Matematički fakultet / Univerzitet u Beogradu

SEMINARSKI RAD

Predmet: Istraživanje podataka – napredni koncepti
predmetni nastavnik: prof dr Nenad Mitić

ANALIZA UPOTREBE KODONA U ZAVISNOSTI OD POZICIJE U OKVIRU GENA

Aleksandar KARTELJ
Matematički fakultet,
Univerzitet u Beogradu, Srbija
kartelj@matf.bg.ac.rs

U Beogradu, 27.9.2011

SADRŽAJ

O DOKUMENTU

U ovom dokumentu opisana je analiza upotrebe kodona u zavisnosti od pozicije na kojoj se taj kodon nalazi u okviru gena. Analiza je sprovedena na genomu jednoćelijskog organizma, bakteriji pod nazivom Escherichia coli. Ovaj organizam je dosta izučavan, i postoji veliki broj mapiranih genoma njegovih različitih varijanti. Korišćena je javno dostupna baza podataka genoma Nacionalnog Centra za Biotehnološke Informacije (eng. National Centre for Biotechnology Information), i u većem delu teksta ću je refereisati skraćeno, sa NCBI baza.

Dokument ima 3 poglavlja, a najpre će biti izložena preciznija formulacija problema i prethodni rezultati. U drugom poglavlju opisaću korišćene resurse, primenjene metode i još neke tehničke aspekte istraživanja, a u trećem će biti izloženi rezultati i zaključci. Na kraju dokumenta je spisak korišćene literature.

UVOD

Amino kiseline su strukturalne jedinice koje izgrađuju proteine. Stapanjem amino kiselina nastaju kratki lanci polimera, koji se nazivaju peptidi, ili duži lanci koji se nazivaju polipeptidi ili proteini. Amino kiseline se obično dele na dve grupe: esencijalne, tj. one koje ljudski organizam nije u stanju da sintetiše; i neesencijalne, koje mogu biti sintetisane u ljudskom organizmu. Postoje 22 amino kiseline, i 20 od njih (osim selencocistina i pirolizina) su kodirane univerzalnim genetskim kodom, dok su preostale 2 kodirane na nešto drugačiji način. Genetski kod je zapisan na mRNA, koja predstavlja kopiju RNA nekog od gena organizma. Prilikom procesa kreiranja proteina, koji se naziva translacija, lanci proteina se grade dodavanjem amino kiselina po redosledu koji kodira mRNA. Jedna amino kiselina kodirana je tripletom nukleotida, a budući da postoje 4 različita nukleotida (Adenin, Citozin, Timin i Guanin), broj mogućih varijacija sa ponavljanjem je $4^3 = 64$. Većina amino kiselina ima višeznačnu reprezanciju, tj. kodirana je sa više različitih kodona. Ovi kodoni se obično razlikuju u jednom nukleotidu na poslednjoj, trećoj poziciji. U Tabeli 1 prikazan je genetski kod (translaciona tablica) koji je specifičan za bakterije i još neke vrste organizama.

Tabela 1: Genetski kod bakterija i prokariotskih virusa (translaciona tablica br. 11, NCBI)

Kod.	Ozn.	Naz.	Kod.	Ozn.	Naz.	Kod.	Ozn.	Naz.	Kod.	Ozn.	Naz.
TTT	F	Phe	TCT	S	Ser	TAT	Y	Tyr	TGT	C	Cys
TTC	F	Phe	TCC	S	Ser	TAC	Y	Tyr	TGC	C	Cys
TTA	L	Leu	TCA	S	Ser	TAA	*	Ter	TGA	*	Ter
TTG	L	Leu	TCG	S	Ser	TAG	*	Ter	TGG	W	Trp
CTT	L	Leu	CCT	P	Pro	CAT	H	His	CGT	R	Arg
CTC	L	Leu	CCC	P	Pro	CAC	H	His	CGC	R	Arg
CTA	L	Leu	CCA	P	Pro	CAA	Q	Gln	CGA	R	Arg
CTG	L	Leu	CCG	P	Pro	CAG	Q	Gln	CGG	R	Arg
ATT	I	Ile	ACT	T	Thr	AAT	N	Asn	AGT	S	Ser
ATC	I	Ile	ACC	T	Thr	AAC	N	Asn	AGC	S	Ser
ATA	I	Ile	ACA	T	Thr	AAA	K	Lys	AGA	R	Arg
ATG	M	Met	ACG	T	Thr	AAG	K	Lys	AGG	R	Arg
GTT	V	Val	GCT	A	Ala	GAT	D	Asp	GGT	G	Gly
GTC	V	Val	GCC	A	Ala	GAC	D	Asp	GGC	G	Gly
GTA	V	Val	GCA	A	Ala	GAA	E	Glu	GGA	G	Gly
GTG	V	Val	GCG	A	Ala	GAG	E	Glu	GGG	G	Gly

U daljem tekstu, različite kodone koji kodiraju istu aminokiselinu, nazivaću kodonskim sinonimima. Upotreba kodonskih sinonima je retko ravnomerna, i obično varira već na nivou vrste organizma (eng. interspecies usage), a potom i u okviru samog genoma (eng. intragenomic usage), tj. među pojedinačnim genima [Sha88]. Postoje tri glavna faktora, koja utiču na upotrebu u okviru samog genoma: odnos količina G i C nukleotida je najčešće povezan sa postojanjem genomskih ostrva, koja mogu biti rezultat horiznotalnog DNK transfera [Law97, Och00, Kar01]; kod mnogih vrsta bakterija uočen je višak G u odnosu na C, što se pripisuje genetskim mutacijama [Lob96, McL98]; konačno, kod brzo razvijajućih bakterija primećena je upotreba kodonskih sinonima koji se od strane tRNK prevode efikasnije i tačnije, jednom rečju, optimalniji su [Ike85, Eyr96]. Obično je različita upotreba kodona proizvod dejstva sva tri faktora [Suz08], a da bi se pouzdano odredili šabloni po kojima se upotreba razlikuje, neophodno je koristiti odgovarajuće statističke metode.

Prvi izazov u primeni odgovarajuće statističke metode je pronalaženje pouzdane i nepristrasne mere ocene značaja kodona. Prva koja se nameće je apsolutna frekvencija (AF_{kod}) kodona u okviru gena ili nekog njegovog dela. Međutim, problem sa ovom merom je što je osetljiva na ukupnu upotrebu aminokiseline u okviru genoma. Iz tog razloga, u ovoj analizi se koristi frekvencija kodona *kod* relativna ukupnoj upotrebi aminokiseline (RF_{kod}):

(1) $$RF_{kod} = \frac{AF_{kod}}{AF_{ak(kod)}} = \frac{N_{kod}}{N_{ak(kod)}}$$

(2) $$AF_{kod} = \frac{N_{kod}}{N_{uk}}$$

(3) $$AF_{ak(kod)} = \frac{N_{ak(kod)}}{N_{uk}}$$

AF_{kod} predstavlja ukupan broj pojavljivanja kodona *kod* (npr. CGA), N_{uk} je ukupan broj kodona u okviru genoma, gena ili dela gena (zavisi koji je od ovih regiona korišćen), a $N_{ak(kod)}$ je broj pojavljivanja aminokiseline u koju se kodon *kod* preslikava. Ranije je spomenuto da se upotreba kodona može analizirati na nivou vrste organizma i na nivou gena. U ovom istraživanju se ide još jedan nivo niže i posmatra upotreba kodona u zavisnosti od dela gena. Postoje dva načina da izvršimo ekvidistantnu podelu gena: prvi je da odredimo segmente fiksne dužine, npr. 300 nukleotida; i drugi, koji podrazumeva proporcionalnu podelu gena na fiksan broj segmenata, npr. 10. Genomi koji su korišćeni, imali su po nekoliko hiljada gena, i pritom je veličina pojedinačnih gena varirala od nekoliko desetina nukleotida, pa do nekoliko hiljada. Zbog velike varijanse dužine gena, prvi način je odmah na početku odbačen, a odabran je upravo drugi pristup koji koristi pomenutu podelu na 10 segmenata. Mislim da bi manji broj segmenata doveo u pitanje interpretaciju rezultata, jer bi se istraživanje svelo na ispitivanje upotrebe u okviru genoma, dok bi dodatno usitnjavanje segmenata, takođe dovelo u pitanje interpretaciju rezultata, ali i dodatno povećalo dimenzionalnost problema. Sledeći korak je podrazumevao analizu matrice relativnih frekvencija, grupisanu prema genomima, genima i delovima u okviru gena. Korišćene su dve mere centralne tendencije: aritmetička i geometrijska sredina.

