Liet. matem. rink, **45***, spec. nr.*, 2005, 363–368

Priklausomybių modeliuotose DNR sekose tyrimas

Tomas REKAŠIUS (VGTU)
el. paštas: tomas.rekasius@fmst.vtu.lt

Reziumė. Daugelyje bioinformatikos publikacijų rašoma apie ilgas nukleotidų priklausomybes DNR sekose, kas turėtų reikšti, jog tokios sekos yra sudėtingos, hierarchinę struktūrą turinčios sistemos. Šiame darbe modeliavimo būdu parodyta, kad aukštos eilės sąveikos gali atsirasti sekai evoliucionuojant pagal labai paprastas taisykles.

1. Ilgos priklausomybės nukleotidų sekose

Ilgų priklausomybių tematika aktuali daugelyje sričių: fizikoje, ekonomikoje, lingvistikoje ir t.t. Bioinformatikoje ji atsiranda tiriant nukleotidų ir amino rūgščių sekų savybes. Paprastai sakoma, kad ilga priklausomybė (long-range correlation) yra viena iš kompleksiškų, hierarchinę struktūrą turinčių sistemų charakteristikų. Pavyzdžiui, tokios sistemos gali būti natūralių kalbų tekstai arba įvairių organizmų genomai.

Publikacijos apie koreliacijas DNR sekose prieštaringos. Li ir kt. bei Peng ir kt. 1992 m. straipsniuose [1, 2] tvirtinama, kad ilga priklausomybė būdinga nekoduojančiai genomo daliai. Tačiau būta ir kitokių rezultatų (pvz., Voss [3], Chatzidimitriou–Dreismann [4] ar Prabhu [5]) parodančių, jog ilga priklausomybė stebima taip pat ir koduojančiose sekose. Buldyrev ir kt. straipsnyje [6] buvo iškeltas uždavinys išspręsti šitą dviprasmišką situaciją ir dviem skirtingais metodais tiriant tuo metu *GenBank* turimas DNR sekas parodyta, kad koreliacinės savybės koduojančiose ir nekoduojančiose sekose skiriasi, o ilga priklausomybė labiau būdinga nekoduojančioms sekoms. Aišku, kad tokie prieštaringi rezultatai gaunami dėl nevienodų tyrimo metodų ir skirtingų DNR sekų požymių bei savybių išryškinimo [7].

Priimta laikyti, kad hieharchinės struktūros, mastelio simetrija ir panašios DNR kompleksiškumą apibūdinančios savybės atsirado dėl globalių genomo pertvarkymų evoliucijos eigoje. Tai gali būti pasikartojančios trumpos nukleotidų sekos, dubliuotos genus koduojančios sekos, ilgų fragmentų įterpimas ar pašalinimas. Tačiau toks aiškinimas nevisada teisingas. Daugumos bakterijų genai – unikalūs, labai nedaug ir tik specifinės paskirties genai yra dubliuojami. Aukštesnių organizmų genai dažnai būna pakartoti kelis kartus. Aišku, kad stambių struktūrų lygyje tokių organizmų genomai organizuoti kiek kitaip nei bakterijų. Tai atsispindi ir jų reagavimu į aplinką: bakterijoms mutavimas yra vienas iš būdų išlikti ir greitai prisitaikyti prie besikeičiančios aplinkos, o aukštesnieji organizmai linkę saugoti savo genetinę informaciją nepakitusią.

Kadangi genų (koduojančių sekų) paskirtis – koduoti baltymus, aišku, kad informacijos pernešimas „uždeda“ tam tikrus apribojimus koduojančiai nukleotidų sekai. Ilgą

