



VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS

**Bioinformatikos pirmojo laboratorinio darbo ataskaita - Kodonų ir dikodonų dažnių
analizė žinduolių ir bakterijų virusuose**

(Analysis of Codon and Dicodon Usage in Mammalian and Bacterial Viruses)

Parengė: Informatikos 4 kurso informatikos studentė Vasarė Pratužaitė

Turinys

<i>Įvadas</i>	3
<i>Užduoties formuluotė</i>	4
<i>Metodai</i>	4
Duomenys	4
ORF paieška ir filtravimas.....	5
Vertimas į aminorūgščių sekas	5
Dažnių skaičiavimas	5
Atstumo matricos sudarymas.....	5
Klasterinė analizė.....	6
<i>Rezultatų išsaugojimas</i>	6
<i>Filogenetinis medžiai ir jų analizė</i>	7
<i>Labiausiai varijuojantys kodonai ir dikodonai</i>	8
Top 10 varijuojantys kodonai.....	8
Top 10 varijuojantys dikodonai.....	8
<i>Išvados</i>	8

Įvadas

Šiame laboratoriniame darbe buvo nagrinėjama kodonų ir dikodonų dažnio analizė žinduolių bei bakterijų virusų genomuose, kurios tikslas – įvertinti ir išanalizuoti šių grupių tarpusavio skirtumus bei nustatyti, ar jos sudaro atskirus klasterius pagal baltymų aminorūgščių sudėtį. Norint tai išsiaiškinti, būtina suprasti, kaip veikia baltymų sintezės procesas ir genetinio kodo principai. Kiekvienas baltymas biologinėse sistemose yra užkoduotas tam tikra nukleotidų seka, kurią sudaro trijų bazių grupės – kodonai. Kiekvienas kodonas atitinka konkrečią aminorūgštį, o kai kurios aminorūgštys gali būti koduojamos keliais skirtingais kodonais. Šis reiškinys vadinamas genetinio kodo degeneracija ir lemia, kad skirtingi organizmai gali naudoti skirtingus kodonus tam pačiam baltymui koduoti. Tokie skirtumai tiriami kodonų naudojimo analizės metodu, kuris leidžia įvertinti, kaip virusai evoliuciškai prisitaikė prie savo šeimininko translacijos aparato. Kadangi baltymų sintezės efektyvumui įtakos turi „šeimininko“ tRNR molekulių prieinamumas, virusai dažnai evoliuciškai prisitaiko prie „šeimininko“ tRNR repertuaro, optimizuodami savo kodonų naudojimo modelius. Be pavienių kodonų analizės, galima tirti ir dikodonų dažnius, t. y. dviejų gretimų aminorūgščių derinius baltymuose. Ši analizė leidžia vertinti subtilius skirtumus baltymų struktūroje, funkcijoje ir sudėtyje, o būtent šie skirtumai gali atskleisti evoliucinius prisitaikymus, baltymų erdvinės struktūros (sulankstymo) savybes bei translacijos greičio ypatumus. Šiame darbe buvo atlikta tiek kodonų, tiek dikodonų dažnių analizė. Gauti rezultatai buvo palyginti tarp žinduolių ir bakterijų virusų, apskaičiuotos atstumo matricos pagal šių dažnių skirtumus, o klasterinė analizė atlikta taikant Neighbour Joining (NJ) metodą. Šioje ataskaitoje terminai *dikodonai* ir *dipeptidai* bei *aminorūgštys* ir *kodonai* vartojami kaip sinonimai, nes analizė buvo atliekama baltymo lygmenyje, o ne nukleotidų sekoje. Biologiniu požiūriu, *kodonas* yra trijų nukleotidų seka, koduojanti vieną aminorūgštį, o *dikodonas* – dviejų iš eilės esančių kodonų derinys DNR arba RNR grandinėje. Kadangi užduotyje buvo nurodyta kodonus ir dikodonus analizuoti baltymų lygmenyje („vienas kodonas – aminorūgštis, vienas dikodonas – dvi aminorūgštys“), šiame kontekste dikodonų analizė atitinka dipeptidų dažnių analizę.

