Liet. matem. rink, **47**, *spec. nr.*, 2007, 21–28

Kai kurie matematikos uždaviniai genetikoje

Marijus RADAVIČIUS (MII), Tomas REKAŠIUS, Jurgita ŽIDANAVIČIŪTĖ (VGTU)
el. paštas: mrad@ktl.mii.lt

1. Įvadas

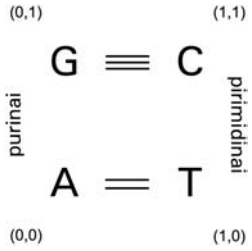
Genetika ir ypač su ja susijęs bioinformatikos mokslas šiuo metu labai sparčiai vystosi. Dėl milžiniško duomenų ir informacijos kiekio šis vystymasis būtų neįmanomas be plataus kompiuterių panaudojimo, naujų algoritmų, o tuo pačiu ir naujų matematinių metodų bei modelių. Šioje apžvalgoje trumpai supažindinama su keliais aktualiais matematiniais uždaviniais, iškylančiais apdorojant genetinius duomenis. Uždavinių pasirinkimą pirmiausia nulėmė autorių moksliniai interesai, tačiau atsižvelgta ir į Lietuvos genetikos ir bioinformatikos mokslų vystymosi artimiausias perspektyvas. Netrukus turėsime įrangą genetiniams (DNR ir baltymų) lustams gaminti. Taigi, šios technologijos pagrindu gautų duomenų apdorojimas jau aktualus ir Lietuvoje. 3.1 skyrelyje pateikta viena iš galimų čia iškylančių uždavinių formuluočių. Neinformatyvių genetinių sekų, kitaip tariant, „genetinio triukšmo“ problema aptariama 3.2 skyrelyje, toliau trumpai aprašytas sekų sulyginimo (alignment) uždavinys. Paskutinis skyrius skirtas genetinių sekų vizualizacijos problemai.

2. Genetinės sekos

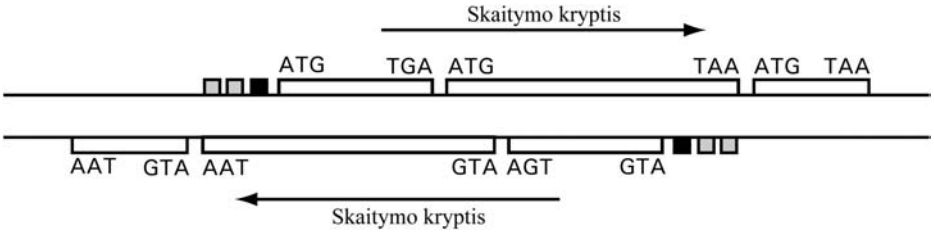
Šiame skyrelyje pateiksime kelis genetikos mokslo faktus, kuriais toliau remsimės. Apsiribosime tik nukleotidų (DNR) sekomis, nors betarpiškai taikymams žymiai aktualesnės baltymų sekos. Formaliai matematinei analizei tarp jų skirtumas mažas: baltymų sekos yra sudarytos iš 20 „raidžių“ alfabeto, o DNR sekos turi tik 4 skirtingas „raides“: jas sudaro nukleotidai adeninas (A), citozinas (C), guaninas (G) ir timinas (T) (arba jo pakaitalas uracilas (U) RNR molekulėje). Kiekvienam iš jų būdingos dvi savybės: *jungčių skaičius* ir *molekulės tipas*. Nukleotidai A ir T turi 2 jungtis, o C ir G turi 3 jungtis, A ir G yra *purinai*, o C ir T yra *pirimidinai*. Nukleotidai dviguboje DNR grandinėje sudaro komplementarias poras: A $\leftrightarrow$ T ir C $\leftrightarrow$ G.

DNR grandinėje nukleotidų sekos skirstomos į *koduojančias (genus)*, *nekoduojančias* ir *reguliavimo* (genų raišką reguliuojančias) sekas. Genas yra baltymą koduojanti nukleotidų seka, o DNR sekos tarp genų vadinamos nekoduojančiomis genomo sekomis. Genai išsidėstę abiejose DNR grandinės pusėse, gali būti pavieniai ar grupėse, gali persikloti vienas su kitu. Tokia organizmo savybes nusakanti nukleotidų sekų visuma vadinama *genomu*.

