Liet. matem. rinkinys, T. 43, spec. nr., 649–654

Palindrominių sekų naikinimo mikroorganizmų genomuose statistinis tyrimas

Tomas REKAŠIUS, Albertas TIMINSKAS (VGTU)
el. paštas: tomas.rekasius@fmst.vtu.lt, timis@ibt.lt

1. Įvadas

Kokybiniai ir kiekybiniai pokyčiai gyvų organizmų genomuose sąlygojo labai didelę gyvosios gamtos įvairovę: nuo mažiausių ir paprasčiausių virusų, iki didžiulių ir itin sudėtingų, kaip *Homo sapiens*. Žinodami veikiančius procesus besikeičiant genetinei informacijai, galėtume prognozuoti kiekvieno iš išorinių ar vidinių veiksnių įtaką šiems procesams.

Iš ankstesnių tyrimų žinoma, kad mikroorganizmų genomuose vienos iš intensyviausiai besikeičiančios – palindrominės sekos. Išrenkant per retai ir per dažnai pasitaikančius nukleotidų derinius, šiame darbe bandoma atsekti galimus palindrominių sekų išnykimo ir atsiradimo kelius bakteriniuose genomuose.

Natūralu tokių sekų ieškoti intensyviai besikeičiančiuose organizmuose. Įvertinant galimą mutacijų skaičių ir standartinį nuokrypį konkrečiam oligonukleotidui, galima „apčiuopti“ biologiškai svarbias vietas genome. Tačiau toks matas neleidžia palyginti bakterijų tarpusavyje. Šiame darbe pasiūlytas naujas metodas mutacijų intensyvumui bakterijų genomuose įvertinti, kuris nepriklauso nuo konkrečių oligonukleotidų ar genomo ilgio ir todėl leidžia palyginti mikroorganizmus tarpusavyje.

Pažymėtina, kad procesai nulemiantys genetinės informacijos raidą sudėtingi, o jų eksperimentinis tyrimas sunkus ir brangus. Tiriant šiuos procesus *in silico* labai efektyviai panaudojami statistinė nukleotidinių derinių genomuose analizė bei matematinis vykstančių procesų modeliavimas. Rezultatai gauti visiems 3–7 nukleotidų ilgio deriniams. Paminėsime, kad visų galimų trinukleotidų skaičius yra 64 (4^3), o heptanukleotidų net 16384 (4^7). Vieno genomo analizė, priklausomai nuo jo ilgio ir kompiuterio procesoriaus tipo, gali užtrukti nuo kelių iki keliolikos valandų.

Šiame darbe ištirti 50 bakterinių organizmų genomai. Duomenys analizei paimti iš *GenBank* duomenų bazių.

2. Tyrimo objektas

Genomas – genetinės informacijos visuma, lemianti kiekvieno organizmo ir rūšies, kuriai jis priklauso, individualumą. Jo pagrindinės charakteristikos yra dydis ir genų skaičius. Sudėtingėjant organizmui, genomas didėja. Bakterijų genomai yra palyginti maži, iki

kelių milijonų nukleotidų porų, todėl yra labai patrauklus objektas tiriant gyvoje ląstelėje vykstančius procesus bei modeliuojant aukštesniųjų organizmų ląstelių veiklą.

DNR molekulę sudaro dvi viena kitai komplementarios polinukleotidinės grandinės (1 pav.). Jos monomerai yra keturi skirtingi nukleotidai: adeninas (žymimas *A*), citozinas (*C*), guaninas (*G*) ir timinas (*T*). *DNR* molekulės grandinėse nukleotidai išsidėstę tam tikra tvarka. Pvz., jei vienoje grandinėje yra adeninas, tai kitoje grandinėje atitinkamoje vietoje bus timinas, citozinas vienoje grandinėje stovės prieš guaniną kitoje.

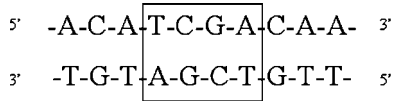
Duomenų bazėse paprastai pateikiama tik viena *DNR* molekulės grandinė. Pvz., bakterijos *Bacillus halodurans* genomo fragmentas atrodo taip: *TGAGGACTTTGAGGATTTT*...

Dėl įvairių vidinių ir išorinių priežasčių *DNR* grandinėse atsiranda klaidų (mutacijų). Mutacijos – natūralus įvykis bet kuriame genome. Tai vienas iš daugelio mechanizmų, sukeliantis natūralius genomo pokyčius. Mutacijų intensyvumas gali parodyti kad, mikroorganizmas gyvena agresyvioj ar besikeičiančioj aplinkoj ir yra priverstas keisti savo genetinę informaciją.