Užduoties formuluotė

1. **Nustatyti koduojančias sekas (ORF):** Pateiktose FASTA formato sekose surandamos visos start (ATG) ir stop (TAA, TAG, TGA) kodonų poros, tarp kurių nėra kitų stop kodonų. Analizė atliekama tiek tiesioginei sekai, tiek jos reverse complement atitikmeniui.
2. **Pasirinkti tinkamas start–stop poras:** Kiekvienam stop kodonui parenkamas toliausiai nuo jo esantis start kodonas, su sąlyga, kad tarp jų nėra kito stop kodono.
3. **Atlikti fragmentų filtravimą:** Iš atrinktų sekų pašalinami visi fragmentai, trumpesni nei 100 bp, nes jie laikomi per trumpi patikimai baltymų kodavimo analizei.
4. **Konvertuoti DNR sekas į baltymo sekas:** Koduojantys fragmentai (ORF'ai) verčiami į atitinkamas aminorūgščių sekas pagal genetinį kodą. Šiame etape analizė atliekama baltymų lygmenyje.
5. **Apskaičiuoti kodonų ir dikodonų (dipeptidų) dažnius:** Kiekvienai sekai nustatomi visi galimi aminorūgščių (kodonų) ir dipeptidų (dviejų aminorūgščių kombinacijų) dažniai. Įtraukiami ir tie atvejai, kai tam tikrų derinių nėra (dažnis = 0).
6. **Sudaryti atstumo matricas ir atlikti palyginimą:** Pagal gautus dažnius apskaičiuojami tarpusavio atstumai tarp sekų (naudojant pasirinktą metriką, pvz., cosine distance). Rezultatai išsaugomi PHYLIP formato matricoje, kurioje pateikiami atstumai tarp visų tiriamų virusų.
7. **Klasterizuoti virusus pagal panašumą:** Naudojant Neighbour Joining metodą (per T-Rex įrankį), sudaromi evoliuciniai medžiai. Pagal juos įvertinama, ar žinduolių ir bakterijų virusai sudaro atskirus klasterius pagal kodonų ir dikodonų dažnius.

Metodai

Analizė buvo paremta automatizuota „Python 3“ kalba parašyta programa, kuri apdorojo pateiktas FASTA formato sekas, naudodama standartines „Python“ bibliotekas bei papildomą „Pandas“ paketą rezultatų apdorojimui ir lentelių formavimui. Gauti duomenys buvo panaudoti tolesnei analizei: iš kodonų ir dikodonų dažnių apskaičiuotos atstumo matricos, kurios vėliau buvo panaudotos „Neighbour Joining“ (NJ) klasterizacijos metodui, siekiant vizualizuoti virusų panašumus pagal jų kodonų ir dikodonų naudojimo dažnius. Filogenetiniai medžiai buvo sukonstruoti naudojant internetinį įrankį „T-Rex“, leidžiantį iš PHYLIP formato matricos sugeneruoti virusų klasterizacijos medžius. Rezultato ieškojau naudodama kosinuso metodą, kurį pasirinkau siekiant rasti atstumus (žr. žemiau).

Duomenys

Analizei buvo panaudoti po keturis žinduolių („mammalian“) ir bakterijų virusų genomus FASTA formatu (šios sekos buvo pridėtos prie užduoties). Kiekviena seka buvo apdorota skirtingų krypties formatu, siekiant aptikti visus galimus koduojančius fragmentus (ORF).

ORF paieška ir filtravimas

Programoje buvo identifikuojamos visos galimos start (ATG, GTG, TTG) ir stop (TAA, TAG, TGA) kodonų poros. Kiekvienam stop kodonui buvo parenkamas toliausiai nuo jo esantis start kodonas, su sąlyga, kad tarp jų nebūtų kito stop kodono. Gauti fragmentai, trumpesni nei 100 bp, buvo atmesti kaip per trumpi patikimai baltymų kodavimo analizei. Paieška atlikta tiek tiesiogiai, tiek atvirkštinėje sekoje, kad būtų aptiktos visos galimos koduojančios sritys.

Vertimas į aminorūgščių sekas

Atrinktos DNR sekos buvo konvertuotos į aminorūgščių sekas pagal standartinį genetinį kodą. Vertimui buvo naudota kodonų lentelė (pagal NCBI standartą), o konversija atlikta su specialiai parašyta funkcija. Šiame etape analizė buvo perkelta į baltymų lygmenį, kad būtų galima skaičiuoti aminorūgščių (kodonų) ir dipeptidų (dviejų iš eilės esančių aminorūgščių) dažnius.

Dažnių skaičiavimas

Kiekvienai sekai buvo apskaičiuoti:

- Aminorūgščių (kodonų) dažniai – visų 20 pagrindinių aminorūgščių pasikartojimai santykinėmis reikšmėmis.
- Dipeptidų (dikodonų) dažniai – visų galimų aminorūgščių porų ($20 \times 20 = 400$ kombinacijų) pasiskirstymas.

Jeigu tam tikrų aminorūgščių ar dipeptidų sekoje nebuvo, jiems priskirta reikšmė 0, kad dažnių vektoriai būtų vienodo ilgio visoms sekoms.