Kiekviena geną koduojanti seka turi START kodoną – trijų nukleotidų „žodį“ ATG, o geno pabaigą žymi vienas iš trijų STOP kodonų: TAA, TAG arba TGA.



1 pav. Nukleotidų savybės.



2 pav. DNR grandinė ir jos sritys: koduojančios, nekoduojančios ir reguliuojančios.

Genomų „statistika“: paprasčiausių organizmų – bakterijų – genome yra apie 0,5–10·10^6 nukleotidų, kuriame yra keletas tūkstančių genų. Žmogaus genome yra apie $3{,}12 \cdot 10^7$ nukleotidų ir apie 30000 genų, o nekoduojančios sekos užima $\approx 97\%$ viso genomo.

3. Aktualūs uždaviniai

Trumpai aptarsime 3 uždavinius: genų raiškos analizė, genetinio triukšmo problema ir genetinių sekų sulyginimas. Pirmasis ir trečiasis uždavinys ypač aktualūs taikymams medicinoje. Genetinių sekų vizualizavimo problemoms skirtas atskiras skyrius.

3.1. Genų raiškos analizė

Šis uždavinys iškilo visų pirma ryšium su nauja *DNR ir baltymų lustų* gamybos technologija, suklestėjusia maždaug prieš 10 metų ir skirta genų raiškos (gene expression), jų aktyvumo tyrimui. Genų raiškos duomenų analizei naudojami įvairūs (matematiniai) metodai: regresinė ir dispersinė analizė, klasterizacija ir klasifikacija, laiko eilučių modeliai, neryškių aibių (fuzzy sets) metodologija ir kiti.

Vienas iš paprasčiausių uždavinio formulavimų atrodytų taip. Turime dvi imtis:

$$\text{n.v.p. } X_1, \dots, X_N \ \sim \mathcal{N}_n(\mu_1, V_1) \ \longleftrightarrow \text{ sergantys vėžiu,}$$
$$\text{n.v.p. } Y_1, \dots, Y_M \ \sim \mathcal{N}_n(\mu_2, V_2) \ \longleftrightarrow \text{ sveiki.}$$

Reikia patikrinti hipotezę H_0: $\mu_1 = \mu_2$ ir nustatyti tas vidurkių μ_1 ir μ_2 komponentes, kurios nelygios, jeigu H_0 atmetama. Tai klasikinis dispersinės analizės uždavinys,

tačiau realiuose tyrimuose, pavyzdžiui, $N = M = 100$, $n = 1500$ ir skiriasi tik 2 ar 3 vidurkių μ_1 ir μ_2 komponentės.

Galimi supaprastinimai: V_i yra diagonalinės, $V_1 = V_2$ arba $V_i = \sigma I_n$ (pastaroji prielaida nerealistiška). Tačiau aišku, kad net ir šiomis prielaidomis supaprastinus uždavinį, jis reikalauja iš principo naujų metodų. Apie genų raiškos analizę rašoma labai daug, plačiau apie tai galima pasiskaityti apžvalginiuose straipsniuose (Wolfgang Huber (2003), Daxin Jiang and Aidong Zhang (2003)).

3.2. Neinformatyvios genetinės sekos: *„genetinis triukšmas“*

Norint atsakyti į klausimą, ar duotoje DNR sekoje yra biologiškai svarbios informacijos, kur ji yra, kaip užkoduota ir kaip ją išgauti, visų pirma reikia atsakyti į klausimą, kaip „atrodo“ simbolių seka, kurioje tos informacijos nėra.

Problema: Ką reiškia „neinformatyvi genetinė seka“? Kaip apibrėžti „genetinį triukšmą“, t.y. seką, kuri neturi genetiškai ar biologiškai reikšmingos informacijos?

Šis, atrodytų, gana teorinis uždavinys turi tiesioginį ryšį su taikymais.

1. DNR sekos fragmento informatyvumas paprastai nustatomas pagal tai, kiek jis tam tikroje metrikoje skiriasi nuo „neinformatyvaus“ tos sekos fragmento. Pastarasis paprastai gaunamas kaip atsitiktinis tiriamo DNR fragmento perstatinys. Ar tai tinkamas būdas generuoti neinformatyvią seką?