(4) $$M_{arit} = \frac{\sum_{kod \in S_{kod}} RF_{kod}}{n}$$

(5) $$M_{geom} = \sqrt[n]{\prod_{kod \in S_{kod}} RF_{kod}}$$

S_{kod} je skup svih kodona koji se pojavljuju u datom segmentu, genu ili genomu, a $n = |S_{kod}|$. Nakon toga, računate su dve vrste odstupanja:

(6) $$D_{arit}(kod) = RF_{kod} - M_{arit}$$

(7) $$D_{geom}(kod) = \frac{RF_{kod}}{M_{geom}}$$

Aritmetičko odstupanje nekog kodona je razlika njegove relativne frekvencije pojavljivanja i aritmetičke sredine relativnih frekvencija svih kodona, dok se geometrijsko odstupanje dobija kao količnik relativne frekvencije kodona i geometrijske sredine relativnih frekvencija svih kodona. U analizi upotrebe kodona, korišćena su oba odstupanja, a kao „interesantni“ su smatrani samo podaci koji su bili dve standardne devijacije udaljeni od aritmetičke odnosno geometrijske sredine.

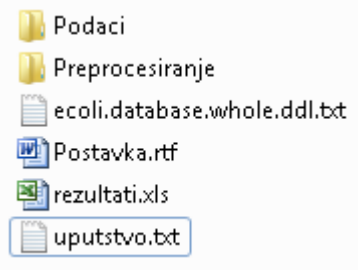
PRIMENJENE METODE

Razmatrani problem zahteva traženje potencijalnih anomalija u pogledu frekvencija kodonskih sinonima za fiksiranu aminokiselinu. Poređenje frekvencija kodona koji ne kodiraju istu aminokiselinu nema smisla, te se polazni problem transformiše u nekoliko instanci problema, za svaku amino kiselinu po jedna. Koristio sam nekoliko tehnika istraživanja podataka, a jedina koja je dovela do nekih zaključaka je „ručna“ analiza aritmetičkih i geometrijskih odstupanja pomenutih u prethodnom poglavlju. Pored ove tehnike, bez mnogo uspeha, primenjene su klaster analiza i klasifikacija.

Analiza upotrebe kodona sprovedena je na 5 bakterija iz grupe Gammaproteobacteria, a tačni nazivi organizama (kako su zavedeni u NCBI bazi genoma) su:

1. Escherichia coli APEC 01
2. Escherichia coli BL21
3. Escherichia coli 536
4. Escherichia coli 55989
5. Escherichia coli 83972

Kao što se vidi, reč je o različitim varijantama bakterije Escherichia coli, koje imaju dosta slične genome. U NCBI bazi genoma, postoji još nekoliko desetina varijanti ove bakterije, ali bi toliko iscrpno ispitivanje zahtevalo mnogo više računarskih resursa, tako da sam se ograničio na ovih 5. Podaci su preuzeti u GenBank formatu, i uključivali su meta informacije o lokacijama gena, i kodirajćim sekvencama, kao i same genomske podatke, tj. nizove nukleotida. Implementirao sam program za preprocesiranje, koji prihvata datoteke u pomenutom formatu, a kao rezultat vraća datoteku, spremnu za masovni unos podataka u DB2 relacionu bazu podataka. Izvršna datoteka programa je „koreni_dir/Podaci/Preprocesiranje.exe“, a izvorni kod se nalazi u direktorijumu „koreni_dir/ Preprocesiranje/“ (koreni_dir je koreni direktorijum arhive koja se nalazi na http://www.matf.bg.ac.rs/~kartelj/doktorske/kodoni.zip, a koja sadrži sve resurse vezane za ovaj projekat). Struktura korenog direktorijuma je kao na Slici 1.:



Slika 1. Struktura korenog direktorijuma

Pokretanjem „koreni_dir/Podaci/Preprocesiranje.cmd“, generišu se izlazne datoteke za svih 5 organizama. Pre unosa podataka u bazu, potrebno je kreirati istu, a zatim pustiti skript

„koreni_dir/ecoli.database.whole.ddl.txt“, koji kreira sve potrebne entitete. Glavni entitet u korišćenoj bazi podataka ima naziv KODON (Tabela 2):

Tabela 2: Entitet KODON

ID	BIGINT	8	No
KOMPLAMENT	CHARACTER	1	No
POCETAK	INTEGER	4	No
KODON	VARCHAR	3	No
POCETAK_GENA	INTEGER	4	No
KRAJ_GENA	INTEGER	4	No
POZICIJA_P8SK	INTEGER	4	Yes
ID_ORGANIZMA	BIGINT	8	Yes

Nakon kreiranja svih entiteta, potrebno je izvršiti uvoz podataka, pokretanjem „koreni_dir/Podaci/load.kodoni.cmd“. Tako uveženi podaci su neprerađeni, tj. potrebno ih je agregirati po raznim kriterijumima. U narednih nekoliko paragrafa, biće ukratko opisani ostali entiteti u bazi, kao i redosled potrebnih upita za punjenje istih. Svi izvršeni upiti se nalaze u datoteci „koreni_dir/Podaci/Postprocesiranje.txt“.

Mapiranje kodona na aminokiseline

U Tabeli 3, prikazana je struktura entiteta KODONAMINO, koji definiše mapiranje kodona na odgovarajuću aminokiselinu. Ovo mapiranje je u skladu sa Tabelom 1.

Tabela 3: Entitet KODONAMINO

ID	BIGINT	8	No
KODON	VARCHAR	3	No
AMINO	CHARACTER	1	No

Određivanje relativnog broja pojavljivanja kodona

Svaki gen je izdeljen na deset segmenata jednake dužine. Narednim upitom, svakom kodonu je određena pozicija u okviru pripadajućeg gena:

```
update kodon
set pozicija_p8sk=
(select (K.pocetak-K.pocetak_gena)*10/(K.kraj_gena-K.pocetak_gena)
from kodon K
where K.id=kodon.id
);
commit work;
```

Sledeći korak je utvrđivanje ukupnog broja kodona po svakom organizmu i amino kiselini unutar njega. Upit kojim se vrši ova operacija je prikazan u nastavku, a entitet u kojem se pohranjuje ova informacija je AMINOUKUPNO (Tabela 4).

```
insert into aminoukupno(id_organizma, amino, ukkodzaamino)
select K.id_organizma, KA.amino, count(*)
```

```
from kodon k join kodonamino ka on k.kodon=ka.kodon
group by K.id_organizma, KA.amino;
commit work;

update aminoukupno
set ukkodona=
(select count(*)
from kodon k where k.id_organizma=aminoukupno.id_organizma);
commit work;
```

Tabela 4: Entitet AMINOUKUPNO

ID	BIGINT	8	No
AMINO	CHARACTER	10	Yes
UKKODZAAMINO	INTEGER	4	Yes
UKKODONA	INTEGER	4	Yes
ID_ORGANIZMA	BIGINT	8	Yes

Na kraju, izračunava se relativni broj pojavljivanja svakog od kodona u zavisnosti od organizma, aminokiseline u koju se mapira i pozicije u okviru gena, primenom sledećeg upita:

```
insert into
kodonaminopozicijaP8SK(KODON,POZICIJA_P8SK,AMINO,OBA,UK_OBA_PROC,
AMINO_OBA_PROC, ID_ORGANIZMA)
select K.kodon, K.pozicija_p8sk, KA.amino, count(*) oba,
count(*)*100.0/AU.ukkodona uk_oba_proc,
count(*)*100.0/AU.ukkodzaamino amino_oba_proc,
K.ID_ORGANIZMA
from kodon K join kodonamino KA on K.kodon=KA.kodon join aminoukupno AU on
        AU.amino=KA.amino and AU.id_organizma=K.id_organizma
group by K.id_organizma,K.kodon, K.pozicija_p8sk, KA.amino,
        AU.ukkodona,AU.ukkodzaamino;
commit work;
```