Atstumo matricos sudarymas

Gauti dažnių vektoriai buvo palyginti tarpusavyje, apskaičiuojant porinius atstumus tarp visų virusų. Atstumo funkcijai buvo pasirinkta kosinusinio panašumo (cosine distance) metrika, kuri vertina vektorių krypties panašumą nepriklausomai nuo absoliučių reikšmių dydžių. Kiekvienam dažnių tipui buvo sudaryta atskira PHYLIP formato atstumo matrica.

Žingsniai:

1. Kiekvienai viruso sekai buvo suformuotas vektorius:

$A = (a_1, a_2, a_3, \dots, a_n)$ ir $B = (b_1, b_2, b_3, \dots, b_n)$, kur a_i ir b_i žymi atitinkamų aminorūgščių (arba dipeptidų) santykinius dažnius.

Panašumas tarp dviejų sekų A ir B apskaičiuojamas pagal formulę:

$$\text{cosine_similarity}(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||}$$

Kur

$$A \cdot B = \sum_{i=1}^n a_i b_i, \text{ o } ||A|| = \sqrt{\sum_{i=1}^n a_i^2}, \text{ ir } ||B|| = \sqrt{\sum_{i=1}^n b_i^2}$$

2. Apskaičiuojamas kosinuso atstumas, nes juo galima įvertinti dažnių vektorių krypties panašumo, nepriklausomai nuo jų absoliutaus dydžio.
 - $d(A, B) = 0$ reiškia visišką panašumą (identiški dažnių profiliai)
 - $d(A, B) \rightarrow 1$ reikia, kad profiniai yra visiškai skirtingi.
 - $D(A, B) = d(B, A)$ – matricos simetriškos
3. Kiekvienai sekai apskaičiuojamas aminorūgščių arba dipeptidų dažnių vektorius A_i .
4. Visos sekos poromis lyginamos tarpusavyje: $D_{i,j} = (A_i, A_j)$
5. Gautos reikšmės įrašomos į simetrišką atstumo matricą D , kur $D_{i,j} = 0$.
6. Matrica išsaugoma PHYLIP formatu, kuris vėliau naudojamas klasterinei analizei.

Klasterinė analizė

Gautos atstumo matricos buvo panaudotos Neighbour Joining (NJ) klasterizacijos metodui, siekiant vizualizuoti virusų panašumus pagal jų kodonų ir dikodonų naudojimo dažnius. Medžiai buvo sukonstruoti naudojant internetinį įrankį T-Rex, kuris leidžia iš PHYLIP formato matricos sugeneruoti filogenetinį medį.

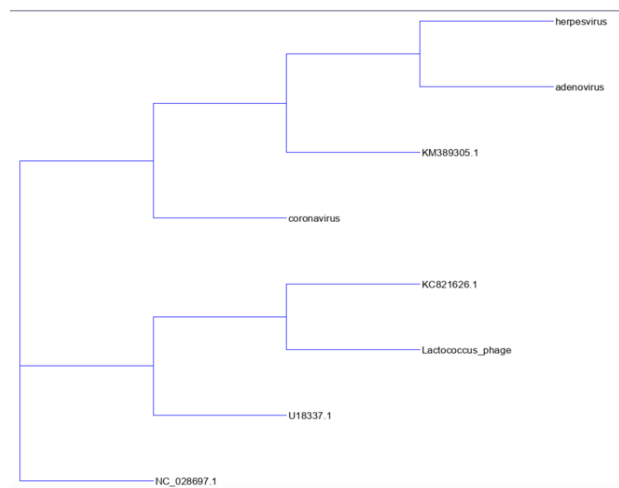
Rezultatų išsaugojimas

Analizės rezultatai (dažnių lentelės ir atstumo matricos) buvo išsaugoti kataloge „results“:

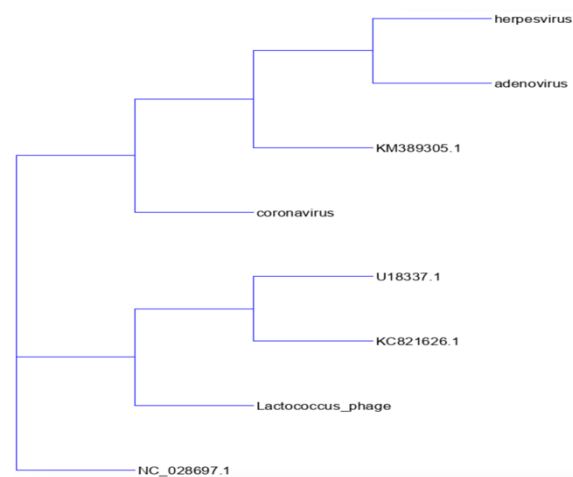
- **aa_freqs.tsv** – aminorūgščių dažniai(kodonai);
- **dipep_freqs.tsv** – dipeptidų dažniai(dikodonai);
- dicodon_freqs.tsv – papildomas;
- **distances_aa.phy** – PHYLIP matrica aminorūgščių dažniams (kodonai);
- **distances_dipep.phy** – PHYLIP matrica dipeptidų dažniams (dikodonai).
- distances_dicodon.phy – papildomas (nukleotidų lygmuo);
- distances_codon.phy – papildomas (nukleotidų lygmuo).

