	GTG	GTT	L.oczek.	Statystyka	C	Pval
21856	53386	106658	37200	54775	74602	0.34	0.00

(c) Test dla aminokwasu V

TAC	TAT	L.oczek.	Statystyka	C	Pval
64071	40870	52471	5129	0.05	0.00

(d) Test dla aminokwasu Y

Tablica A.2.1: Wyniki testu zgodności dla muszki owocowej.

B Testowanie zgodności z rozkładami

B.1 E. Coli

	1	2	3	4	5	Stat	C	Pval
Liczba GGT	7638	8603	7833	7567	7778			
L.Oczek. GGT	7807	7908	7863	7902	7937	82	2.09×10^{-3}	5.53×10^{-17}
Liczba GGC	8552	10253	9203	8922	8721			
L.Oczek. GGC	9055	9141	9013	9250	9189	202	4.44×10^{-3}	1.06×10^{-42}
Liczba GGA	3093	3126	2916	2592	2812			
L.Oczek. GGA	2905	2899	2908	2908	2917	68	4.68×10^{-3}	5.77×10^{-14}
Liczba GGG	3708	4117	4016	3842	3760			
L.Oczek. GGG	3826	3935	3893	3882	3903	21	1.11×10^{-3}	2.41×10^{-4}
Liczba G	17513	27539	26123	24987	22890			
L.Oczek. G	22409	24094	24745	23539	24288	1809	1.52×10^{-2}	0

(a) Test dla aminokwasu G

	1	2	3	4	5	Stat	C	Pval
Liczba CAT	4330	4252	4216	4076	3965			
L.Oczek. CAT	4086	4182	4209	4188	4172	28	1.39×10^{-3}	8.03×10^{-6}
Liczba CAC	3084	3129	3084	2916	2996			
L.Oczek. CAC	3037	3043	3032	3042	3051	10	6.73×10^{-4}	3.65×10^{-2}
Liczba H	7769	6959	7378	6888	7054			
L.Oczek. H	7243	7205	7349	7275	7270	71	1.98×10^{-3}	8.69×10^{-15}

(b) Test dla aminokwasu H

	1	2	3	4	5	Stat	C	Pval
Liczba ATT	9113	10082	10073	9719	9165			
L.Oczek. ATT	9322	9904	9466	9817	9641	71	1.48×10^{-3}	1.15×10^{-14}
Liczba ATC	7510	8315	7739	7604	7688			
L.Oczek. ATC	7613	7734	7871	7810	7826	55	1.42×10^{-3}	3.15×10^{-11}
Liczba ATA	2655	1920	1481	1250	1752			
L.Oczek. ATA	1794	1805	1824	1815	1817	663	7.32×10^{-2}	3.49×10^{-142}
Liczba I	16273	19399	22762	19149	18483			
L.Oczek. I	18316	19931	19139	19442	19258	963	1.00×10^{-2}	2.79×10^{-207}

(c) Test dla aminokwasu I

	1	2	3	4	5	Stat	C	Pval
Liczba AAA	11111	12287	10220	11051	10752			
L.Oczek. AAA	10682	11252	11108	11184	11192	202	3.65×10^{-3}	1.24×10^{-42}
Liczba AAG	4296	3608	3335	3403	3579			
L.Oczek. AAG	3570	3678	3659	3662	3649	197	1.08×10^{-2}	1.41×10^{-41}
Liczba K	15542	16057	13801	13607	14635			
L.Oczek. K	14042	15136	14868	14667	14966	376	5.11×10^{-3}	2.75×10^{-80}

(d) Test dla aminokwasu K

Tablica B.1.1: Wyniki testu zgodności dla bakterii Ecoli.

	1	2	3	4	5	Stat	C	Pval
Liczba TTA	5591	4647	4262	3705	4223			
L.Oczek. TTA	4453	4477	4482	4522	4491	471	2.10×10^{-2}	7.74×10^{-101}
Liczba TTG	4728	4325	3996	3925	4047			
L.Oczek. TTG	4198	4200	4199	4205	4217	106	5.04×10^{-3}	5.09×10^{-22}
Liczba CTT	3987	3899	3652	3480	3622			
L.Oczek. CTT	3664	3770	3731	3743	3730	56	3.02×10^{-3}	1.72×10^{-11}
Liczba CTC	3365	3492	3455	3406	3375			
L.Oczek. CTC	3414	3377	3431	3437	3432	6	3.51×10^{-4}	1.99×10^{-1}
Liczba CTA	1353	1223	1297	1196	1254			
L.Oczek. CTA	1257	1269	1261	1266	1268	14	2.23×10^{-3}	7.03×10^{-3}
Liczba CTG	13039	18659	18091	18032	15520			
L.Oczek. CTG	15600	17630	16374	16934	16801	829	9.95×10^{-3}	3.55×10^{-178}
Liczba L	18433	44829	38979	33110	33413			
L.Oczek. L	33146	34297	33464	33906	33956	10701	6.34×10^{-2}	0