Određivanje proseka i odstupanja

Najpre se vrši izračunavanje aritmetičke i geometrijske sredine, kao što je opisano u (4) i (5). Kako DB2 nema ugrađenu agregatnu funkciju za izračunavanje proizvoda, za izračunavanje geometrijske sredine, korišćena je sledeća transformacija:

$$M_{geom} = \sqrt[n]{\prod_{kod \in S_{kod}} RF_{kod}} = \exp\left(\log\left(\sqrt[n]{\prod_{kod \in S_{kod}} RF_{kod}}\right)\right)$$

$$= \exp\left(\log\left(\prod_{kod \in S_{kod}} RF_{kod}\right)^{\frac{1}{n}}\right) = \exp\left(\frac{\log\left(\prod_{kod \in S_{kod}} RF_{kod}\right)}{n}\right) \quad (8)$$

$$= \exp\left(\frac{\sum_{kod \in S_{kod}} \log(RF_{kod})}{n}\right)$$

Upit za izračunavanje geometrijske sredine iz tog razloga izgleda ovako:

```
update kodonaminopozicijap8sk
set amino_oba_geom=
(select exp(sum(log(K.amino_oba_proc))/count(*))
from kodonaminopozicijap8sk K
where K.id_organizma=kodonaminopozicijap8sk.id_organizma and
            K.amino=kodonaminopozicijap8sk.amino and
            K.pozicija_p8sk=kodonaminopozicijap8sk.pozicija_p8sk);
commit work;
```

Upit za izračunavanje aritmetičke sredine je intuitivniji:

```
update kodonaminopozicijap8sk
set amino_oba_arit=
(select sum(K.amino_oba_proc)/count(*)
from kodonaminopozicijap8sk K
where K.id_organizma=kodonaminopozicijap8sk.id_organizma and
            K.amino=kodonaminopozicijap8sk.amino and
            K.pozicija_p8sk=kodonaminopozicijap8sk.pozicija_p8sk);
commit work;
```

Odstupanja se računaju u skladu sa formulama (6) i (7):

```
update kodonaminopozicijap8sk
set amino_oba_geom_odst=amino_oba_proc/amino_oba_geom;
commit work;

update kodonaminopozicijap8sk
set amino_oba_arit_odst=amino_oba_proc-amino_oba_arit;
commit work;
```

Tabelarni prikaz podataka

Poslednji korak je „izvlačenje“ agregiranih podataka u pregledne tabele aritmetičkih i geometrijskih proseka i odstupanja. Naredni upit vraća aritmetička i geometrijska odstupanja prikazana tako, da svaki red odgovara konkretnom kodonu u okviru pripadajuće amino kiseline, po svakom organizmu zasebno. U kolonama su prikazana aritmetička i geometrijska odstupanja po svakom od deset delova gena.

```
select id_organizma, amino, kodon,
sum(case pozicija_p8sk when 0 then amino_oba_geom_odst else 0 end) godst0,
sum(case pozicija_p8sk when 0 then amino_oba_arit_odst else 0 end) aodst0,
sum(case pozicija_p8sk when 1 then amino_oba_geom_odst else 0 end) godst1,
sum(case pozicija_p8sk when 1 then amino_oba_arit_odst else 0 end) aodst1,
sum(case pozicija_p8sk when 2 then amino_oba_geom_odst else 0 end) godst2,
sum(case pozicija_p8sk when 2 then amino_oba_arit_odst else 0 end) aodst2,
sum(case pozicija_p8sk when 3 then amino_oba_geom_odst else 0 end) godst3,
sum(case pozicija_p8sk when 3 then amino_oba_arit_odst else 0 end) aodst3,
sum(case pozicija_p8sk when 4 then amino_oba_geom_odst else 0 end) godst4,
sum(case pozicija_p8sk when 4 then amino_oba_arit_odst else 0 end) aodst4,
sum(case pozicija_p8sk when 5 then amino_oba_geom_odst else 0 end) godst5,
sum(case pozicija_p8sk when 5 then amino_oba_arit_odst else 0 end) aodst5,
sum(case pozicija_p8sk when 6 then amino_oba_geom_odst else 0 end) godst6,
sum(case pozicija_p8sk when 6 then amino_oba_arit_odst else 0 end) aodst6,
sum(case pozicija_p8sk when 7 then amino_oba_geom_odst else 0 end) godst7,
sum(case pozicija_p8sk when 7 then amino_oba_arit_odst else 0 end) aodst7,
sum(case pozicija_p8sk when 8 then amino_oba_geom_odst else 0 end) godst8,
sum(case pozicija_p8sk when 8 then amino_oba_arit_odst else 0 end) aodst8,
sum(case pozicija_p8sk when 9 then amino_oba_geom_odst else 0 end) godst9,
sum(case pozicija_p8sk when 9 then amino_oba_arit_odst else 0 end) aodst9
 from kodonaminopozicijaP8SK
group by id_organizma, amino, kodon;
commit work;
```

Upit za tabelarni prikaz aritmetičkih i geometrijskih sredina se postavlja na vrlo sličan način, tj. samo se svako pojavljivanje kolone „amino_oba_geom_odst" zameni sa „amino_oba_geom" i svako pojavljivanje „amino_oba_arit_odst" zameni sa „amino_oba_arit".
Ovako struktuirani podaci su izvezeni u Excel dokument „koreni_dir/rezultati.xls". U Tabeli 5, prikazan je deo tih podataka, a radi se o odstupanjima kodona za amino kiselinu Alanin u svih 5 razmatranih organizama u prvih 6 delova gena (ima ukupno 10 delova).

Tabela 5: Odstupanja za Alanin u prvih 6 delova gena

org	am	kod	G0	A0	G1	A1	G2	A2	G3	A3	G4	A4	G5	A5
1	a	GCA	0.718	-0.728	0.673	-0.913	0.657	-0.944	0.658	-0.976	0.668	-0.929	0.680	-0.894
1	a	GCC	1.306	0.622	1.438	0.934	1.406	0.858	1.427	0.910	1.389	0.799	1.397	0.827
1	a	GCG	1.287	0.579	1.295	0.588	1.303	0.610	1.347	0.715	1.389	0.799	1.360	0.737
1	a	GCT	0.829	-0.473	0.798	-0.610	0.831	-0.524	0.791	-0.649	0.776	-0.669	0.774	-0.670
2	a	GCA	0.714	-0.759	0.643	-0.981	0.647	-0.970	0.652	-0.995	0.647	-0.977	0.694	-0.858
2	a	GCC	1.309	0.649	1.401	0.827	1.398	0.799	1.392	0.794	1.370	0.728	1.354	0.729
2	a	GCG	1.280	0.581	1.351	0.708	1.401	0.807	1.446	0.927	1.450	0.917	1.389	0.814
2	a	GCT	0.836	-0.471	0.822	-0.554	0.789	-0.637	0.763	-0.726	0.778	-0.668	0.766	-0.684
3	a	GCA	0.711	-0.761	0.662	-0.941	0.658	-0.933	0.656	-0.980	0.654	-0.960	0.681	-0.881
3	a	GCC	1.317	0.657	1.443	0.939	1.422	0.880	1.428	0.913	1.386	0.783	1.391	0.814
3	a	GCG	1.287	0.586	1.306	0.608	1.304	0.600	1.345	0.709	1.399	0.815	1.342	0.696
3	a	GCT	0.830	-0.482	0.801	-0.607	0.820	-0.547	0.794	-0.642	0.789	-0.638	0.787	-0.629
4	a	GCA	0.711	-0.758	0.652	-0.968	0.653	-0.954	0.661	-0.965	0.667	-0.938	0.680	-0.886
4	a	GCC	1.303	0.624	1.426	0.897	1.405	0.847	1.431	0.915	1.400	0.815	1.389	0.807
4	a	GCG	1.293	0.601	1.327	0.657	1.324	0.653	1.349	0.715	1.412	0.844	1.352	0.720
4	a	GCT	0.835	-0.467	0.811	-0.585	0.823	-0.547	0.784	-0.665	0.758	-0.721	0.783	-0.641