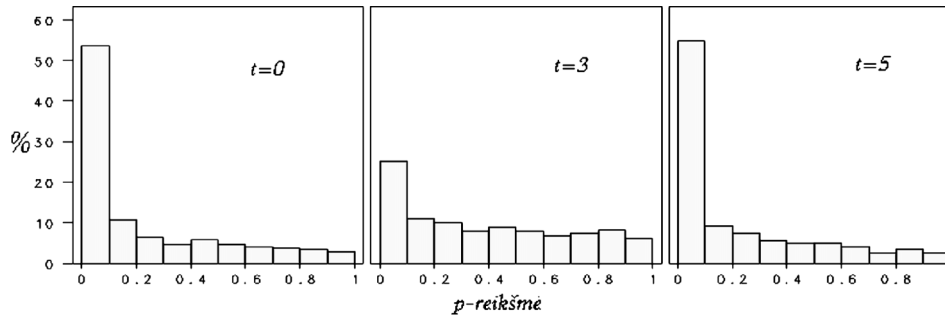
laiką manyta, kad nekoduojančios sekos yra nereikšmingos arba ne tokios reikšmingos. Tačiau dabar žinoma, kad nekoduojančios sekos yra ne mažiau svarbios ir susijusios su genų veiklos reguliavimo mechanizmais [8]. Jose yra daug kitų specialios paskirties sekų. Kita vertus nekoduojančioms sekoms nėra taip svarbu išlaikyti esamą būseną. Nežymūs pasikeitimai nesukelia jokių dramatiškų pasekmių, todėl jos nėra varžomos tokių stiprių apribojimų kaip koduojančios sekos.

2. DNR grandinės evoliuciniai modeliai

DNR sekų evoliucija ilgą laiką buvo modeliuojama kaip nepriklausomai vienas nuo kito mutuojančių nukleotidų evoliucija. Laikoma, kad bet kurio iš DNR sekos paimto vieno nukleotido mutavimas nepriklauso nuo kaimyninių nukleotidų ir laike sudaro Markovo grandinę. Priklausomai nuo to, kaip sudaroma tikimybių perėjimo matrica gaunami Jukes–Cantor (1969) arba Kimura (1980) modeliai, o juos apibendrina HKY (Hasegava, Kishino, Yano 1985) modelis.

Kad kaimyniniai nukleotidai nėra nepriklausomi, galima parodyti naudojant χ^2 požymių nepriklausomumo kriterijų. Kaip pavyzdys paimtos bakterijos *Bordetella bronchiseptica* nekoduojančios genomo sekos (*GenBank* duomenų bazė). Ilgesnėms kaip 200 nukleotidų sekoms sudarytos 4×4 dydžio gretimų nukleotidų dažnių lentelės ir paskaičiuotos χ^2 nepriklausomumo testo *p-reikšmės*. Tokios pat dažnių lentelės sudaromos nukleotidams, kuriuos skiria 3 ir 5 nukleotidų tarpas. Sekos ilgio apribojimas taikomas tam, kad nepritrūktų duomenų dažnių lentelėms užpildyti ir kriterijaus rezultatai būtų patikimesni.

Iš 1 pav. matyti, kad kriterijaus *p-reikšmės* pasiskirsčiusios netolygiai. Galima pastebėti, jog atmetamų hipotezių skaičius, priklausomai nuo atstumo tarp nukleotidų, keičiasi. Čia išsiskiria atvejis $t = 3$, kur *p-reikšmių* pasiskirstymas akivaizdžiai tolygesnis, kas reiškia, jog priklausomybės tarp nukleotidų silpnėja. Tai galima paaiškinti tuo, kad nemažą dalį nukleotidų priklausomybių sudaro priklausomybės kodonų (amino rūgštis koduojančių nukletoditų tripletai) viduje. Didelis *p-reikšmių* mažesnių už 0,05 skaičius reiškia, kad hipotezė apie nukletotidų nepriklausomumą daugeliu atveju atmetama, ir bendru atveju nukleotidai net ir nekoduojančiose genomo sekose yra priklausomi.



1 pav. χ^2 nepriklausomumo testo *p-reikšmių* histogramos.

Nepriklausomai vienas nuo kito mutuojančių nukleotidų DNR modeliai patogūs dėl savo paprastumo, jais lengva skaičiuoti sekos tikėtinumą ir sudarinėti filogenetinius medžius ar kitaip klasifikuoti organizmus. HKY modeliu generuojama DNR seka gana tiksliai atkartoja di-tri-tetranukleotidų sudėtį. Tačiau ilgesnių simbolių sekų dažniuose atsiranda tendencijos, kokių realiame genome nėra.