Taškinės mutacijos atsiranda kai vienoje grandinėje nukleotidas pasikeičia į tokį, kuris nėra komplementarus. Gaunama dviguba *DNR* grandinė su neatiktimis. Ląstelės dalijimosi metu tokia neatitiktis gali būti užfiksuota. Dėl taškinių mutacijų baltymą koduojančioje sekoje gana dažnai pakinta šių baltymų amino rūgščių seka.

Aiškiausi iki šiol pastebėti kitimo taikiniai – palindrominės sekos. Tai tokio tipo sekos, kurios abiejose komplementariose grandinėse iš 5’ į 3’ galą skaitomos vienodai (1 pav.).

Nustatyta, kad palindromuose pokyčiai vyksta intensyviausiai, bent jau prokariotiniuose organizmuose. Daugelis per retai bakteriniuose genomuose esančių žodžių yra trumpi palindromai („žodžiu“ vadiname nukleotidų derinį – oligomerą). Manoma, kad palindrominių sekų išnykimas genomuose susijęs su restrikcijos-modifikacijos (*R-M*) fermentais. Pagal tokį modelį, patekusi į organizmą *R-M* sistema iššaukia reakciją naikinti tos sistemos taikinius – palindromines sekas. Sąsaja tarp palindromų naikinimo ir genus koduojančių *R-M* sistemų buvimo bakteriniame genome buvo detalizuota aptikus, kad modifikuotas citozinas gali virsti į timiną. Iškelta hipotezė, kad tokių restrikcijos taikinių (palindromų) išnykimas gali būti paaiškinamas atsitiktinėmis ir vėliau užfiksuotomis modifikavimo klaidomis. Suprantama, kad užtenka vienos tokios klaidos, kad palindromas virstų į nepalindrominę seką. Tokiu būdu genome sumažėja palindrominių ir padidėja kitokių sekų skaičius.



1 pav. Palindrominė seka.

3. Matematinis modelis

Tegul seka S žymi genomo grandinę. Tada n nukleotidų ilgio seka S aprašoma taip: $S = S_1S_2 \dots S_n : S_i \in \{A, C, G, T\},\ i = 1 \dots n$. Trumpesnė seka s yra L ilgio oligomeras ir aprašo genomo grandinės S fragmentą: $s = s_1s_2 \dots s_L : s_i \in \{A, C, G, T\},\ i = 1 \dots L$. Laikoma, kad nukleo rūgščių sekos tenkina apibendrintą Markovo grandinės modelį, o tai reiškia, kad jos turi baigtinę „atmintį“. Stacionariai k eilės Markovo grandinei teisinga lygybė: $p(S_N|S_{N-1}, \dots, S_1) = p(S_N|S_{N-1}, \dots, S_{N-k})$. Tinkamos eilės k parinkimas yra atskiras ir gana svarbus uždavinys [2]. Nuo to priklauso modelio jautrumas ir kaip tiksliai aptinkamos biologiškai „įdomios“ vietos genome.

Teorinis sekos $s = (s_1s_2 \dots s_L)$ dažnis $k = L - 2$ eilės Markovo grandinei skaičiuojamas taip:

$$F(s_1s_2 \dots s_L) = \frac{F(s_1s_2 \dots s_{L-1})F(s_2s_3 \dots s_L)}{F(s_2s_3 \dots s_{L-1})}. \tag{1}$$

Tarkim, kad genomo seka S yra $k = L - 2$ eilės Markovo grandinė. Tada iš duomenų suskaičiavus $L - 1$ ir $L - 2$ ilgio sekų $s_k = (s_1s_2 \dots s_{L-1})$, $s_d = (s_2s_3 \dots s_L)$ ir $s_v = (s_2s_3 \dots s_{L-1})$ dažnius, sekos s dažnio įvertis yra:

$$E(s_1s_2 \dots s_L) = \frac{f(s_1s_2 \dots s_{L-1})f(s_2s_3 \dots s_L)}{f(s_2s_3 \dots s_{L-1})}. \tag{2}$$

Dydis $d(s)$ vadinamas normalizuotu nuokrypiu:

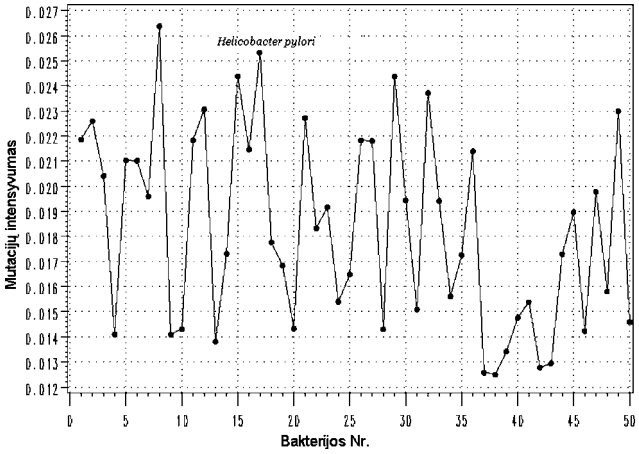
$$d(s_1s_2 \dots s_L) = \frac{f(s_1s_2 \dots s_L) - E(s_1s_2 \dots s_L)}{\sqrt{E(s_1s_2 \dots s_L)}}. \tag{3}$$

$d(s)$ skaičiavimas remiasi puasonine aproksimacija ir parodo kiek daug seka s neatitinka Markovo modelio. Laikoma, kad oligomerai s, kuriems $|d(s)| < a$, mikroorganizmo genome yra neinformatyvūs. Oligomerai, kuriems $|d(s)| \geqslant a$, yra potencialiai biologiškai aktyvios vietos genome. Tačiau tinkamos a reikšmės parinkimas priklauso ir nuo sekos s ilgio, ir nuo viso genomo ilgio. Todėl lyginti organizmus pagal $d(s)$ reikšmę tam pačiam oligomerui nėra korektiška. Iškyla poreikis įsivesti universalų matą mutacijų intensyvumui genomuose įvertinti, tokį, kuris leistų palyginti skirtingus organizmus.

4. Mutacijų intensyvumas

Žinant mutacijų intensyvumą galima įvertinti informacijos keitimo procesus organizme, kelti klausimą apie minimalų informacijos kitimo lygį, kuris būtinas išgyventi. Šiuo atveju mutacijų intensyvumas yra kriterijus atrinkti bakterijas palindrominių sekų analizei. Šiame darbe mutacijų intensyvumą genome siūlome skaičiuoti taip:

$$I = \frac{1}{L_{\max} - L_{\min} + 1} \sum_{j=L_{\min}}^{L_{\max}} I_j, \quad I_j = \frac{1}{jn} \sum_{s} (E(s) - f_s)_+. \tag{4}$$



2 pav. Bakteriju genomų mutacijų intensyvumo grafikas.

$L_{\min}(L_{\max})$ yra minimalus (maksimalus) oligomerų ilgis, n yra analizuojamo genomo ilgis, oligomero s dažnis f_s randamas iš duomenų. Šiame darbe $L_{\min} = 3$ ir $L_{\max} = 7$. Ilgesnėms nei 7 nukleotidai sekoms dažnių f_s ir $E(s)$ skaičiavimai reikalauja daug kompiuterio resursų ir laiko. Pvz., kai $L = 8$, tokių sekų yra net 65536 (4^8).

Kiekvienam organizmui paskaičiavę mutacijų intensyvumą I, galime juos palyginti tarpusavyje (2 pav.), grupuoti organizmus į klases pagal tai, kaip intensyviai vyksta mutacijos jų genomuose.

Tos pačios bakterijos genome skirtinguose nukleotidų deriniuose mutacijos vyksta nevienodai aktyviai. Nors daugumoje oligomerų mutacijų dažnis nėra didelis, tačiau galima pastebėti, kad kai kurie nukleotidų deriniai keičiasi intensyviai ir kryptingai.

5. Palindrominių sekų analizė

Palindrominių sekų analizei paimta viena iš intensyviausiai mutuojančių bakterija *Helicobacter pylori*. Jos genomo ilgis apie $1,6 \times 10^6$ nukleotidų, nukleotidai *C* ir *G* sudaro apie 39,2% genomo.

Visų bakterijų genomams sudaromos tri-, tetra-, penta- ir t.t. ilgio nukleotidų kombinacijų dažnių lentelės (1 lentelė).