5 a	GCA	0.712	-0.757	0.667	-0.931	0.661	-0.931	0.668	-0.952	0.666	-0.918	0.686	-0.871
5 a	GCC	1.317	0.659	1.442	0.945	1.429	0.900	1.429	0.928	1.380	0.769	1.383	0.795
5 a	GCG	1.281	0.574	1.291	0.580	1.306	0.607	1.327	0.675	1.390	0.792	1.351	0.719
5 a	GCT	0.832	-0.477	0.806	-0.594	0.810	-0.576	0.790	-0.650	0.782	-0.643	0.781	-0.643

REZULTATI I ZAKLJUČAK

Glavni zadatak u analizi dobijenih podataka, bilo je uočavanje pravilnosti upotrebe kodona na nekim genskim lokacijama za datu aminokiselinu u okviru svih 5 organizama. U gotovo svim situacijama, kada bi pravilnost bila uočena, ona je bila zastupljena u svih 5 razmatranih genoma. Ipak, bilo je i sporadičnih slučajeva kada je u nekim genomima ta pravilnost bila jače izražena. Postupak uočavanja pravilnosti izvršen je na sledeći način:

1. Na prethodno opisanoj tabeli odstupanja, izvršeno je sortiranje podataka u opadajućem poretku prema vrednosti geometrijskog odstupanja za kolonu 1, tj. za početni deo gena.
2. Sva odstupanja koja su veća od 2, markirana su prvom bojom, a sva koja su manja od 0.5, markirana su drugom bojom.
3. Zatim je izvršeno sortiranje po aritmetičkom odstupanju, takođe u opadajućem redosledu i to po prvoj koloni.
4. Sva aritmetička odstupanja koja su veća od 2 su markirana trećom bojom, a sva koja su manja od -2, markirana su četvrtom bojom.
5. Koraci 1-4 ponovljeni su za sve kolone, odnosno za podatke iz svih delova gena.
6. Na kraju izvršeno je sortiranje tabele prema kodonu, organizmu i amino kiselini.

Korišćene boje i njihova interpretacija, prikazane su u Tabeli 6:

Tabela 6: Boje i njihova interpretacija

	Amino. geom. odst. > 2
	Amino. geom. odst.<0.5
	Amino. arit. odst.>2
	Amino. arit. odst. <-2

Na ovaj način omogućeno je lakše uočavanje kodona i genskih lokacija od interesa. U narednih nekoliko paragrafa, opisaću uočene pravilnosti.

Cistein (C)

Cistein je kodiran kodonima TGC i TGT. Primećeno je aritmetičko odstupanje od 2.06 u okviru drugog (A2) dela gena i 2.01 u šestog (A6) delu gena drugog razmatranog organizma (delovi gena su numerisani brojevima 0-9).

Tabela 7: Statistički značajno odstupanje u 2. i 6. delu gena (Escherichia BL21)

org	am	kod	G0	A0	G1	A1	G2	A2	G3	A3	G4	A4	G5	A5	G6	A6
1	c	TGC	1.42	1.82	1.47	1.86	1.44	1.76	1.50	1.85	1.49	1.86	1.48	1.81	1.49	1.79
1	c	TGT	0.70	-1.82	0.68	-1.86	0.69	-1.76	0.67	-1.85	0.67	-1.86	0.68	-1.81	0.67	-1.79
2	c	TGC	1.42	1.74	1.45	1.78	1.55	2.06	1.49	1.87	1.48	1.81	1.49	1.85	1.53	2.01
2	c	TGT	0.70	-1.74	0.69	-1.78	0.64	-2.06	0.67	-1.87	0.68	-1.81	0.67	-1.85	0.65	-2.01
3	c	TGC	1.41	1.73	1.46	1.84	1.43	1.77	1.49	1.83	1.47	1.82	1.46	1.73	1.47	1.75
3	c	TGT	0.71	-1.73	0.68	-1.84	0.70	-1.77	0.67	-1.83	0.68	-1.82	0.69	-1.73	0.68	-1.75
4	c	TGC	1.39	1.66	1.42	1.70	1.45	1.81	1.47	1.78	1.48	1.87	1.45	1.73	1.44	1.66
4	c	TGT	0.72	-1.66	0.71	-1.70	0.69	-1.81	0.68	-1.78	0.67	-1.87	0.69	-1.73	0.70	-1.66
5	c	TGC	1.42	1.76	1.47	1.86	1.46	1.83	1.48	1.83	1.45	1.76	1.47	1.75	1.46	1.75
5	c	TGT	0.71	-1.76	0.68	-1.86	0.68	-1.83	0.67	-1.83	0.69	-1.76	0.68	-1.75	0.68	-1.75

Slično se dobija i za vrednost u koloni A6. Zaključak je da organizam Escherichia coli BL21

uslovno poseduje značajnu* povećanu upotrebu kodona TGC u okviru 2. i 6. dela gena. Ipak, cilj ove analize nije utvrđivanje pravilnosti svojstvenih samo za neke od organizama i lokacija, već onih pravilnosti koje su zajedničke za sve organizme Escherichia coli.

Lizin (K)

Lizin je kodiran kodonima AAA i AAG. Odnos upotrebe AAA kodona u odnosu na AAG je skoro konstantan u svim delovima gena, zbog toga ova, iako uneravnotežena upotreba, nema nikakav značaj kada je u pitanju razmatrani problem (Tabela 8).

Tabela 8: AAA i AAG imaju slična odstupanja u svim delovima gena

org	am	kod	A0	A1	A2	A3	A4	A5	A6	A7	A8	A9
	1k	AAA	2.16	2.03	2.11	2.03	2.13	2.19	2.16	2.24	2.17	2.14
	1k	AAG	-2.16	-2.03	-2.11	-2.03	-2.13	-2.19	-2.16	-2.24	-2.17	-2.14
	2k	AAA	2.14	2.15	2.21	2.23	2.30	2.19	2.26	2.18	2.15	2.25
	2k	AAG	-2.14	-2.15	-2.21	-2.23	-2.30	-2.19	-2.26	-2.18	-2.15	-2.25
	3k	AAA	2.19	2.09	2.11	2.05	2.14	2.38	2.15	2.27	2.16	2.15
	3k	AAG	-2.19	-2.09	-2.11	-2.05	-2.14	-2.38	-2.15	-2.27	-2.16	-2.15
	4k	AAA	2.22	2.08	2.04	1.96	2.05	2.17	2.06	2.20	2.11	2.03
	4k	AAG	-2.22	-2.08	-2.04	-1.96	-2.05	-2.17	-2.06	-2.20	-2.11	-2.03
	5k	AAA	2.26	2.01	2.12	2.04	2.21	2.25	2.15	2.26	2.19	2.16
	5k	AAG	-2.26	-2.01	-2.12	-2.04	-2.21	-2.25	-2.15	-2.26	-2.19	-2.16

Leucin (L)

Leucin je kodiran sledećim kodonima: TTA, TTG, CTT, CTC, CTA i CTG. Tabela 9 prikazuje aritmetička i geometrijska odstupanja za ovu amino kiselinu.