1976–1977 m. (G. J. Russell) biocheminiais tyrimais buvo parodyta, kad dinukleotidų dažnių rinkinys (*AA*, *AC*, ..., *TT*) skirtingose genomo dalyse išlieka panašus, gimininguų organizmų atitinkami dažniai taip pat panašūs, tačiau mažai gimininguų organizmų šitas dažnių rinkinys akivaizdžiai skiriasi. Dar vėliau (1995–1997 m., S. Karlin) tas pats buvo patvirtinta jau statistiniais tyrimais. Tokiu būdu parodyta, kad DNR sekose kaimyniniai nukleotidai gali daryti įtaką mutacijos tipui ir mutavimo intensyvumui, todėl dinukleotidų dažniai nėra tolygiai pasiskirstę: kai kurių jų „trūksta“, kitų yra „daugiau nei turėtų būti“. Toks „genomo parašas“ yra pakankamai stabili charakteristika ir gali būti naudojamas organizmų genetiniam atstumui (gimininguмui) nustatyti. Daugelyje organizmų (ypač aukštesniųjų) dinukleotidas *CG* (žymima *CpG*) yra retesnis nei turėtų būti. Egzistuoja biologinis mechanizmas, dėl kurio kopijuojant DNR grandinę nukleotidų pora *CG* klaidingai nuskaitoma ir mutuoja į *CA* arba *TG*.

Ir nors kaimyninių nukleotidų įtaka mutavimo procese suprasta gana seniai, tačiau taip vadinami kontekstiniai (context-dependent) modeliai pasirodė visai neseniai. Tokie yra Arndt ir kt. (2003), Hwang ir Green (2004), Siepel ir Haussler (2004) modeliai nukleotidams arba Jensen ir Pedersen (2000, 2001), Christensen ir kt. (2004) modeliai kodonų mutavimui. Jie konstruojami taip, kad stacionarus skirstinys būtų Markovo, be to visa nukleotidų seka yra viena grandinės būsena, o nukleotidų priklausomybės laike ir erdvėje gesimas eksponentinis [9].

3. Modelio aprašymas ir gauti rezultatai

DNR molekulė sudaryta iš dviejų komplementarių grandinių, kurią kiekvieną sudaro nukleotidų $\{A, C, G, T\}$ seka. Priešingose grandinėse nukleotidas A visada sudaro porą su T, ir C su G. Duomenų bazėse paprastai pateikiama tik viena DNR grandinė.

Egzistuoja keletas plačiai žinomų ir biologinę interpretaciją turinčių būdų, kaip nukleotidų $\{A, C, G, T\}$ aibę atvaizduoti į skaičių aibę, pvz. į aibę $\{0, 1\}$. Natūraliai nukleotidai skirstomi į dvi grupes: purinus (nukleotidai A ir G) ir pirimidinus (nukleotidai C ir T). Kitas dažnas variantas yra nukleotidus suskirstyti pagal jungčių, kuriomis jie jungiasi tarp DNR grandinių, skaičių: dviem jungtimis A su T ir trimis jungtimis C su G. Nukleotido būsenų aibės susiaurinimas gali turėti įtakos priklausomybėms tarp artimų sekos elementų, bet mažai įtakoja ilgos priklausomybės charakteristikas [10]. Tačiau modeliavimui šitas faktas nėra taip svarbu, nes nagrinėsime sekas, kurių elementai gali įgyti tik dvi būsenas.

Toliau aptarsime binarinės sekos $S = (x_0, x_1, x_2, \ldots, x_n, x_{n+1})$ evoliucionavimo modelį (Glauberio dinamika), kur $x_k \in \{0, 1\}$, sekos galai x_0 ir x_{n+1} yra fiksuoti, o nukleotido x_k mutavimas dar priklauso ir nuo gretimų nukleotidų x_{k-1} bei x_{k+1}. Tegu $t^{(k)} = (t_1^{(k)} t_2^{(k)} t_3^{(k)})$, $k = 1, 2, \ldots, n$ yra tripletas su viduriniu nukleotidu k-tojoje sekos S pozicijoje. Kadangi tripleto vidurinio nukleotido mutavimo tikimybė priklauso nuo dviejų kaimyninių nukleotidų, o kiekvienas iš jų įgyja tik dvi reikšmes, tai iš viso yra 2^3 skirtingų tripletų. Visus juos galima sunumeruoti nuo 000, 001 iki 111 (1 lentelė).