1 lentelė. Nukleotidinių kombinacijų dažnių lentelė

Nr.	Oligomeras	f_s	$E(s)$	$f_s/E(s)$	$d(s)$
1	*AAAA*	79658	83252	0,96	12,46
2	*AAAC*	28568	24716	1,16	24,50
…	…	…	…	…	…
255	*TTTG*	33769	34518	0,98	4,03
256	*TTTT*	79658	83252	0,96	12,46

2 lentelė. Palindromo išnykimas

Oligomeras	$f_s/E(s)$	$d(s)$
TCGA	0,13	62,07
Galima mutacijos kryptis		
G*CGA / TCG****C***	1,34	34,28
A*CGA / TCG****T***	1,31	23,71

3 lentelė. Palindromo atsiradimas

Oligomeras	$f_s/E(s)$	
	min	max
CTA / TAG	0,58	0,78
Galima mutacijos kryptis		
C*TG / CA****G***	1,26	1,55

Jei santykis $f_s/E(s)$ didesnis už vienetą, sakoma, kad tokia nukleotidų kombinacija mikroorganizmo genome yra dažnesnė nei atsitiktinė. Dažnai palindrominei sekai šis santykis yra mažesnis už vienetą. Tai reiškia, kad šita nukleotidų kombinacija genome yra „naikinama“.

Remiantis tuo, kad dviejų mutacijų tikimybė trumpoje sekoje labai maža, galimos mutacijų kryptys nustatomos iš dažnių lentelės išrenkant tokias nukleotidų kombinacijas, kurios nuo nagrinėjamo palindromo skiriasi tik vienu nukleotidu. Pvz., bakterijos *Helicobacter pylori* genome rečiausias tetranukleotidas yra palindromas *TCGA*. Vienu nukleotidu nuo šito palindromo skiriasi ir yra dažnos keturios kitos nukleotidinės kombinacijos, po dvi kiekvienoje *DNR* grandinėje (2 lentelė).

Nors daugumoje mikroorganizmų palindrominės sekos gana intensyviai mutuoja ir virsta į kitas sekas, tačiau vyksta ir atvirkštinis procesas – palindromų atsiradimas. Pvz., trinukleotidas *CTA* (*TAG* komplementarioj grandinėj) tarp tirtų organizmų yra vienas iš rečiausių oligonukleotidų nepalindromų (3 lentelė). Jis yra genomo sekos, koduojančios baltymą, *stop* kodonas. Suprantama, kad tokios nukleotidinės kombinacijos skaičius genome yra ribotas. Tačiau įdomu, kad vienu nukleotidu besiskiriantys trinukleotidai *CTG* ir *CAG* (komplementarioj grandinėj) yra gana dažnai pasitaikantys palindromai.

Iš lentelėse pateiktų rezultatų galima atsekti palindrominių sekų išnykimo ir atsiradimo kelius. Tokia informacija gali pasitarnauti biologams tiriantiems informacijos kitimo procesus mikroorganizmų genomuose.

6. Išvados

- Pasiūlytas matas mutacijų intensyvumui bakterijų genomuose įvertinti, kuris leidžia pagal šį kriterijų palyginti skirtingus organizmus, juos klasifikuoti. Parametrų

parinkimo klausimas nenagrinėtas.

- Mutacijos organizmuose vyksta nevienodai intensyviai, tačiau ryškios priklausomybės nuo agresyvios ir didelę įtaką genomui darančios aplinkos nepastebėta.
- Rezultatai rodo, kad palindrominės sekos atsiranda ar išnaikinamos specifiniu būdu, tačiau skirtinguose organizmuose tie procesai vyksta nevienodai intensyviai
- Analizuotuose genomuose dauguma aptiktų *R-M* sistemų dar nėra identifikuotos, nenustatytos jų *DNR* taikinio sekos ir specifiškumas, todėl nustačius būdingiausius virsmus sudaromos prielaidos fermentų, dalyvaujančių genetinės informacijos kaitoje paieškai ir eksperimentiniams tyrimams.

Literatūra

[1] S. Kandrotienė, *Palindrominių sekų naikinimo mikroorganizmų genomuose tyrimai*, Baigiamais magistro darbas: bioinžinerija, Vilnius (2002).

[2] E.E. Stuckle, C. Emmrich, U. Grob, P. Nielsen, Statistical analysis of nucleotide sequences, *Nucleic Acids Research*, **18**(22), 6641–6647 (1990).

[3] E.M. Panina, A.A. Mironov, M.S. Genfand, Statistical analysis of complete bacterial genomes: avoidance of palindromes and restriction-modification systems, *Molecular Biology*, **34**(2), 215–221 (2000).

[4] *GenBank duomenų bazė*. `http://www.ncbi.nlm.nih.gov/genomes/Complete.html`.

Statistical analysis on avoidance of palindromic sequences in genomes of microorganisms

T. Rekašius, A. Timinskas

The statistical analysis of oligonucleotide combinations and mathematical modelling of outgoing processes is very successfully used in the research. It present research we applied Markov chain model. Using the model few new methodologies for research of oligonucleotidic composition within DNA sequences we created.