Tabela 9: Značajno mala upotreba CTG u prvom delu gena

org	am	kod	G0	A0	G1	A1	G2	A2	G3	A3	G4	A4	G5	A5	G6	A6	G7	A7	G8	A8	G9	A9
1	l	CTA	0.24	-1.61	0.19	-1.43	0.18	-1.45	0.17	-1.45	0.18	-1.43	0.18	-1.44	0.19	-1.42	0.19	-1.42	0.18	-1.43	0.18	-1.35
1	l	CTC	1.18	-0.06	1.48	0.09	1.51	0.04	1.49	0.02	1.46	-0.01	1.50	0.02	1.43	-0.04	1.43	-0.04	1.39	-0.08	1.19	-0.23
1	l	CTG	2.49	2.11	3.68	2.69	4.13	2.95	4.08	2.91	4.09	2.89	4.18	2.96	4.14	2.97	4.10	2.92	4.06	2.91	3.58	2.41
1	l	CTT	1.01	-0.35	1.14	-0.32	1.06	-0.46	1.09	-0.43	1.06	-0.46	1.03	-0.50	1.02	-0.50	1.00	-0.52	1.01	-0.50	1.16	-0.26
1	l	TTA	1.29	0.13	0.65	-0.90	0.63	-0.95	0.62	-0.95	0.60	-0.97	0.63	-0.94	0.61	-0.96	0.64	-0.91	0.65	-0.91	0.71	-0.76
1	l	TTG	1.08	-0.22	1.29	-0.13	1.36	-0.13	1.39	-0.09	1.47	-0.01	1.40	-0.09	1.43	-0.05	1.44	-0.03	1.48	0.02	1.58	0.20
2	l	CTA	0.27	-1.60	0.20	-1.42	0.20	-1.42	0.18	-1.45	0.18	-1.42	0.20	-1.39	0.19	-1.41	0.21	-1.39	0.21	-1.38	0.20	-1.33
2	l	CTC	1.08	-0.20	1.41	0.01	1.41	-0.01	1.40	-0.07	1.40	-0.10	1.43	-0.02	1.41	-0.08	1.32	-0.15	1.32	-0.12	1.12	-0.31
2	l	CTG	2.48	2.19	3.82	2.84	3.93	2.91	4.15	3.03	4.32	3.07	4.13	2.98	4.26	3.04	4.19	3.06	4.05	2.97	3.65	2.49
2	l	CTT	0.97	-0.39	1.03	-0.44	1.01	-0.48	1.04	-0.49	1.06	-0.47	0.99	-0.51	1.01	-0.52	0.92	-0.60	0.94	-0.55	1.06	-0.38
2	l	TTA	1.33	0.22	0.67	-0.87	0.67	-0.87	0.66	-0.92	0.66	-0.90	0.63	-0.92	0.64	-0.92	0.66	-0.89	0.68	-0.84	0.72	-0.74
2	l	TTG	1.07	-0.22	1.32	-0.11	1.31	-0.13	1.37	-0.11	1.33	-0.18	1.34	-0.13	1.39	-0.10	1.42	-0.04	1.36	-0.07	1.64	0.27
3	l	CTA	0.25	-1.63	0.20	-1.43	0.18	-1.45	0.17	-1.46	0.19	-1.42	0.18	-1.42	0.17	-1.43	0.19	-1.40	0.18	-1.42	0.19	-1.34
3	l	CTC	1.15	-0.10	1.52	0.15	1.51	0.05	1.49	0.01	1.47	0.01	1.49	0.01	1.46	-0.03	1.38	-0.07	1.38	-0.08	1.17	-0.25
3	l	CTG	2.46	2.12	3.62	2.65	4.06	2.92	4.06	2.86	4.04	2.88	4.17	2.92	4.18	2.93	4.05	2.89	4.01	2.86	3.65	2.47
3	l	CTT	0.99	-0.37	1.10	-0.35	1.02	-0.51	1.09	-0.43	1.01	-0.50	1.00	-0.52	1.03	-0.50	0.95	-0.55	0.99	-0.52	1.15	-0.28
3	l	TTA	1.31	0.17	0.63	-0.91	0.66	-0.91	0.62	-0.95	0.60	-0.96	0.63	-0.92	0.62	-0.94	0.66	-0.88	0.66	-0.89	0.68	-0.79

* Pokazalo se da je ovo odstupanje statistički značajno za 90% interval poverenja, ukoliko pretpostavimo normalan raspored. Npr., aritmetička sredina i standardna devijacija u okviru A2 regiona za kodon TGC ako uzmemo u obzir sve organizme su 1.84 i 0.12. Normalizacijom vrednosti 2.06 dobijamo (2.06-1.84)/0.12=1.74, a pridružena p-vrednost je 8.186%. Međutim, budući da je uzorak mali, svega 5 organizama, ne može se smatrati relevantnim.

3 l	TTG	1.10	-0.19	1.31	-0.11	1.38	-0.10	1.44	-0.04	1.45	-0.01	1.42	-0.07	1.46	-0.03	1.46	0.01	1.51	0.06	1.58	0.19
4 l	CTA	0.25	-1.63	0.18	-1.45	0.19	-1.44	0.18	-1.43	0.19	-1.42	0.19	-1.41	0.19	-1.42	0.21	-1.38	0.20	-1.40	0.20	-1.32
4 l	CTC	1.19	-0.03	1.57	0.17	1.51	0.06	1.48	0.01	1.48	0.02	1.47	0.00	1.42	-0.05	1.41	-0.04	1.41	-0.02	1.19	-0.22
4 l	CTG	2.45	2.11	3.70	2.67	4.03	2.93	4.10	2.92	4.06	2.91	4.17	2.98	4.15	2.96	4.06	2.94	3.93	2.85	3.58	2.42
4 l	CTT	1.02	-0.33	1.16	-0.30	1.03	-0.48	1.05	-0.47	1.00	-0.51	1.02	-0.50	1.05	-0.46	0.96	-0.53	0.97	-0.53	1.12	-0.29
4 l	TTA	1.27	0.11	0.64	-0.91	0.63	-0.94	0.61	-0.96	0.62	-0.94	0.60	-0.96	0.61	-0.96	0.66	-0.88	0.64	-0.90	0.70	-0.76
4 l	TTG	1.07	-0.23	1.26	-0.19	1.34	-0.13	1.41	-0.06	1.40	-0.06	1.36	-0.12	1.41	-0.07	1.34	-0.11	1.43	0.00	1.54	0.18
5 l	CTA	0.24	-1.64	0.19	-1.43	0.18	-1.45	0.17	-1.47	0.19	-1.41	0.18	-1.41	0.18	-1.42	0.19	-1.40	0.18	-1.43	0.19	-1.34
5 l	CTC	1.15	-0.11	1.52	0.12	1.52	0.05	1.50	0.00	1.44	-0.02	1.43	-0.05	1.43	-0.05	1.38	-0.09	1.39	-0.09	1.18	-0.25
5 l	CTG	2.53	2.21	3.75	2.72	4.10	2.95	4.15	2.95	4.09	2.93	4.19	2.97	4.19	2.96	4.10	2.93	4.09	2.90	3.66	2.50
5 l	CTT	1.01	-0.36	1.13	-0.34	1.02	-0.50	1.09	-0.45	1.04	-0.48	1.02	-0.50	1.02	-0.50	0.98	-0.52	1.00	-0.52	1.14	-0.29
5 l	TTA	1.29	0.12	0.61	-0.94	0.64	-0.93	0.62	-0.97	0.59	-0.98	0.64	-0.92	0.62	-0.94	0.67	-0.87	0.66	-0.89	0.70	-0.78
5 l	TTG	1.09	-0.22	1.31	-0.13	1.36	-0.13	1.43	-0.06	1.43	-0.04	1.40	-0.08	1.44	-0.05	1.41	-0.05	1.50	0.03	1.54	0.15

Ključno zapažanje, kada je u pitanju ova amino kiselina, je da je upotreba kodona CTG značajno niža u okviru prvog dela gena. Npr., kod prvog organizma, geometrijsko i aritmetičko odstupanje su 2.49 i 2.11 u prvom delu gena, a već u narednom se beleži nagli porast na 3.68 i 2.69. Odstupanja kulminiraju negde na sredini genoma, i u 6. delu gena iznose: 4.18 i 2.96. Pred sam kraj genoma beleži se blagi pad upotrebe ovog kodona, ali ne toliko značajan, koliko na samom početku. Ova pravilnost uočena je u svim razmatranim organizmima.