2. DNR sekos tikimybinis modelis nusakomas jos evoliucijos laike modeliu, o šis yra *filogenetinių medžių* sudarymo pagrindas. Filogenetiniai medžiai – tai biologinių rūšių atsiradimo ir raidos pavaizdavimas grafu. Biologinės rūšys yra pavaizduotos kaip to grafo viršūnės ir kelias grafe tarp dviejų rūšių tuo ilgesnis, kuo ilgesnis pagal priimtą evoliucijos modelį (vidutinis) laikas, reikalingas vienai iš jų transformuotis į kitą. Metodas iš esmės remiasi evoliucijos apverčiamumo prielaida, dėl to filogenetinį grafą galima laikyti neorientuotu ir kelio ilgis tarp viršūnių nepriklauso nuo to, kuri iš jų yra pradinė.

Evoliucija laike ir erdvėje. Natūralu laikyti, kad DNR sekos evoliuciją laike aprašo diskretaus laiko homogeninė Markovo grandinė

$$X(t) = X_{[1,n]}(t) := \{X_l(t),\ l = 1, \dots, n\} \in \mathcal{A}^n, \quad t \in T.$$

Čia $T = \{0, 1, \dots\}$ yra laikas, $\mathcal{A}$ yra kiekvienos komponentės $x_l(t)$ galimų būsenų aibė (alfabetas); DNR sekoms $\mathcal{A} = \{A, C, G, T\}$. Pažymėkime raide Π Markovo grandinės $\{X(t),\ t \in T\}$ perėjimo tikimybių matricą. Patogu naudoti tokią terminologiją: DNR sekos *evoliuciją laike* aprašo perėjimas

$$X(t) \xrightarrow{\Pi} X(t+1),$$

o jos *evoliuciją erdvėje* fiksuotu laiko momentu t nusako taisyklė, pagal kurią prie duotos k ilgio sekos $X_{[1,k]}(t)$ prisijungia naujas elementas $X_{k+1}(t)$. Evoliucija erdvėje tam tikra prasme imituoja DNR grandinės susidarymą. *Stacionariu* atveju atsitiktinės sekos $X(t)$ skirstinys nepriklauso nuo t. Pažymėkime jį Q, ir tegu X yra pagal jį

pasiskirsčiusi atsitiktinė seka iš $\mathcal{A}^n$. Tuomet galima nagrinėti sekos $X = (X_l,\ l = 1, \ldots, n)$ *evoliuciją erdvėje*: $X_l \to X_{l+1}$.

Duomenys apie DNR sekos evoliuciją laike yra retenybė, paprastai turime stebėjimus tik *vienu laiko momentu*, t.y. tik vieną sekos X realizaciją.

Uždavinys. Iš Markovo evoliucijos laike stacionaraus skirstinio atstatyti jos perėjimo matricą arba tos matricos savybes.

Šis uždavinys yra *nekorektiškasis* (ill-posed). Bendru atveju jis turi (be galo) daug sprendinių, ir sprendinio radimui reikia papildomų prielaidų arba reguliarizavimo metodų. Kai seka $\{X(t),\ t \in T\}$ apverčiama ir Π priklauso tik nuo (santykinai) mažo nežinomų parametrų skaičiaus, tai sprendinį nesunku gauti iš *lokalaus balanso lygties*: $Q_x \Pi_{xy} = \Pi_{xy} Q_y\ \forall x, y \in \mathcal{A}^n$.

Neinformatyvios genetinės sekos matematinis modelis turėtų būti suderintas su žinomais genetikos mokslo faktais. Siūlomas „genetinio triukšmo“ apibrėžimas remiasi tokiomis biologinėmis prielaidomis.

1. *Nekoduojantys* DNR fragmentai (tarpai) tiesiogiai neįtakoja (primityvios) biologinės rūšies (ir individo) išlikimo. Vadinasi, jų savybes sąlygoja daugiau negyvosios gamtos dėsningumai, o ne biologinė atranka. Jie nėra tokie reikšmingi biologiškai. Vadinasi, neinformatyvių sekų reiktų „ieškoti“ nekoduojančiose DNR sekose.

2. Nekoduojančių DNR fragmentų evoliucija laike turi *paprastą* struktūrą ir ją veikia tik *lokalūs* atsitiktiniai veiksniai. Pavyzdžiui, laikoma, kad nėra DNR fragmentų įterpimų ar išmetimų, o vyksta tik nukleotidų pasikeitimas, ir kiekvieno nukleotido pasikeitimo tikimybė priklauso tik nuo jo artimiausiųjų kaimynų (paprasta lokali evoliucija). Nors ši prielaida yra gana natūrali ir argumentuota, tačiau ji nėra nei būtina, nei pilnai pagrįsta.