(e) Test dla aminokwasu L

	1	2	3	4	5	Stat	C	Pval
Liczba ATG	11539	7568	8011	7850	9122			
L.Oczek. ATG	8772	8764	8779	8923	8850	1240	2.81×10^{-2}	2.45×10^{-267}
Liczba M	11621	7568	8011	7850	9122			
L.Oczek. M	8809	8796	8805	8938	8838	1282	2.90×10^{-2}	1.86×10^{-276}

(f) Test dla aminokwasu M

	1	2	3	4	5	Stat	C	Pval
Liczba AAT	7264	6410	5848	5469	5965			
L.Oczek. AAT	6146	6144	6192	6251	6220	342	1.11×10^{-2}	7.61×10^{-73}
Liczba AAC	7191	7168	6824	7037	7009			
L.Oczek. AAC	6877	7066	7135	7087	7062	30	8.55×10^{-4}	4.63×10^{-6}
Liczba N	12267	15276	12905	12354	13383			
L.Oczek. N	12513	13890	13236	13235	13377	210	3.17×10^{-3}	2.42×10^{-44}

(g) Test dla aminokwasu N

	1	2	3	4	5	Stat	C	Pval
Liczba CCT	2495	2417	2374	2242	2345			
L.Oczek. CCT	2356	2364	2387	2388	2375	18	1.58×10^{-3}	8.69×10^{-4}
Liczba CCC	1890	1946	1811	1712	1791			
L.Oczek. CCC	1806	1838	1835	1835	1832	19	2.16×10^{-3}	5.51×10^{-4}
Liczba CCA	2938	2755	2656	2647	2711			
L.Oczek. CCA	2703	2747	2752	2750	2752	28	2.07×10^{-3}	1.07×10^{-5}
Liczba CCG	6587	8346	7798	7015	7119			
L.Oczek. CCG	7273	7313	7523	7330	7424	246	6.69×10^{-3}	3.24×10^{-52}
Liczba P	13142	15131	14167	14713	14442			
L.Oczek. P	13557	14644	14426	14529	14516	36	5.06×10^{-4}	2.58×10^{-7}

(h) Test dla aminokwasu P

Tablica B.1.1: Wyniki testu zgodności dla bakterii Ecoli.

	1	2	3	4	5	Stat	C	Pval
Liczba CAA	5844	4493	4685	4294	4566			
L.Oczek. CAA	4739	4765	4772	4816	4788	341	1.43×10^{-2}	1.01×10^{-72}
Liczba CAG	9849	9779	9813	9328	9512			
L.Oczek. CAG	9321	9910	9702	9630	9716	46	9.66×10^{-4}	1.80×10^{-9}
Liczba Q	15646	15095	12994	13586	14842			
L.Oczek. Q	13678	14770	14545	14641	14608	535	7.42×10^{-3}	1.11×10^{-114}

(i) Test dla aminokwasu Q

	1	2	3	4	5	Stat	C	Pval
Liczba CGT	6978	6902	6444	6323	6553			
L.Oczek. CGT	6611	6650	6613	6647	6678	52	1.58×10^{-3}	1.17×10^{-10}
Liczba CGC	6789	7137	7217	6527	6600			
L.Oczek. CGC	6619	6999	6856	6912	6882	59	1.72×10^{-3}	4.50×10^{-12}
Liczba CGA	1442	1336	1230	1137	1269			
L.Oczek. CGA	1269	1290	1288	1282	1283	44	6.93×10^{-3}	5.24×10^{-9}
Liczba CGG	2269	2362	1944	1837	2019			
L.Oczek. CGG	2079	2091	2078	2087	2093	93	8.98×10^{-3}	2.14×10^{-19}
Liczba AGA	1313	1005	867	685	901			
L.Oczek. AGA	948	956	951	958	955	231	4.85×10^{-2}	6.52×10^{-49}
Liczba AGG	782	599	616	479	581			
L.Oczek. AGG	607	612	610	614	611	81	2.67×10^{-2}	7.38×10^{-17}
Liczba R	17454	21175	16764	18754	17996			
L.Oczek. R	18321	18150	18474	18464	18773	740	8.03×10^{-3}	6.31×10^{-159}

(j) Test dla aminokwasu R

	1	2	3	4	5	Stat	C	Pval
Liczba TCT	3211	2889	2644	2610	2863			
L.Oczek. TCT	2836	2832	2842	2857	2847	85	6.05×10^{-3}	9.57×10^{-18}
Liczba TCC	2792	3133	3004	2811	2938			
L.Oczek. TCC	2885	2984	2911	2948	2948	19	1.35×10^{-3}	5.60×10^{-4}
Liczba TCA	3201	2687	2477	2313	2666			
L.Oczek. TCA	2665	2640	2684	2681	2672	174	1.31×10^{-2}	8.88×10^{-37}
Liczba TCG	2831	2987	2778	2831	2879			
L.Oczek. TCG	2840	2866	2851	2879	2868	7	5.44×10^{-4}	9.97×10^{-2}
Liczba AGT	3220	3179	3073	2810	3042			
L.Oczek. AGT	3046	3069	3070	3069	3067	35	2.34×10^{-3}	3.13×10^{-7}
Liczba AGC	5060	5386	5442	5098	5127			
L.Oczek. AGC	5204	5151	5308	5191	5256	22	8.75×10^{-4}	1.35×10^{-4}
Liczba S	18519	19115	19560	20678	20110			
L.Oczek. S	18985	19293	20054	19754	19918	70	7.17×10^{-4}	1.97×10^{-14}