Filogenetinis medžiai ir jų analizė

Filogenetinis medis, sukurtas remiantis aminorūgščių dažnių skirtumais tarp žinduolių ir bakterijų virusų, atskleidė aiškią dviejų pagrindinių grupių diferenciaciją. Viršutinėje medžio dalyje išsidėsto žinduolių virusai (herpesvirus, adenovirus, coronavirus ir KM389305.1), kurie sudaro glaudų klasterį, rodantį aminorūgščių panašumus. Šių virusų artimumas rodo, kad jų translacijos ir baltymų sintezės sistemos yra panašiai prisitaikiusios prie žinduolių šeimininkų tRNR repertuaro. Apatinėje dalyje matyti atskira grupė, kurią sudaro bakteriofagai (Lactococcus_phage, U18337.1, KC821626.1 ir NC_028697.1). Šie virusai tarpusavyje yra labai artimi, o tai patvirtina, kad jų baltymų sudėtis ir kodonų naudojimas yra būdingi bakterinių virusų evoliucinei linijai. Didžiausi atstumai stebimi tarp bakterinių ir žinduolių virusų klasterių (ypač tarp herpesvirus ir Lactococcus_phage), kas rodo ryškius skirtumus jų aminorūgščių naudojimo dažniuose. Tokie skirtumai atspindi evoliucinius prisitaikymus prie skirtingų šeimininkų - žinduolių ir bakterijų - translacijos įvykių. Pagal gautus duomenis (žr. pav. 1 ir pav. 2) galima pamatyti, kad skirtumų nėra. Taip yra todėl, nes buvo taikomas kosinuso metodas, kuris vertina kampą tarp vektoriaus (kryptį) ir parodo, kiek dvi sekos turi panašų santykinį kodonų ir dikodonų pasiskirstymą, bet nėra jautrus absoliučiam dažnių dydžiui, jei būtų taikomas euklidinio atstumo metodas, reikšmės išsiskirtų, nes būtų naudojamas matuojant geometrinį atstumą tarp taškų erdvėje. Manheteno atstumo skaičiavimo metodas padėtų taip pat gauti kitokius medžius, nes būtų sumuojami visi absoliutūs dažnių skirtumai.



pav. 2 Kodonų filogenetinis medis



pav. 1 Dikodonų filogenetinis medis

Labiausiai varijuojantys kodonai ir dikodonai

Top 10 varijuojantys kodonai	
R	0.000466
I	0.000444
K	0.000362
P	0.000232
A	0.000176
N	0.000159
G	0.000115
Y	0.000109
L	0.000105
C	0.000072

Lentelė 1

Top 10 varijuojantys dikodonai	
RR	0.000036
AA	0.000011
II	0.000009
LK	0.000008
PR	0.000007
RG	0.000007
NI	0.000006
PP	0.000006
LI	0.000006
LR	0.000006

Lentelė 2

Išvados

1. Atlikta analizė parodė aiškų skirtumą tarp žinduolių ir bakterinių virusų pagal kodonų bei dikodonų (dipeptidų) dažnius, o analizijuojant filogenetinius medžius buvo pastebėta, kad medžiai rodo, jog šios dvi grupės formuoja atskirus klasterius, o tai patvirtina jų evoliucinį prisitaikymą prie skirtingų šeimininkų translacijos sistemų.
2. Didžiausia variancija pastebėta tarp kodonų R (arginas), I (izoleucinas) ir K (lizinas), bei dikodonų RR, AA, II, o tai rodo, jog šie elementai labiausiai prisideda prie virusų baltymų skirtumų tarp šeimininkų grupių.
3. Naudojant kosinuso atstumo metodą buvo įvertintas panašumas tarp sekų pagal kryptį (santykių dažnių pasiskirstymą), todėl skirtumai tarp grupių buvo švelnesni. Tačiau, jei būtų naudotas Euklidinis ar Manheteno metodas, filogenetiniai medžiai būtų rodę aiškesnę diferenciaciją.
4. Žinduolių ir bakterijų virusai pasižymi skirtingais kodonų bei dikodonų naudojimo modeliais, kurie atspindi jų evoliucinį prisitaikymą prie skirtingų šeimininkų.