1 lentelė. Tripletai t, $\overline{t}$ ir konkretus perėjimo tikimybių rinkinys

t	000	001	010	011	100	101	110	111
$\overline{t}$	010	011	000	001	110	111	100	101
π_t	θ_2	θ_3	θ_0	θ_1	θ_6	θ_7	θ_4	θ_5
$\pi_t^{(1)}$	0,70	0,50	0,20	0,50	0,50	0,20	0,50	0,70
$\pi_t^{(2)}$	0,30	0,50	0,80	0,50	0,50	0,80	0,50	0,30

Tokio modelio perėjimo tikimybių matrica $P = \{p_{ij}\}$, $i, j = 0, 1, \ldots, 7$ nusakoma 8 parametrais ir lygi

$$p_{ij} = \begin{cases} \theta_j, & |i - j| = 2, \\ 1 - \theta_j, & i = j, \\ 0, & \text{kitais atvejais.} \end{cases} \tag{1}$$

3.1. Perėjimo tikimybės ir modeliavimas

Sekos S evoliucijos modeliavimui buvo pasirinkta keletas tripletų t perėjimo tikimybių rinkinių $\pi = \{\theta_i\}$. Šiuo atveju jie sudaromi laikantis tam tikrų taisyklių. Pvz., rinkinyje $\pi^{(1)}$ tripletai t_{010} ir t_{101} linkę likti esamose būsenose, tripletų porų t_{100} ir t_{001} bei t_{110} ir t_{011} perėjimo tikimybės dėl simetrijos vienodos, o tripletai t_{000} ir t_{111} linkę mutuoti. Tokiu būdu stengiamasi, kad sekoje nebūtų daug iš eilės einančių vienodų nukleotidų. Rinkinyje $\pi^{(2)}$ – atvirkščiai, tripletai t_{000} ir t_{111} linkę pasilikti savo būsenose, o tripletai t_{010} ir t_{101} linkę mutuoti (1 lentelė).

Pradinė seka S užpildoma atsitiktiniais 0 ir 1. Kiekvienos iteracijos metu vienas atsitiktinai pasirinktas tripletas $t^{(k)}$ su tam tikra tikimybe π_t pereina į kitą tripletą $\overline{t}^{(k)}$, nuo kurio skiriasi tik viduriniu nukleotidu $t_2^{(k)}$, arba lieka toje pačioje būsenoje (1 lentelė). Mutavimo pozicija k tolygiai pasiskirsčiusi per visą sekos S ilgį. Atliekama 10^7 iteracijų.

Tokį modelį galima interpretuoti kaip vienmatį Isingo modelio atvejį. Pagal Hamersley–Clifford teorema tam, kad nukleotidų seka S tenkintų 1-os eilės Markovo savybę būtina, kad tos sekos tikimybinis skirstinys butu Gibbso su neaukšesnėmis kaip antros eilės sąveikomis. Tai reiškia, kad

$$\frac{\pi_{\overline{t}}}{\pi_t} = \exp\left\{U(t) - U(\overline{t})\right\}, \tag{2}$$

$$U(t) = \psi_1(t_2) + \psi_2(t_1, t_2) + \psi_2(t_2, t_3), \quad \psi_1(0) = 0, \ \ \psi_2(0, 0) = 0. \tag{3}$$

Vadinasi, nukleotidų sekos stacionarusis skirstinys nusakomas 4 parametrais:

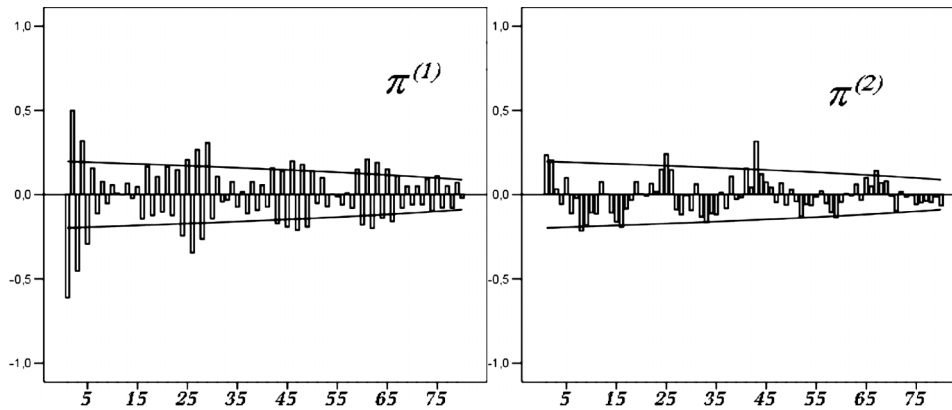
$$a = \psi_1(1), \ \ b_1 = \psi_2(0, 1), \ \ b_2 = \psi_2(1, 0), \ \ c = \psi_2(1, 1). \tag{4}$$

Reikia pabrėžti, kad priešingai nuo aukščiau paminėtų modelių, čia nenagrinėjamas *CpG* efektas.