Glutamin (Q)

Tabela 10: Manja aritmetička odstupanja na početku i kraju gena

org	am	kod	G0	A0	G1	A1	G2	A2	G3	A3	G4	A4	G5	A5	G6	A6	G7	A7	G8	A8	G9	A9
1	q	CAA	0.66	-1.86	0.57	-2.55	0.58	-2.55	0.57	-2.59	0.56	-2.55	0.55	-2.70	0.57	-2.61	0.58	-2.48	0.58	-2.43	0.66	-1.92
1	q	CAG	1.51	1.86	1.75	2.55	1.71	2.55	1.75	2.59	1.77	2.55	1.81	2.70	1.76	2.61	1.73	2.48	1.71	2.43	1.52	1.92
2	q	CAA	0.66	-1.89	0.60	-2.40	0.59	-2.49	0.59	-2.48	0.59	-2.47	0.60	-2.39	0.58	-2.48	0.58	-2.50	0.59	-2.32	0.65	-1.94
2	q	CAG	1.50	1.89	1.67	2.40	1.71	2.49	1.71	2.48	1.69	2.47	1.68	2.39	1.72	2.48	1.71	2.50	1.69	2.32	1.53	1.94
3	q	CAA	0.67	-1.81	0.58	-2.50	0.59	-2.51	0.58	-2.55	0.57	-2.48	0.56	-2.67	0.57	-2.57	0.59	-2.41	0.60	-2.34	0.65	-2.00
3	q	CAG	1.49	1.81	1.72	2.50	1.70	2.51	1.73	2.55	1.75	2.48	1.80	2.67	1.74	2.57	1.69	2.41	1.68	2.34	1.54	2.00
4	q	CAA	0.66	-1.90	0.57	-2.55	0.58	-2.57	0.56	-2.64	0.56	-2.57	0.54	-2.80	0.57	-2.58	0.57	-2.56	0.58	-2.51	0.64	-2.01
4	q	CAG	1.52	1.90	1.76	2.55	1.74	2.57	1.78	2.64	1.78	2.57	1.85	2.80	1.75	2.58	1.76	2.56	1.74	2.51	1.55	2.01
5	q	CAA	0.66	-1.88	0.58	-2.52	0.59	-2.53	0.58	-2.53	0.57	-2.54	0.55	-2.70	0.57	-2.55	0.59	-2.41	0.59	-2.39	0.65	-1.99
5	q	CAG	1.51	1.88	1.73	2.52	1.71	2.53	1.73	2.53	1.76	2.54	1.81	2.70	1.74	2.55	1.69	2.41	1.70	2.39	1.53	1.99

Glutamin kodiraju dva kodona: CAA i CAG. Aritmetička odstupanja, kao što se vidi u Tabeli 10, jače su izražena nego geometrijska. Naime, ovde se radi o nešto manjoj apsolutnoj upotrebi Glutamina u početnim i krajnjim delovima kodona. To se može jasno videti iz Tabele 11, koja prikazuje neagregirane podatke za prvi i drugi deo gena, kao i prosečne vrednosti upotrebe kodona. Kolona „**% od uk**“ prikazuje ukupnu upotrebu kodona u celom genomu (ne samo u odnosu na pripadajuću amino kiselinu). Postavlja se pitanje, da li je aritmetičko odstupanje u ovom slučaju nepristrasno?

Tabela 11: Neagregirani podaci za prvi i drugi deo gena

org	am	poz	kod	br. poj	% od uk	% od am	geom.	geom odst	arit	arit odst
1	q	0	CAG	5422	0.37	6.66	4.42	1.51	4.80	1.86
1	q	0	CAA	2391	0.16	2.94	4.42	0.66	4.80	-1.86
2	q	0	CAG	5034	0.37	6.76	4.49	1.50	4.87	1.89
2	q	0	CAA	2223	0.16	2.98	4.49	0.66	4.87	-1.89

3 q	0 CAG	5397	0.37	6.59	4.43	1.49	4.78	1.81
3 q	0 CAA	2432	0.17	2.97	4.43	0.67	4.78	-1.81
4 q	0 CAG	5687	0.37	6.73	4.43	1.52	4.83	1.90
4 q	0 CAA	2471	0.16	2.92	4.43	0.66	4.83	-1.90
5 q	0 CAG	5609	0.37	6.67	4.40	1.51	4.79	1.88
5 q	0 CAA	2444	0.16	2.91	4.40	0.66	4.79	-1.88
1 q	**1 CAG**	**6168**	**0.42**	**7.57**	**4.33**	**1.75**	**5.02**	**2.55**
1 q	1 CAA	2014	0.14	2.47	4.33	0.57	5.02	-2.55
2 q	**1 CAG**	**5575**	**0.41**	**7.48**	**4.48**	**1.67**	**5.08**	**2.40**
2 q	1 CAA	1997	0.15	2.68	4.48	0.60	5.08	-2.40
3 q	**1 CAG**	**6167**	**0.42**	**7.54**	**4.38**	**1.72**	**5.04**	**2.50**
3 q	1 CAA	2079	0.14	2.54	4.38	0.58	5.04	-2.50
4 q	**1 CAG**	**6387**	**0.42**	**7.56**	**4.30**	**1.76**	**5.00**	**2.55**
4 q	1 CAA	2068	0.13	2.45	4.30	0.57	5.00	-2.55
5 q	**1 CAG**	**6357**	**0.42**	**7.56**	**4.37**	**1.73**	**5.04**	**2.52**
5 q	1 CAA	2118	0.14	2.52	4.37	0.58	5.04	-2.52

Iz Tabele 11, vidi se da je je na poziciji 0, npr. kod prvog organizma, odnos relativnog broja pojavljivanja kodona CAG i CAA jednak 6.66/2.94=2.27 (specijalno, ovaj odnos je jednak odnosu geometrijskih odstupanja, kada imamo samo dva elementa, kao što je ovde slučaj). Na poziciji 1, imamo odnos 7.57/2.47=3.06, tj. dosta veću nesrazmernost. Dakle, i na osnovu geometrijskih odstupanja, može se zaključiti da se radi o značajnoj pravilnosti. Zbog toga, ne odbacujem ovu pravilnost, već samo sugerišem, da je možda, odabrani interval za geometrijska odstupanja $(-\infty,0.5]\cup[2,+\infty)$ neadekvatan.

Seratonin (S)

Utvrđena je pravilnost za kodon AGC, kada je u pitanju nulti deo gena. U tom delu, geometrijsko odstupanje po ovom kodonu je značajno manje nego u ostalim delovima (Tabela 12). Npr., u nultom delu gena prvog organizma beleži se geometrijsko odstupanje od 1.57, što je dosta manje od od 2.1 u sledećem delu, ili čak 2.26 u središnjem delu. Pravilnost je uočena u svim organizmima.