(k) Test dla aminokwasu S

	1	2	3	4	5	Stat	C	Pval
Liczba ACT	3231	2884	2826	2852	2975			
L.Oczek. ACT	2939	2936	2977	2957	2957	41	2.80×10^{-3}	2.32×10^{-8}
Liczba ACC	7132	7992	7094	7327	7468			
L.Oczek. ACC	7340	7369	7431	7418	7453	74	2.02×10^{-3}	2.14×10^{-15}
Liczba ACA	3393	2673	2415	2145	2690			
L.Oczek. ACA	2637	2674	2667	2669	2667	343	2.58×10^{-2}	3.66×10^{-73}
Liczba ACG	5023	4879	5171	4811	4796			
L.Oczek. ACG	4914	4936	4937	4941	4950	22	9.06×10^{-4}	1.71×10^{-4}
Liczba T	16560	19464	18345	16865	18543			
L.Oczek. T	17643	17534	18647	17680	18300	324	3.61×10^{-3}	5.46×10^{-69}

(l) Test dla aminokwasu T

Tablica B.1.1: Wyniki testu zgodności dla bakterii Ecoli.

	1	2	3	4	5	Stat	C	Pval
Liczba GTT	5839	6357	5936	5686	5662			
L.Oczek. GTT	5880	5796	5963	5943	5896	75	2.55×10^{-3}	1.83×10^{-15}
Liczba GTC	4484	5031	5040	4871	4637			
L.Oczek. GTC	4793	4787	4834	4817	4829	49	2.05×10^{-3}	4.97×10^{-10}
Liczba GTA	3523	3658	3672	3447	3537			
L.Oczek. GTA	3497	3584	3593	3584	3575	9	5.10×10^{-4}	5.86×10^{-2}
Liczba GTG	8384	9326	8443	8010	8431			
L.Oczek. GTG	8251	8570	8670	8497	8603	106	2.49×10^{-3}	4.93×10^{-22}
Liczba V	17201	25451	22762	26145	22415			
L.Oczek. V	20570	24109	22757	23354	23197	986	8.65×10^{-3}	3.16×10^{-212}

(m) Test dla aminokwasu V

	1	2	3	4	5	Stat	C	Pval
Liczba TGG	5090	5879	4841	4648	4792			
L.Oczek. TGG	5019	5074	5026	5057	5071	184	7.29×10^{-3}	9.95×10^{-39}
Liczba W	5090	5585	4770	4839	4966			
L.Oczek. W	5120	5216	5120	5225	5197	75	2.92×10^{-3}	1.46×10^{-15}

(n) Test dla aminokwasu W

	1	2	3	4	5	Stat	C	Pval
Liczba TAT	5338	5633	5448	5107	5348			
L.Oczek. TAT	5364	5338	5369	5414	5385	35	1.31×10^{-3}	4.12×10^{-7}
Liczba TAC	3942	4264	3944	3911	3946			
L.Oczek. TAC	3929	4065	3989	4019	4002	13	6.96×10^{-4}	7.54×10^{-3}
Liczba Y	9084	10100	9369	9116	9212			
L.Oczek. Y	9373	9453	9309	9536	9424	76	1.62×10^{-3}	1.16×10^{-15}

(o) Test dla aminokwasu Y

Tablica B.1.1: Wyniki testu zgodności dla bakterii Ecoli.

B.2 Muszka owocowa

	1	2	3	4	5	Stat	C	Pval
Liczba GGT	16013	10608	9129	8872	11590			
L.Oczek. GGT	11044	11426	11210	11228	11302	3182	5.66×10^{-2}	0
Liczba GGC	31236	23633	22064	21132	23661			
L.Oczek. GGC	23445	24943	24440	24294	24602	3336	2.74×10^{-2}	0
Liczba GGA	20338	11340	10387	9832	12959			
L.Oczek. GGA	12952	12927	12875	13058	13041	5685	8.77×10^{-2}	0
Liczba GGG	4624	3627	3063	2918	3631			
L.Oczek. GGG	3552	3569	3587	3577	3575	523	2.93×10^{-2}	5.61×10^{-112}
Liczba G	63797	55328	46585	42517	52430			
L.Oczek. G	49860	51970	52472	53947	52410	7194	2.76×10^{-2}	0

(a) Test dla aminokwasu G

	1	2	3	4	5	Stat	C	Pval
Liczba CAT	12392	7671	7178	7423	8928			
L.Oczek. CAT	8588	8839	8678	8750	8735	2304	5.29×10^{-2}	0
Liczba CAC	16313	11933	11211	10913	12358			
L.Oczek. CAC	12322	12749	12438	12601	12616	1697	2.71×10^{-2}	0
Liczba H	30756	17694	18768	17514	21588			
L.Oczek. H	20897	21413	21455	21170	21494	6271	5.89×10^{-2}	0