3. Tariama, kad nagrinėjamos DNR sekos evoliucijos laike procesas yra nusistovėjęs, *stacionarus*. Nestacionariu atveju, kai viskas keičiasi, bet kuri genetinės sekos vieta gali pasidaryti informatyvia.

Nekoduojančio DNR fragmento nukleotidų paprastos lokalios evoliucijos laike stacionarųjį skirstinį natūralu traktuoti kaip „neinformatyvų“.

APIBRĖŽIMAS. Tegu DNR sekos $X(t) \in \mathcal{A}^n$ evoliucija laike $\{X(t), t \in T\}$ yra (diskretaus laiko) homogeniška Markovo grandinė su (paprastos struktūros ir lokalios sąveikos) perėjimo tikimybių matrica Π. Tarkime, kad egzistuoja jos stacionarusis skirstinys Q aibėje $\mathcal{A}^n$. Tada atsitiktinė seka X su skirstiniu Q vadinama *neinformatyviaja* arba tiesiog *genetiniu triukšmu*.

TEIGINYS (žr., pvz., Jensen (2005); laikas tolydus). *Tarkime, kad DNR evoliucija laike $X(t), t \in T$, yra apverčiama Markovo grandinė ir kiekvieno nukleotido mutavimo tikimybės priklauso tik nuo jo gretimų nukleotidų ir nepriklauso nuo jo vietos. Tada jos stacionarusis skirstinys Q aprašo 1-os eilės Markovo grandinę X (erdvėje).*

Išvada. Iš ergodinės teoremos Markovo grandinėms išplaukia, kad sekos X narių tarpusavio priklausomybė gęsta eksponentiškai atstumo tarp jų pozicijų sekoje atžvilgiu.

Ši išvada prieštarauja žinomiems stilizuotiems faktams (stylized facts). Skaitlingi (empiriniai) tyrimai rodo, kad tarp nukleotidų yra „ilga priklausomybė“ tiek koduojančiose, tiek ir nekoduojančiose DNR sekos srityse ir pastarosiose ji išreikšta labiau. Hursto eksponentė, nusakanti ilgos priklausomybės gesimo greitį, įvertinama naudojant R/S analizę. Tirti ilgas priklausomybes ypač mėgsta fizikai (žr., pavyzdžiui, Yu, Anh ir Lau (2001), Usatenko ir Yampol'skii (2003) ir nuorodas šiuose straipsniuose).

Gal mutacijos priklauso ne vien nuo artimiausiųjų kaimynų? O gal evoliucija laike $\{X(t), t \in T\}$ neapverčiama? Kol kas atsakymas į šį klausimą autoriams nežinomas, tačiau jie linkę laikytis antrosios prielaidos.

3.3. Sekų sulyginimas

Genetikoje ir medicinoje labai aktualus uždavinys: pagal baltymo kodą atstatyti jo biologines funkcijas.

Genetikos ir bioinformatikos aksioma teigia, kad panašios genetinės (baltymų) sekos turi panašią (erdvinę baltymų) molekulės struktūrą, o panašios struktūros sąlygoja ir panašią paskirtį bei panašias biologines funkcijas. Ši aksioma greičiau uždavinio formulavimas: kaip išmatuoti sekų panašumą pagal jų biologinę prasmę?

Dažniausiai (bio)informatikoje taikomas *Levenshtein'o atstumas* bei jo modifikacijos ir apibendrinimai (Levenshtein (1965), Navaro (2001), Deonier *et al.* (2005)).

Levenshtein'o atstumas $d(x, y)$ tarp sekų $x \in \mathcal{A}^m$ ir $y \in \mathcal{A}^n$ (nebūtinai vienodo ilgio!) apibrėžiamas kaip minimalus redagavimo (edit) operacijų, kurias reikia atlikti, norint iš sekos x gauti seką y, skaičius. Galimos operacijos: simbolio įterpimas, ištrynimas arba pakeitimas kitu simboliu. Todėl jis dar dažnai vadinamas redagavimo atstumu (edit distance). DNR sekoms šios operacijos atitinka nukleotidų mutacijas. Jeigu lyginamos sekos yra vienodo ilgio ir jose galimi tik simbolių pakeitimai, tai Levenshtein'o atstumas sutampa su *Hamming'o atstumu*, kuris yra lygus tų sekų, lyginant jas paelemenčiui, nesutampančių simbolių skaičiui.