Tabela 12: Manja upotreba AGC kodona na početku

org am	kod	G0	A0	G1	A1	G2	A2	G3	A3	G4	A4	G5	A5	G6	A6	G7	A7	G8	A8	G9	A9
1 s	AGC	1.57	1.04	2.10	1.50	2.15	1.51	2.16	1.49	2.26	1.58	2.17	1.52	2.14	1.50	2.14	1.48	2.21	1.63	2.16	1.59
1 s	AGT	1.04	0.02	1.25	0.22	1.21	0.14	1.26	0.20	1.22	0.13	1.27	0.22	1.28	0.24	1.26	0.21	1.27	0.22	1.29	0.27
1 s	TCA	1.06	0.04	0.54	-0.85	0.51	-0.88	0.49	-0.89	0.51	-0.85	0.47	-0.94	0.51	-0.89	0.50	-0.88	0.50	-0.92	0.49	-0.95
1 s	TCC	1.03	-0.01	1.19	0.13	1.22	0.15	1.25	0.18	1.19	0.10	1.14	0.03	1.17	0.08	1.20	0.12	1.20	0.12	1.06	-0.08
1 s	TCG	0.73	-0.58	0.81	-0.45	0.87	-0.35	0.84	-0.40	0.86	-0.37	0.92	-0.29	0.90	-0.31	0.90	-0.31	0.83	-0.43	0.86	-0.38
1 s	TCT	0.77	-0.51	0.74	-0.55	0.72	-0.58	0.71	-0.58	0.70	-0.59	0.74	-0.55	0.69	-0.62	0.68	-0.62	0.71	-0.61	0.81	-0.45
2 s	AGC	1.64	1.15	2.12	1.54	2.12	1.50	2.14	1.51	2.22	1.60	2.15	1.54	2.22	1.61	2.29	1.68	2.22	1.61	2.15	1.57
2 s	AGT	1.02	-0.03	1.15	0.07	1.20	0.14	1.15	0.07	1.16	0.07	1.16	0.09	1.28	0.25	1.14	0.03	1.24	0.18	1.29	0.28
2 s	TCA	1.00	-0.07	0.52	-0.89	0.49	-0.92	0.52	-0.84	0.53	-0.83	0.53	-0.85	0.51	-0.89	0.50	-0.89	0.52	-0.88	0.57	-0.80
2 s	TCC	1.03	-0.01	1.24	0.20	1.17	0.09	1.19	0.12	1.13	0.03	1.14	0.06	1.07	-0.07	1.18	0.08	1.15	0.05	1.04	-0.10
2 s	TCG	0.71	-0.63	0.90	-0.32	0.90	-0.30	0.88	-0.32	0.87	-0.34	0.89	-0.31	0.89	-0.34	0.90	-0.31	0.93	-0.27	0.84	-0.39
2 s	TCT	0.82	-0.41	0.71	-0.60	0.76	-0.51	0.74	-0.53	0.74	-0.52	0.75	-0.52	0.73	-0.56	0.72	-0.58	0.64	-0.70	0.73	-0.56
3 s	AGC	1.61	1.10	2.10	1.49	2.16	1.50	2.17	1.49	2.27	1.60	2.12	1.46	2.13	1.49	2.19	1.52	2.21	1.62	2.22	1.68
3 s	AGT	1.03	-0.01	1.24	0.20	1.19	0.12	1.29	0.24	1.22	0.14	1.28	0.25	1.27	0.22	1.27	0.21	1.30	0.26	1.33	0.32
3 s	TCA	1.06	0.04	0.52	-0.87	0.51	-0.88	0.49	-0.91	0.52	-0.84	0.47	-0.93	0.51	-0.89	0.48	-0.93	0.49	-0.94	0.50	-0.94
3 s	TCC	1.04	0.01	1.19	0.13	1.25	0.20	1.29	0.24	1.18	0.08	1.14	0.04	1.24	0.18	1.23	0.15	1.18	0.09	1.05	-0.10

3 s	TCG	0.75	-0.56	0.79	-0.47	0.85	-0.38	0.81	-0.44	0.86	-0.36	0.90	-0.30	0.89	-0.33	0.88	-0.35	0.85	-0.40	0.82	-0.45
3 s	TCT	0.74	-0.58	0.79	-0.48	0.73	-0.56	0.70	-0.61	0.68	-0.62	0.77	-0.50	0.67	-0.66	0.70	-0.61	0.70	-0.63	0.78	-0.51
4 s	AGC	1.67	1.19	2.08	1.48	2.08	1.46	2.12	1.44	2.27	1.60	2.14	1.49	2.19	1.56	2.21	1.57	2.18	1.62	2.12	1.55
4 s	AGT	1.03	-0.02	1.16	0.10	1.17	0.10	1.28	0.24	1.25	0.17	1.26	0.21	1.23	0.17	1.29	0.23	1.22	0.16	1.28	0.28
4 s	TCA	1.05	0.02	0.55	-0.82	0.53	-0.85	0.50	-0.89	0.51	-0.86	0.49	-0.90	0.51	-0.88	0.49	-0.92	0.53	-0.87	0.53	-0.87
4 s	TCC	1.04	0.00	1.22	0.19	1.22	0.18	1.23	0.17	1.17	0.06	1.15	0.06	1.17	0.08	1.18	0.08	1.16	0.07	1.06	-0.06
4 s	TCG	0.70	-0.64	0.82	-0.42	0.88	-0.32	0.87	-0.35	0.83	-0.41	0.91	-0.29	0.92	-0.28	0.88	-0.36	0.86	-0.37	0.83	-0.41
4 s	TCT	0.75	-0.55	0.75	-0.53	0.72	-0.57	0.69	-0.61	0.72	-0.57	0.72	-0.56	0.67	-0.65	0.70	-0.61	0.70	-0.61	0.79	-0.48
5 s	AGC	1.61	1.08	2.15	1.55	2.11	1.46	2.21	1.53	2.24	1.58	2.13	1.50	2.13	1.47	2.16	1.51	2.18	1.60	2.19	1.63
5 s	AGT	1.02	-0.04	1.21	0.14	1.26	0.21	1.27	0.21	1.25	0.18	1.27	0.23	1.28	0.24	1.30	0.26	1.26	0.22	1.33	0.33
5 s	TCA	1.06	0.04	0.53	-0.86	0.52	-0.87	0.50	-0.89	0.52	-0.85	0.49	-0.91	0.49	-0.91	0.49	-0.92	0.51	-0.91	0.53	-0.87
5 s	TCC	1.05	0.02	1.21	0.16	1.22	0.16	1.29	0.23	1.21	0.12	1.15	0.06	1.21	0.14	1.22	0.15	1.22	0.15	1.03	-0.13
5 s	TCG	0.73	-0.59	0.81	-0.45	0.86	-0.36	0.82	-0.43	0.84	-0.40	0.92	-0.28	0.91	-0.31	0.86	-0.38	0.85	-0.40	0.83	-0.42
5 s	TCT	0.76	-0.52	0.75	-0.54	0.69	-0.61	0.68	-0.64	0.68	-0.63	0.71	-0.59	0.68	-0.64	0.70	-0.62	0.68	-0.66	0.75	-0.55

Tirozin (T)

Za Tirozin je utvrđena pravilnost po kodonu ACC u svim razmatranim organizmima. U početnim i krajnjim delovima gena, geometrijsko odstupanje naglo opada. Tako npr. u nultom delu gena, prvog organizma, geometrijsko odstupanje je 1.78, a već u sledećem dolazi do povećanja. U pretposlednjem delu gena, odstupanje je malo manje od 2, tako da se može reći da je promena odstupanja značajna samo u prvom i poslednjem delu gena.