(b) Test dla aminokwasu H

	1	2	3	4	5	Stat	C	Pval
Liczba ATT	11978	11579	11023	9849	10793			
L.Oczek. ATT	10995	11034	11079	11056	11056	253	4.59×10^{-3}	1.32×10^{-53}
Liczba ATC	18755	19987	16489	15808	16327			
L.Oczek. ATC	17340	17460	17383	17650	17531	801	9.18×10^{-3}	2.88×10^{-172}
Liczba ATA	7690	6831	6189	5432	6399			
L.Oczek. ATA	6482	6474	6528	6537	6517	451	1.39×10^{-2}	2.27×10^{-96}
Liczba I	36054	38097	33614	34417	32947			
L.Oczek. I	34896	34851	34519	35621	35252	555	3.17×10^{-3}	5.69×10^{-119}

(c) Test dla aminokwasu I

	1	2	3	4	5	Stat	C	Pval
Liczba AAA	14104	10594	8804	7849	10057			
L.Oczek. AAA	10142	10319	10370	10262	10313	2365	4.60×10^{-2}	0
Liczba AAG	36023	29194	26799	22667	26780			
L.Oczek. AAG	27946	28135	28225	28747	28407	3825	2.70×10^{-2}	0
Liczba K	40718	49865	35695	29911	36682			
L.Oczek. K	36473	40183	38453	39101	38672	5287	2.74×10^{-2}	0

(d) Test dla aminokwasu K

Tablica B.2.1: Wyniki testu zgodności dla muszki owocowej.

	1	2	3	4	5	Stat	C	Pval
Liczba TTA	4659	2658	2184	2008	2794			
L.Oczek. TTA	2840	2868	2869	2859	2865	1598	1.12×10^{-1}	0
Liczba TTG	12423	12478	10690	10475	11157			
L.Oczek. TTG	11359	11524	11386	11465	11486	316	5.52×10^{-3}	3.62×10^{-67}
Liczba CTT	6423	5708	4817	4338	5361			
L.Oczek. CTT	5300	5323	5337	5352	5333	508	1.91×10^{-2}	8.38×10^{-109}
Liczba CTC	11480	11882	10366	10047	9898			
L.Oczek. CTC	10596	10779	10717	10825	10753	321	5.99×10^{-3}	2.18×10^{-68}
Liczba CTA	6724	5908	5582	5209	5837			
L.Oczek. CTA	5821	5872	5857	5841	5867	221	7.57×10^{-3}	8.45×10^{-47}
Liczba CTG	33591	30947	31272	27218	27533			
L.Oczek. CTG	29913	29177	31086	30161	30221	1087	7.22×10^{-3}	4.66×10^{-234}
Liczba L	60323	74067	64120	69345	63797			
L.Oczek. L	62572	69478	64229	67269	68109	721	2.17×10^{-3}	8.87×10^{-155}

(e) Test dla aminokwasu L

	1	2	3	4	5	Stat	C	Pval
Liczba ATG	23463	17183	15175	14256	17622			
L.Oczek. ATG	17109	17883	17399	17647	17658	3322	3.79×10^{-2}	0
Liczba M	23478	17183	15175	14256	17622			
L.Oczek. M	17115	17888	17403	17649	17656	3331	3.80×10^{-2}	0

(f) Test dla aminokwasu M

	1	2	3	4	5	Stat	C	Pval
Liczba AAT	18356	18049	16309	15251	16593			
L.Oczek. AAT	16490	17096	17136	16880	16953	468	5.55×10^{-3}	3.49×10^{-100}
Liczba AAC	25636	16997	17606	16215	18984			
L.Oczek. AAC	18997	18904	19291	19110	19134	3099	3.25×10^{-2}	0
Liczba N	42417	36818	31854	32497	36410			
L.Oczek. N	34381	37037	36351	35816	36424	2743	1.52×10^{-2}	0

(g) Test dla aminokwasu N

	1	2	3	4	5	Stat	C	Pval
Liczba CCT	6233	3945	3339	3452	4246			
L.Oczek. CCT	4231	4234	4243	4257	4248	1311	6.18×10^{-2}	9.15×10^{-283}
Liczba CCC	15135	13875	12568	12463	13049			
L.Oczek. CCC	13131	13514	13603	13403	13436	471	7.03×10^{-3}	1.07×10^{-100}
Liczba CCA	17005	9582	8591	8248	11324			
L.Oczek. CCA	10946	10843	11022	10950	10986	4713	8.61×10^{-2}	0
Liczba CCG	19344	14128	12828	12403	14655			
L.Oczek. CCG	14406	14803	14614	14841	14691	2341	3.19×10^{-2}	0
Liczba P	53682	48174	35992	33234	45331			
L.Oczek. P	42059	44502	42193	43592	44086	6923	3.20×10^{-2}	0

(h) Test dla aminokwasu P

Tablica B.2.1: Wyniki testu zgodności dla muszki owocowej.