Genetikoje ir bioinformatikoje skaičiuojant atstumą tarp baltymų ar DNR sekų natūralu skirtingiems simboliams priskirti skirtingus svorius. Garsiojo *BLAST* algoritmo versijos remiasi Levenshtein'o tipo atstumu su specialiai baltymams sudaryta svorių matrica BLOSUM (Henikoff and Henikoff (1992)).

Toliau aptarsime kitą uždavinį, kuriame atstumas tarp simbolių sekų vaidins svarbų vaidmenį.

4. Genomo parašas

Ar galima pamatyti visą genomą?

Atsižvelgiant į dvi nukleotidų savybes, jungčių skaičių ir molekulės tipą, natūralu raides pervesti į skaičius, apibrėžiant abipus vienareikšmį atvaizdavimą $Z\colon \mathcal{A} \to \mathcal{A}_1 \times \mathcal{A}_1$, $\mathcal{A}_1 := \{0, 1\}$,

$$\big\{Z(A), Z(C), Z(G), Z(T)\big\} = \big\{(0,0), (1,1), (0,1), (1,0)\big\}. \tag{1}$$

Tuomet nukleotidų seką $s = (s_1, \ldots, s_n) \in \mathcal{A}^n$ galima sutapatinti su dvimate 0 ir 1 seka $Z(s) = (Z(s_1), \ldots, Z(s_n))$, o į ją galima žiūrėti kaip atitinkamo taško vienetiniame kvadrate užrašymą dvejetainėje sistemoje. Duotam $k \in \mathbf{N}$ apibrėžkime

$U_k\colon \mathcal{A}_1^n \to [0,1]$, $U := U_\infty$,

$$U_k(z) = \sum_{j=1}^{k\wedge n} z_{n+1-j} \cdot 2^{-j}, \quad z = (z_1, \ldots, z_n) \in \mathcal{A}_1^n. \tag{2}$$

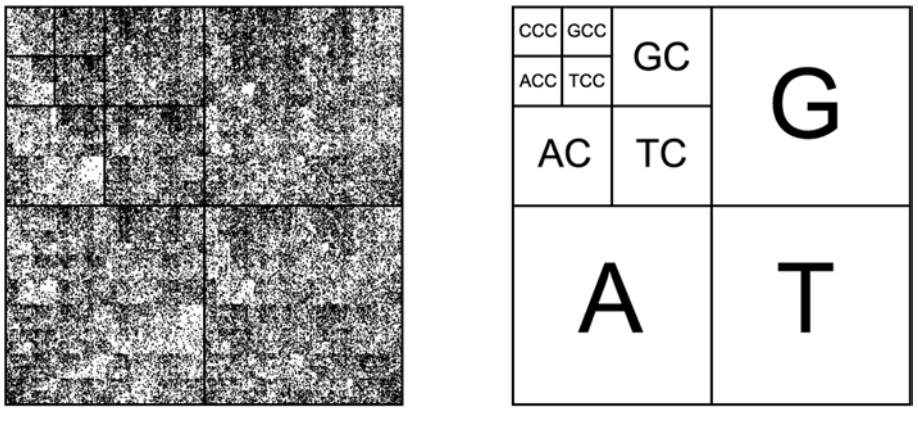
Tuomet $U(Z(s))$ yra tik taškas vienetiniame kvadrate $[0,1] \times [0,1]$, bet *vienareikšmiškai nusako visą seką* $s \in \mathcal{A}^n$. Sekos $s, r \in \mathcal{A}^n, n > 10$, kurių pirmieji 10 narių sutampa, vizualiai neatskiriamos, nes atstumas tarp $U(Z(s))$ ir $U(Z(r))$ yra mažesnis už 10^{-3}. Formaliai sutapatinkime seką $s \in \mathcal{A}^n$ su jos pratęsimu iki begalinės sekos $s^{(A)} = \{\ldots, s_{-1}^{(A)}, s_0^{(A)}, s_1^{(A)}, \ldots\} \in \mathcal{A}^\infty$, kuriai $s_j^{(A)} =$'A', kai $j < 1$ arba $j > n$, ir sudarykime dvimatę „laiko eilutę“