Tabela 13: Smanjena upotreba ACC kodona na početku i na kraju gena

org	am	kod	G0	A0	G1	A1	G2	A2	G3	A3	G4	A4	G5	A5	G6	A6	G7	A7	G8	A8	G9	A9
1	t	ACA	0.67	-0.62	0.53	-0.90	0.52	-0.90	0.49	-0.95	0.51	-0.92	0.51	-0.89	0.51	-0.91	0.50	-0.91	0.55	-0.83	0.57	-0.73
1	t	ACC	1.78	1.10	2.11	1.52	2.12	1.50	2.07	1.46	2.18	1.57	2.12	1.46	2.09	1.45	2.05	1.41	1.95	1.31	1.59	0.81
1	t	ACG	1.51	0.69	1.49	0.58	1.53	0.62	1.53	0.64	1.56	0.64	1.57	0.65	1.60	0.72	1.53	0.63	1.46	0.56	1.34	0.44
1	t	ACT	0.79	-0.44	0.75	-0.56	0.71	-0.61	0.79	-0.50	0.75	-0.57	0.72	-0.59	0.72	-0.60	0.76	-0.53	0.76	-0.51	0.78	-0.41
1	t	TAC	0.86	-0.32	0.92	-0.29	0.95	-0.25	0.94	-0.27	0.97	-0.23	0.98	-0.21	0.95	-0.26	0.97	-0.21	0.98	-0.18	1.00	-0.08
1	t	TAT	0.81	-0.40	0.88	-0.36	0.88	-0.36	0.86	-0.39	0.80	-0.49	0.84	-0.41	0.85	-0.40	0.86	-0.38	0.86	-0.35	1.04	-0.03
2	t	ACA	0.64	-0.68	0.50	-0.93	0.50	-0.92	0.49	-0.96	0.49	-0.94	0.49	-0.93	0.49	-0.92	0.54	-0.85	0.50	-0.91	0.55	-0.76
2	t	ACC	1.81	1.16	2.10	1.50	2.12	1.51	2.07	1.46	2.07	1.43	2.08	1.44	2.04	1.39	2.10	1.49	1.97	1.32	1.64	0.86
2	t	ACG	1.51	0.68	1.49	0.56	1.41	0.45	1.46	0.53	1.52	0.61	1.48	0.55	1.47	0.54	1.45	0.51	1.49	0.59	1.42	0.53
2	t	ACT	0.81	-0.41	0.75	-0.56	0.76	-0.53	0.73	-0.59	0.77	-0.52	0.73	-0.58	0.79	-0.47	0.72	-0.58	0.73	-0.56	0.80	-0.39
2	t	TAC	0.90	-0.28	0.96	-0.23	0.99	-0.18	1.03	-0.13	1.00	-0.17	1.02	-0.15	0.98	-0.19	0.98	-0.19	1.05	-0.08	0.96	-0.15
2	t	TAT	0.77	-0.48	0.89	-0.35	0.89	-0.33	0.90	-0.32	0.84	-0.41	0.89	-0.33	0.88	-0.34	0.86	-0.37	0.88	-0.34	1.01	-0.08
3	t	ACA	0.66	-0.64	0.53	-0.88	0.53	-0.88	0.50	-0.94	0.51	-0.92	0.54	-0.85	0.51	-0.91	0.51	-0.90	0.54	-0.83	0.57	-0.74
3	t	ACC	1.77	1.08	2.12	1.53	2.16	1.58	2.10	1.49	2.18	1.57	2.03	1.39	2.06	1.42	2.07	1.43	1.91	1.26	1.60	0.83
3	t	ACG	1.50	0.67	1.45	0.52	1.50	0.58	1.50	0.58	1.52	0.59	1.52	0.62	1.58	0.69	1.50	0.59	1.46	0.56	1.33	0.42
3	t	ACT	0.82	-0.39	0.74	-0.56	0.73	-0.58	0.80	-0.47	0.76	-0.55	0.72	-0.58	0.74	-0.57	0.75	-0.54	0.77	-0.48	0.82	-0.36
3	t	TAC	0.87	-0.31	0.93	-0.28	0.92	-0.30	0.95	-0.25	0.98	-0.22	0.97	-0.20	0.96	-0.24	0.95	-0.24	0.97	-0.18	1.01	-0.07
3	t	TAT	0.80	-0.42	0.89	-0.33	0.86	-0.40	0.84	-0.41	0.80	-0.48	0.85	-0.38	0.86	-0.39	0.88	-0.35	0.87	-0.33	0.99	-0.09
4	t	ACA	0.64	-0.67	0.52	-0.91	0.53	-0.87	0.50	-0.93	0.52	-0.89	0.53	-0.88	0.52	-0.88	0.50	-0.91	0.54	-0.84	0.56	-0.74
4	t	ACC	1.78	1.10	2.11	1.53	2.06	1.46	2.06	1.46	2.12	1.49	2.06	1.42	2.02	1.36	2.00	1.36	1.95	1.30	1.61	0.83
4	t	ACG	1.49	0.64	1.47	0.54	1.45	0.53	1.50	0.60	1.55	0.65	1.55	0.66	1.54	0.65	1.48	0.57	1.47	0.57	1.35	0.44
4	t	ACT	0.80	-0.43	0.76	-0.53	0.74	-0.55	0.80	-0.48	0.74	-0.57	0.71	-0.60	0.73	-0.57	0.79	-0.47	0.76	-0.50	0.82	-0.36
4	t	TAC	0.89	-0.28	0.94	-0.26	0.96	-0.23	0.97	-0.22	0.98	-0.21	0.97	-0.21	0.97	-0.21	0.96	-0.22	0.97	-0.19	0.99	-0.11
4	t	TAT	0.83	-0.38	0.88	-0.36	0.89	-0.33	0.84	-0.42	0.81	-0.47	0.86	-0.39	0.87	-0.35	0.89	-0.33	0.87	-0.34	1.02	-0.05
5	t	ACA	0.67	-0.62	0.53	-0.89	0.54	-0.86	0.48	-0.96	0.52	-0.91	0.55	-0.84	0.53	-0.87	0.52	-0.87	0.55	-0.83	0.58	-0.72
5	t	ACC	1.78	1.10	2.12	1.54	2.13	1.54	2.09	1.47	2.18	1.58	2.07	1.44	2.03	1.39	2.06	1.42	1.93	1.28	1.62	0.85
5	t	ACG	1.50	0.67	1.47	0.55	1.48	0.56	1.53	0.62	1.52	0.59	1.50	0.58	1.56	0.67	1.49	0.58	1.47	0.58	1.33	0.41

5 t	ACT	0.80	-0.43	0.73	-0.59	0.74	-0.56	0.79	-0.50	0.76	-0.55	0.72	-0.58	0.71	-0.60	0.76	-0.52	0.75	-0.53	0.79	-0.39
5 t	TAC	0.87	-0.31	0.93	-0.27	0.91	-0.30	0.96	-0.24	0.95	-0.25	0.97	-0.21	0.98	-0.20	0.94	-0.26	0.98	-0.17	1.01	-0.06
5 t	TAT	0.80	-0.41	0.89	-0.34	0.86	-0.38	0.86	-0.39	0.81	-0.47	0.85	-0.39	0.85	-0.39	0.87	-0.35	0.87	-0.34	1.00	-0.09

Zaključak

U okviru ovog seminarskog rada, izvršena je analiza upotrebe genoma u zavisnosti od njegove lokacije u okviru gena. Geni su podeljeni na 10 delova jednake dužine, na taj način je svakom kodonu u okviru genoma dodeljena lokacija (broj između 0 i 9). Podaci su agregirani prema organizmima, amino kiselinama i lokacijama, a potom su izračunate potrebne statistike. Sva merenja su sprovedena na relativnim frekvencijama kodona u okviru pripadajuće amino kiseline. Tako organizovani podaci doprineli su nepristrasnosti korišćenih statističkih mera i ocena. Aritmetičko odstupanje se pokazalo kao manje relevantan kriterijum od geometrijskog odstupanja, jer je u nekim situacijama prenaglašavalo značajnost stvarnog odstupanja. Međutim, ni u jednom slučaju nije utvrđena velika neuravnoteženost između ove dve mere. Utvrđeno je nekoliko pravilnosti, čija biološka, odnosno genetička interpretacija nije utvrđena (jer ne posedujem potrebno stručno znanje iz pomenutih oblasti). Od bitnijih zaključaka navodim sledeće: kod Leucina je utvrđena umanjena upotreba CTG kodona u okviru početnog dela gena; upotreba CAA i CAG (Glutamin) je takođe značajno varirala na početku gena; kod Tirozina i Seratonina su utvrđene pravilnosti u pogledu upotrebe AGC i ACC kodona na početnim i krajnjim delovima. U pogledu načina analize podataka, moje mišljenje je da su korišćenje metode odgovorile na postavljeni problem, iako su prilično bazične u pogledu istraživanja podataka.

LITERATURA

[Eyr96] Eyre-Walker, A. 1996, Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy?, Mol Biol Evol, 13, 864–872.

[Ike85] Ikemura, T. 1985, Codon usage and tRNA content in unicellular and multicellular organisms, Mol Biol Evol, 2, 13–34.

[Kar01] Karlin, S. 2001, Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes, Trends Microbiol., 9, 335–343.

[Law97] Lawrence, J. G. and Ochman, H. 1997, Amelioration of bacterial genomes: rates of change and exchange, J. Mol. Evol., 44, 383–397.

[Lob96] Lobry, J. R. 1996, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol., 13, 660–665.

[McL98] McLean, M. J., Wolfe, K. H. and Devine, K. M. 1998, Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes, J. Mol. Evol., 47, 691–696.

[Och00] Ochman, H., Lawrence, J. G. and Groisman, E. A. 2000, Lateral gene transfer and the nature of bacterial innovation, Nature, 405, 299–304.

[Sha88] Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C.,Wolfe, K. H. and Wright, F. 1988, Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable withinspecies diversity, Nucleic Acids Res., 16, 8207–8211.

[Suz08] Haruo Suzuki, Celeste J. Brown, Larry J. Forneq, and Eva M. Top. 2008, Comparison of Correspondence Analysis Methods for Synonymous Codon Usage in Bacteria, Oxford Journals, DNA Research, 15, 357-365.