	1	2	3	4	5	Stat	C	Pval
Liczba CAA	20414	10814	9576	9209	12375			
L.Oczek. CAA	12318	12570	12418	12595	12484	7127	1.14×10^{-1}	0
Liczba CAG	47645	29115	22613	25032	31879			
L.Oczek. CAG	30287	31456	31302	31951	31286	14043	8.99×10^{-2}	0
Liczba Q	69467	38591	33305	31318	45991			
L.Oczek. Q	42899	42654	45387	43951	43813	23800	1.09×10^{-1}	0

(i) Test dla aminokwasu Q

	1	2	3	4	5	Stat	C	Pval
Liczba CGT	8865	6901	6347	6624	7212			
L.Oczek. CGT	7177	7175	7176	7220	7199	552	1.54×10^{-2}	3.42×10^{-118}
Liczba CGC	17126	17355	14288	13411	14251			
L.Oczek. CGC	14934	15475	15351	15373	15294	945	1.24×10^{-2}	2.59×10^{-203}
Liczba CGA	7600	6263	5970	6029	6483			
L.Oczek. CGA	6433	6476	6478	6477	6478	289	8.94×10^{-3}	2.18×10^{-61}
Liczba CGG	8388	6327	6087	5383	6597			
L.Oczek. CGG	6547	6514	6597	6558	6564	773	2.36×10^{-2}	4.28×10^{-166}
Liczba AGA	4736	3210	2788	2640	3437			
L.Oczek. AGA	3336	3372	3368	3368	3364	853	5.08×10^{-2}	1.65×10^{-183}
Liczba AGG	5520	4020	4080	3920	4325			
L.Oczek. AGG	4370	4347	4386	4384	4376	397	1.82×10^{-2}	7.58×10^{-85}
Liczba R	49771	48194	39822	37388	41008			
L.Oczek. R	41899	44367	42109	44432	43397	3181	1.47×10^{-2}	0

(j) Test dla aminokwasu R

	1	2	3	4	5	Stat	C	Pval
Liczba TCT	6573	4082	3677	3797	4526			
L.Oczek. TCT	4511	4538	4537	4535	4531	1271	5.61×10^{-2}	5.86×10^{-274}
Liczba TCC	19401	14063	13457	12439	15509			
L.Oczek. TCC	14865	14907	15077	15000	15017	2059	2.75×10^{-2}	0
Liczba TCA	8905	6050	4908	5054	6368			
L.Oczek. TCA	6200	6267	6287	6271	6258	1728	5.52×10^{-2}	0.00×10^{0}
Liczba TCG	18832	15632	13887	13706	15890			
L.Oczek. TCG	15562	15334	15772	15595	15681	1149	1.48×10^{-2}	1.23×10^{-247}
Liczba AGT	11730	9523	8084	7507	9316			
L.Oczek. AGT	9119	9243	9288	9260	9247	1244	2.70×10^{-2}	4.27×10^{-268}
Liczba AGC	26398	15939	14491	14258	18618			
L.Oczek. AGC	17838	17627	18218	17966	18052	5814	6.48×10^{-2}	0
Liczba S	86020	71700	56716	60569	67615			
L.Oczek. S	66513	66908	69086	71170	68943	9883	2.88×10^{-2}	0

(k) Test dla aminokwasu S

	1	2	3	4	5	Stat	C	Pval
Liczba ACT	8256	5681	5166	4911	5988			
L.Oczek. ACT	5993	5973	6016	6002	6017	1186	3.96×10^{-2}	1.26×10^{-255}
Liczba ACC	20682	16508	15302	15672	17099			
L.Oczek. ACC	16627	17238	17278	17021	17096	1353	1.59×10^{-2}	1.06×10^{-291}
Liczba ACA	14271	7698	6913	6709	9449			
L.Oczek. ACA	8974	9035	8945	9059	9025	4415	9.80×10^{-2}	0
Liczba ACG	14905	12828	12752	11372	12777			
L.Oczek. ACG	12736	13003	13019	12893	12981	559	8.66×10^{-3}	7.15×10^{-120}
Liczba T	53759	46966	36336	42719	45159			
L.Oczek. T	41418	48278	45222	44960	45078	5571	2.48×10^{-2}	0

(l) Test dla aminokwasu T

Tablica B.2.1: Wyniki testu zgodności dla muszki owocowej.

	1	2	3	4	5	Stat	C	Pval
Liczba GTT	8465	7479	7044	6688	7524			
L.Oczek. GTT	7349	7506	7448	7433	7461	266	7.16×10^{-3}	1.79×10^{-56}
Liczba GTC	11712	10932	10486	9953	10303			
L.Oczek. GTC	10635	10574	10783	10687	10705	194	3.65×10^{-3}	4.72×10^{-41}
Liczba GTA	4777	4628	4320	3683	4448			
L.Oczek. GTA	4338	4375	4393	4368	4380	168	7.72×10^{-3}	1.96×10^{-35}
Liczba GTG	24474	21115	21368	18871	20830			
L.Oczek. GTG	20937	21454	21499	21430	21336	921	8.64×10^{-3}	4.34×10^{-198}
Liczba V	42607	48718	45270	38841	43664			
L.Oczek. V	42867	42652	45430	44060	44098	1487	6.79×10^{-3}	8.09×10^{-321}