$$x(t) = x(t|s,k) = U_k\big(Z(L^t s)\big), \quad t = 0, 1, \ldots. \tag{3}$$

Čia L^t žymi postūmio operatorių $\big(L^t r\big)_j = r_{j-t}$, $r \in \mathcal{A}^\infty$. Gauname taškų vienetiniame kvadrate seką $X = X(s,k) = \big\{x(t|s,k),\ t = 0, 1, \ldots, n\big\}$, kuri vadinama *genomo s parašu* (genome signature, Russel *et al.* (1976), Jeffrey (1990)). Formaliai X apibrėžime reiktų imti $k = \infty$, bet $X(s,\infty) \approx X(s,k)$ pakankamai dideliems k, sakykim $k > 10$.

3 pav. aiškiai matosi fraktalinė struktūra. Atstumas vienetiniame kvadrate tarp taškų, atitinkančių sekas 'AAAAAAAAAAC' ir 'CCCCCCCCCCA', mažesnis už 10^{-3}, o sekoms 'CAAAAAAAAAA' ir 'ACCCCCCCCCC' jis didesnis už $\sqrt{2}(1 - 2 \cdot 10^{-3})$. Tarkime, kad $d(s,r)$ yra duotas (natūralus) atstumas tarp sekų $s, r \in \mathcal{A}^n$. Tuomet prasminga reikalauti, kad atvaizdavimas $U \circ Z\colon \mathcal{A}^n \to [0,1]^2$ ir jam atvirkštinis būtų tolydūs. Tačiau taip nėra, ir tai pagrindinis „genomo parašo“ trūkumas.

Šią problemą galima suvesti į tokį žinomą matematinį uždavinį. Tegu duoti: atstumų (skirtingumo) matas $d(z,u)$, $z, u \in \mathcal{A}_1^k$, klasė Ψ (paprastų) atvaizdavimų $\psi\colon \mathcal{A}_1^k \to \mathbf{R}$ bei nuostolių funkcija $\ell\big(\cdot,\cdot\big)$. Paprastai $\ell\big(a,b\big) = |a-b|$ arba $\ell\big(a,b\big) = (a-b)^2$.



3 pav. Genomo parašas.

Reikia rasti uždavinio

$$\sum_{z,u\in\mathcal{A}_1^k} \ell\big(|\psi(z)-\psi(u)|, d(z,u)\big) \longrightarrow \min_{\psi\in\Psi} \tag{4}$$

sprendinį. Apytikriai imant, reikia rasti tokį $\psi^* \in \Psi$, kad $|\psi^*(z) - \psi^*(u)| \approx d(z,u)\ \forall\ u, z \in \mathcal{A}_1^k$. Statistikoje tai yra *daugiamačio mastelio parinkimo* (multidimensional scaling) uždavinys (klasikiniu atveju $\ell(a,b) = (a-b)^2$, o Ψ yra tiesiniai atvaizdavimai). Informatikoje (computer science) atitinkamas uždavinys dažnai formuluojamas truputį kitaip: kaip apytikslės įdėties (approximate embending) teoremos. Pačios reikšmės $\psi^*(z),\ z \in \mathcal{A}_1^k$, nėra labai svarbios, svarbesnė yra aibės $\mathcal{A}_1^k$ elementų tvarka, kurią nusako funkcija ψ^*. Patogu apibrėžti naują transformaciją G_k taip, kad ji išlaikytų ψ^* nusakytą $\mathcal{A}_1^k$ elementų tvarką ir reikšmės $G_k(z),\ z \in \mathcal{A}_1^k$, būtų tolygiai išsidėsčiusios intervale $[0,1]$. Gauname taškų vienetiniame kvadrate seką