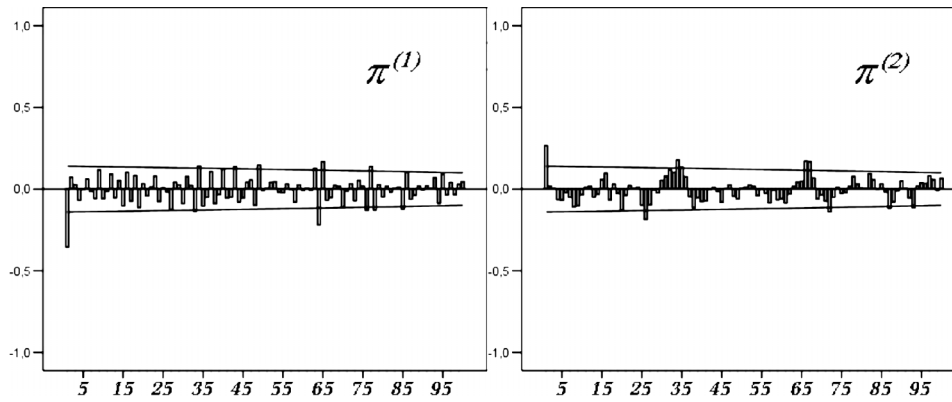
3.2. Koreliacijos modeliuotose sekose

Modeliuotos sekos S nukleotidų koreliacijos rodo, kad sąveikos yra aukštesnės nei antros eilės. 2 ir 3 pav. pavaizduotos 100 ir 200 nukleotidų sekų, generuotų su skirtingais perėjimo tikimybių rinkiniais, autokoreliacinės funkcijos $r(\tau)$ grafikai.

Iš jų matyti, kad nors ir silpnos, tačiau reikšmingos koreliacijos gali pasirodyti ir prie kelias dešimtis siekiančių τ reikšmių. Neturint koreliacinės funkcijos išraiškos, negalima pasakyti, ar ji yra integruojama ir ar sekoje stebimos tolimos sąveikos yra ilga priklausomybė pagal apibrėžimą. Tačiau pagrindinis tikslas buvo parodyti, kad tolimos sąveikos atsiranda nebūtinai pagal sudėtingas taisykles besivystančiose ir sudėtingą struktūrą turinčiose sekose.



2 pav. Autokoreliacijos funkcija 100 nukleotidų sekai, tikimybės $\pi^{(1)}$ ir $\pi^{(2)}$.



3 pav. Autokoreliacijos funkcija 200 nukleotidų sekai, tikimybės $\pi^{(1)}$ ir $\pi^{(2)}$.

Literatūra

1. W. Li, K. Kaneko, Long-range correlation and partial 1/f spectrum in a noncoding DNA sequence, *Europhys. Lett.*, **17**, 655–660 (1992).
2. C.-K. Peng, S. Buldyrev, A.L. Goldberg, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Long-range correlations in nucleotide sequences, *Nature*, **356**, 168–169 (1992).
3. R. Voss, Evolution of long-range fractal correlations and 1/f noise in DNA base sequences, *Phys. Rev. Lett*, **68**, 3805–3808 (1992).
4. C.A. Chatzidimitriou–Dreismann, D. Larhammar, Long-range correlations in DNA, *Nature*, **361**, 212–213 (1993).
5. V.V. Prabhu, J.M. Claverie, Correlations in intronless DNA, *Nature*, **359**, 782–782 (1992).
6. S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, H.E. Stanley long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis, *PRE*, **51**, 5084–5091 (1995).
7. Zu-Guo Yu, Vo Anh, Ka-Sing Lau, Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome, *Physica*, **A 301**(1–4), 351–361 (2001).
8. V. Rančelis, *Genetika*, Vilnius (2000).
9. J.L. Jensen, Context dependent DNA evolutionary models, *Research Reports*, **458** (2005).
10. O.V. Usatenko, V. A. Yampol'skii, Binary N-step Markov chains and long-range correlated systems, *Phys. Rev. Lett*, **90** (2003).

SUMMARY

T. Rekašius. Research on dependencies in simulated DNA sequences

Many publications in bio-informatics deal with long-range correlation of nucleotides in DNA sequences, which implies that such sequences are complicated systems with hierarchical structure. Applying the method of simulation, this work reveals that long-range correlation might develop in the course of evolution of a sequence under very general rules.

Keywords: context-dependent DNA model, long-range correlation.