(m) Test dla aminokwasu V

	1	2	3	4	5	Stat	C	Pval
Liczba TGG	8826	7692	6607	5975	5892			
L.Oczek. TGG	6905	7029	7020	7021	7015	956	2.73×10^{-2}	8.98×10^{-206}
Liczba W	8826	7395	6679	6054	6038			
L.Oczek. W	7091	7029	7149	7097	7113	794	2.24×10^{-2}	1.21×10^{-170}

(n) Test dla aminokwasu W

	1	2	3	4	5	Stat	C	Pval
Liczba TAT	9865	8670	7720	6940	7675			
L.Oczek. TAT	8050	8288	8169	8170	8191	669	1.64×10^{-2}	1.61×10^{-143}
Liczba TAC	15115	14522	12017	10789	11628			
L.Oczek. TAC	12760	12677	12921	12827	12884	1212	1.89×10^{-2}	2.80×10^{-261}
Liczba Y	24814	23671	19063	18242	19151			
L.Oczek. Y	20379	21570	20876	21155	21018	1895	1.80×10^{-2}	0

(o) Test dla aminokwasu Y

Tablica B.2.1: Wyniki testu zgodności dla muszki owocowej.

C Testowanie niezależności

C.1 E. Coli

	beg	mid	end	Stat	C	Pval
CCT	4369	7261	7091	656	8.21×10^{-3}	1.57×10^{-138}
CCC	3610	4560	4167			
CCA	11099	11282	10584			
CCG	4374	5957	5571			

(a) Test dla aminokwasu P

	beg	mid	end	Stat	C	Pval
CAA	8273	8453	8400	483	7.74×10^{-3}	8.41×10^{-106}
CAG	15562	10849	10965			

(b) Test dla aminokwasu Q

	beg	mid	end	Stat	C	Pval
CGT	1656	6300	6807	10718	7.76×10^{-2}	0
CGC	956	5773	6170			
CGA	2121	9070	9102			
CGG	11153	12955	12716			
AGA	3292	10783	10572			
AGG	10646	9093	8933			

(c) Test dla aminokwasu R

	beg	mid	end	Stat	C	Pval
TCT	8534	9203	9109	1009	8.55×10^{-3}	1.63×10^{-210}
TCC	5356	5140	5511			
TCA	4775	7905	7757			
TCG	5313	5714	5271			
AGT	4707	7783	7530			
AGC	5162	6797	6448			

(d) Test dla aminokwasu S

	beg	mid	end	Stat	C	Pval
ACT	4823	6483	6821	1634	1.82×10^{-2}	0
ACC	12409	8493	8043			
ACA	8224	8576	8626			
ACG	5193	6071	6095			

(e) Test dla aminokwasu T

	beg	mid	end	Stat	C	Pval
GTT	6015	5791	6363	373	3.72×10^{-3}	1.53×10^{-77}
GTC	8419	6506	6307			
GTA	13271	9662	9681			
GTG	10313	9050	9055			

(f) Test dla aminokwasu V

	beg	mid	end	Stat	C	Pval
TAT	6302	6423	6273	10	2.47×10^{-4}	4.30×10^{-3}
TAC	8711	8249	8184			

(g) Test dla aminokwasu Y

Tablica C.1.1: Tablice kontyngencji dla bakterii E.Coli.

C.2 Muszka owocowa

	beg	mid	end	Stat	C	Pval
CCT	18139	31688	32031	9274	3.76×10^{-2}	0
CCC	22478	15595	15988			
CCA	24302	22625	24300			
CCG	7116	16413	16103			

(a) Test dla aminokwasu P

	beg	mid	end	Stat	C	Pval
CAA	21544	31953	30414	7786	3.91×10^{-2}	0
CAG	51750	31419	31962			

(b) Test dla aminokwasu Q

	beg	mid	end	Stat	C	Pval
CGT	6025	18971	18977	12131	3.60×10^{-2}	0
CGC	6870	21235	23029			
CGA	10625	27631	26250			
CGG	25450	25630	26434			
AGA	11447	23880	25023			
AGG	11526	14168	14242			

(c) Test dla aminokwasu R

	beg	mid	end	Stat	C	Pval
TCT	31084	30559	32857	8403	2.29×10^{-2}	0
TCC	15495	15600	15772			
TCA	10892	20542	20303			
TCG	26259	21441	21050			
AGT	26560	20864	21321			
AGC	7116	15160	14814			

(d) Test dla aminokwasu S

	beg	mid	end	Stat	C	Pval
ACT	16117	22037	23013	5390	2.30×10^{-2}	0
ACC	29298	20343	20629			
ACA	21854	20017	20953			
ACG	9399	15356	15281			

(e) Test dla aminokwasu T

	beg	mid	end	Stat	C	Pval
GTT	6948	8199	8470	2667	1.40×10^{-2}	0
GTC	17947	16374	16205			
GTA	35002	21684	22490			
GTG	12045	12790	12153			

(f) Test dla aminokwasu V

	beg	mid	end	Stat	C	Pval
TAT	19979	13355	13024	199	2.45×10^{-3}	6.04×10^{-44}
TAC	13302	11004	10524			

(g) Test dla aminokwasu Y

Tablica C.2.1: Tablice kontyngencji dla bakterii muszki owocowej.