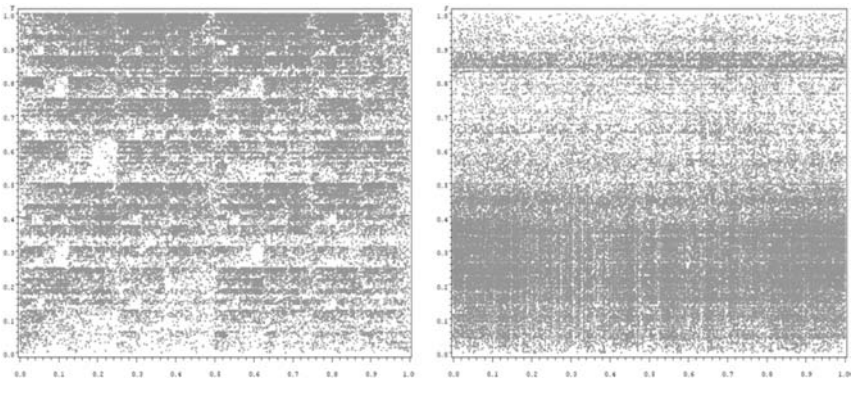
$$Y = Y(s,k) = \Big\{ G_k\Big(Z\big(L^t s\big)\Big),\ t = 0,1,\ldots,n \Big\}, \quad s \in \mathcal{A}^n,$$

kurią vadinsime *modifikuotu sekos s parašu*.

Problema. Kaip parinkti atstumą $d(z,u)$? Deja, ir Hamming'o ir Levenshtein'o atstumas šiuo atveju netinka, nes redagavimo operacijų suma (beveik) nepriklauso nuo redaguojamų vietų konteksto (gretimų simbolių).

Diskretus Sobolevo normos analogas. Tarkime, kad realaus kintamojo t funkcija f yra m kartų diferencijuojama ir jos išvestinės $f^{(j)}(t) = \partial^j f(t)/\partial t^j$ kvadratu sumuojamos. Sobolevo tipo norma

$$\|f\|^2 := \sum_{j=0}^{m} w_j \int \big|f^{(j)}\big|^2.$$



4 pav. Genomo parašo ir modifikuoto genomo parašo palyginimas.

Čia $w_0 > 0$, $w_m > 0$, $w_j \geqslant 0$, $j = \overline{1, (m-1)}$, yra duoti svoriai. Diskretus šios normos analogas $\|x\|$ sekai $x \in \mathbf{R}^k$ gaunamas išvestines pakeičiant sekos x narių atitinkamos eilės skirtumais. Apibrėžkime

$$d(z, u) := \|z - u\|, \quad z, u \in \mathcal{A}_1^k.$$

Modifikuoto genomo parašo iliustracija pateikta 4 pav. Jis yra žymiai glodesnis ir turi daug aiškesnę struktūrą negu tradicinis genomo parašas.

Literatūra

1. R.C. Deonier, S. Tavare, M.S. Waterman, *Computational Genome Analysis*, Springer (2005).
2. R. Durban, S.R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis – Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge (1998).
3. S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci. U S A*, **89**(22), 10915–10919 (1992).
4. W. Huber, A.V. Heydebreck, M. Vingron, Analysis of microarray gene expression data, In: *Handbook of Statistical Genetics*, 2nd edition, vol. 1, Wiley (2003), 162–187.
5. H.J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Res.*, **18**, 2163–2170 (1990).
6. J.L. Jensen, *Context Dependent DNA Evoliutionary Models, Research Reports 458*, Department of Mathematical Sciences, University of Aarhus (2005).
7. D. Jiang, A. Zhang, Cluster Analysis for Gene Expression Data: A Survey, *Technical Report 2002-06*, State University of New Your at Buffalo (2002).
8. V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, **10**, 707–710 (1966).
9. G. Navaro, A guided tour to approximate string matching, *ACM Computing Surveys*, **33**(1), 31–88 (2001).
10. T. Rekašius, Priklausomybių modeliuotose DNR sekose tyrimas, *Liet. matem. rink.*, **45**, spec. nr., 363–368 (2005).
11. O.V. Usatenko, V.A. Yampolskii, Binary N-step Markov chains and long-range correlated systems, *Phys Rev Lett*, **90**, id. 110601 (2003).
12. Z.-G. Yu, V. Anh, K.-S. Lau, Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome, *Physica A*, **301**(1–4), 351–361 (2001).

SUMMARY

M. Radavičius, T. Rekašius, J. Židanavičiūtė. Some mathematical problems in genetics

After an introduction to genetic basics three problems are briefly discussed: microarray data analysis, a definition of noninformative DNA sequence, and genetic sequence alignment. More atention is paid to DNA sequence visualization and regularization of a genome signature.

Keywords: alignment, discrete Sobolev norm, DNA sequence, genetic noise, Markov chain, microarray, multidimensional scaling, noninformative sequence, regularization, stationary distribution.