D Plik z kodem

```
#Kod genetyczny
genomeEcoli = readDNAStringSet("GCF_000008865.2_ASM886v2_cds_
    from_genomic.fna", format = "fasta")
genomeDrosChrX = readDNAStringSet("Drosophila melanogaster chrX
    cds.fna", format = "fasta")
genomeDrosChrX = deleteSeq(genomeDrosChrX)
#Podzielenie DNA na 3 czesci: poczatek, srodek i koniec
divideDNA3 = function(DNA){
  width = width(DNA)
  beg = DNAStringSet()
  mid = DNAStringSet()
  end = DNAStringSet()

  for (i in 1:length(DNA)) {
    n = width[i]/3
    beg = c(beg, subseq(DNA[i], 1, n))
    mid = c(mid, subseq(DNA[i], n+1, 2*n))
    end = c(end, subseq(DNA[i], 2*n+1, 3*n))
  }

  return(list(beg, mid, end))
}
#Szukanie kodonow odpowiadajacych aminokwasowi
whichCodon = function(AA){
  names(GENETIC_CODE[which(GENETIC_CODE == AA)])
}
#Czestosc kodonow kodujaca jeden aminokwas
codonFreq = function(DNA, AA){
  freq = trinucleotideFrequency(DNA, step=3)
  indicies = which( colnames(freq) %in% whichCodon(AA) )
  return(freq[,indicies])
}
#Usuwanie niedostepnych sekwencji
deleteSeq <- function(DNA) {
  DNA[width(DNA)>9]
}
#Testowanie niezaleznosci kodonu od polozenia
testCodonInd <- function(DNA, AA) {
  contingency = data.frame()
  dividedDNA = divideDNA3(DNA)

  for (DNA in dividedDNA) {
    sumFreq = apply(codonFreq(DNA, AA), MARGIN=2, sum)
    contingency = rbind(contingency, sumFreq)
  }
  colnames(contingency) = whichCodon(AA)
```

```
45    rownames(contingency) = c("beg", "mid", "end")
      contingency = t(contingency)
47    test = chisq.test(contingency)
      pval = test$p.val
49    stat = unname(test$statistic)
      c = discrepancyCoeff(stat, sum(contingency))
51    nr = nrow(contingency)
      contingency = cbind(contingency, c(stat,rep(NA, nr-1)), c(c,
         rep(NA, nr-1)), c(pval,rep(NA, nr-1)))
53    nc = ncol(contingency)
      colnames(contingency)[(nc-2):nc] = c("Stat", "C", "Pval")
55    return(contingency)
    }
57  #Testowanie zgodnosci AA
    testCodonFreq <- function(DNA, AA) {
59    contingency = apply(codonFreq(DNA, AA), MARGIN=2, sum)
      expected = sum(contingency)/length(contingency)
61    test = chisq.test(contingency)
      stat = test$statistic
63    pval = test$p.value
      c = discrepancyCoeff(stat, sum(contingency))
65    data = t(c(contingency, L.oczekiwanych=expected, Stat=unname(
         stat), C=c, Pval=pval))
      return(as.data.frame(data))
67  }
    #Wystepowanie aminokwasow
69  AAPos <- function(dna, AA) {
      greg = gregexpr(AA,translate(dna))[[1]]
71    3*greg[1:length(greg)]-2
    }
73  #Wystepoiwanie kodonow w sekwencji kodujacej
    codonPos <- function(dna, AA) {
75    cod = whichCodon(AA)
      listPos = list()
77    pos = codons(dna)

79    for (c in cod) {
        listPos[[c]] = start(pos[as.data.frame(pos)[,1]==c,])
81    }
      return(listPos)
83  }
    #Zamiana pozycji na przerwy
85  breaks <- function(X) {
      n = length(X)
87    return( X - c(0,X[1:n-1]) )
    }
89  #Przerwy w wystepowaniu aminokwasow w sekwencji kodujacej
    AABreaks <- function(dna, AA)
```

```
91    breaks(AAPos(dna, AA))
  #Przerwy w wystepowaniu kodonow w sekwencji kodujacej
93  codonBreaks <- function(dna, AA)
      lapply(codonPos(dna, AA), breaks)
95  #Przerwy w wystepowaniu aminokwasow w calym genomie
    genomeAABreaks <- function(DNA, AA) {
97    breaks = c()
      for (i in 1:length(DNA)) {
99      breaks = c(breaks, AABreaks(DNA[[i]],AA))
      }
101   return(breaks)
    }
103 #Przerwy w wystepowaniu kodonow w calym genomie
    genomeCodonBreaks <- function(DNA, AA) {
105   codons = whichCodon(AA)
      listPos = list()
107   for (i in 1:length(DNA)) {
        listBreaks = codonBreaks(DNA[[i]], AA)
109     for (c in codons) {
          listPos[[c]] = c(listPos[[c]], listBreaks[[c]])
111     }
      }
113   return(listPos)
    }
115 #Przedzialy do rownego rozdzielenia rozkladu
    quantIntervals <- function(qdist,theta,n) {
117   p = 1/n
      rev(qdist(1:n*p, theta, lower.tail = FALSE))
119 }
    #Testowanie rozkladu geometrycznego
121 testCodonDist <- function(DNA, AA, n=5, qdist=qgeom, MLE=
      function(X)1/mean(X)) {
      codonBr = genomeCodonBreaks(DNA,AA)
123   AABr = list(genomeAABreaks(DNA, AA))
      names(AABr) = AA
125   data = data.frame()

127   for (cod in c(whichCodon(AA), AA)) {
        breaks = append(codonBr, AABr)[[cod]]
129     mle = MLE(breaks)
        int = quantIntervals(qdist, mle, n)
131
        count = c()
133     for (i in 1:(length(int)-1))
          count[i] = sum( int[i]<=breaks & breaks<int[i+1])
135     count[n] = sum(int[n]<=breaks)
```

```
137       prob = c(pgeom(int[-1]-1,mle) - pgeom(int[-n]-1,mle),pgeom(
              int[n]-1,mle, lower.tail = FALSE))
          test = chisq.test(count,p=prob)
139       pval= test$p.value
          stat = unname(test$statistic)
141       c = discrepancyCoeff(stat, length(breaks))

143       dataCod = data.frame(c(count,NA,NA,NA), c(prob*length(breaks
              ), stat, c, pval))
          colnames(dataCod) = c(paste("Liczba", cod), paste("L.Oczek."
              , cod))
145       data = rbind(data, t(dataCod))
      }
147   colnames(data) = c(paste(1:(ncol(data)-3)), "Stat", "C", "Pval
          ")
      return(data)
149 }
    #Tablica przejsc z kodonow
151 transitionMatrix <- function(dna, AA) {
      listPos = codonPos(dna, AA)
153   k = length(whichCodon(AA))
      transMat = matrix(0,k,k)
155   posVec = NULL
      for (i in 1:k)
157     posVec[listPos[[i]]] = i
      posVec = posVec[!is.na(posVec)]
159   for (i in 1:k) {
        nextCod = c(posVec[which(posVec==i)+1], 1:k)
161     row = table(nextCod)-1
        transMat[i,] = row
163   }
      rownames(transMat) = whichCodon(AA)
165   colnames(transMat) = whichCodon(AA)
      return(transMat)
167 }
    #Tablica przejsc z kodonow dla genomu
169 transitionGenome <- function(DNA, AA) {
      k = length(whichCodon(AA))
171   transMat = matrix(0,k,k)
      for (i in 1:length(DNA)) {
173     transMat = transMat + transitionMatrix(DNA[[i]],AA)
      }
175   transMat = diag(1/rowSums(transMat)) %*% transMat
      rownames(transMat) =whichCodon(AA)
177   return(transMat)
    }
179 #Wspolczynnik rozbieznosci
    discrepancyCoeff <- function(chi, N) unname(chi/N)
```

Literatura

[1] D. Chen and D. E. Texada. Low-usage codons and rare codons of escherichia coli. *Gene Ther. Mol. Biol*, 10:1–12, 2006.

[2] N. Cressie and T. R. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464, 1984.

[3] B. B. Khomtchouk, C. Wahlestedt, and W. Nonner. A global perspective of codon usage. *BioRxiv*, 2016.

[4] K. Kirilov and I. Ivanov. A programme for determination of codons and codons context frequency of occurrence in sequenced genomes. *Biotechnology & Biotechnological Equipment*, 26(5):3310–3314, 2012.

[5] J. Koronacki and J. Mielniczuk. *Statystyka: dla studentów kierunków technicznych i przyrodniczych*. Wydawnictwa Naukowo-Techniczne, 2009.

[6] H. Mirsafian, A. Mat Ripen, A. Singh, P. H. Teo, A. F. Merican, and S. B. Mohamad. A comparative analysis of synonymous codon usage bias pattern in human albumin superfamily. *The Scientific World Journal*, 2014, 2014.

[7] D. S. Moore. Measures of lack of fit from tests of chi-squared type. *Journal of statistical planning and inference*, 10(2):151–166, 1984.

[8] Y. Nakamura and T. Ikemura. Fop (frequency of optimal codon usage): Www service with its distribution analysis. *Genome Informatics*, 6:166–167, 1995.

[9] G. A. Palidwor, T. J. Perkins, and X. Xia. A general model of codon bias due to gc mutational bias. *PLoS One*, 5(10), 2010.

[10] T. E. F. Quax, N. J. Claassens, D. Söll, and J. van der Oost. Codon bias as a means to fine-tune gene expression. *Molecular cell*, 59(2):149–161, 2015.

[11] H. M. Salim and A. R. Cavalcanti. Factors influencing codon usage bias in genomes. *Journal of the Brazilian Chemical Society*, 19(2):257–262, 2008.

[12] A. Uddin. Indices of codon usage bias. *Proteom. Bioinform*, 10(6), 2017.

Spis tablic

Spis rysunków