Liet. matem. rink, **46***, spec. nr.*, 2006, 390–396

Nukleotidų sekų vizualizacija

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

Reziumė. Šiame darbe apžvelgiami genetinių sekų ir jų tyrimų rezultatų vizualizavimo uždaviniai ir su jais susijusios problemos. Pateikiamas binarinių kodų išrikiavimo ant tiesės metodas ir juo paremta „genomo parašo“ modifikacija.

1. Genetinių sekų vizualizavimo uždavinys

Per paskutinius keliasdešimt metų fundamentalūs atradimai molekulinės biologijos srityje padarė ją centrine biologijos mokslų disciplina. Ankstesnis dėmesys nuo vieno konkretaus geno identifikavimo persikelia prie didelių galimybių, kurios tapo įmanomos iššifravus ištisus organizmų genomus. Tai savo ruožtu atvėrė kelią naujoms, taip vadinamoms *post-genomics* eros technologijoms. Jos dažnai remiasi kompiuterine viso genomo analize, t.y. bioinformatika.

Genetinių sekų duomenų bazėse, tokiose kaip *GenBank*, *DDBJ* ar *EMBL*, nukleotidų ir amino rūgščių sekų nuolat daugėja ir šie skaičiai milžiniški. 2006-ųjų metų pradžioje *GenBank* turėjo apytiksliai 59 750 386 305 nukleotidų iš 54 584 635 sekų. Šiuo metu žinomi daugiau nei 2000 virusų ir 1100 bakterijų genomai, taip pat nuolat vykdomi projektai aukštesnių gyvūnų genomams iššifruoti. Net paprasčiausios gyvybės formos – viruso genomas gali būti labai didelis ir viršyti $3,5 \cdot 10^5$ nukleotidų. Bakterijų genomai turi maždaug nuo $0,5 \cdot 10^6$ iki $10 \cdot 10^6$ nukleotidų. Žmogaus genomas yra apie $3,12 \cdot 10^7$ nukleotidų ilgio ir turi apie 30 000 genų. Atvaizduoti tokios apimties duomenis ar jų statistikas didelė problema. Beveik visos bioinformatikos duomenų bazės turi nukleo ar amino rūgščių sekų vizualizacijos priemones, tyrimams palengvinti sudarinėjami detalūs viso genomo „žemėlapiai“. Naudojantis jais galima detaliai pažvelgtį į nedidelį sekos fragmentą, tačiau negalima pamatyti visumos, negalima pamatyti tik tai sekai būdingų savybių ir vizualiai ją atskirti nuo kitomis savybėmis pasižyminčių sekų. Trumpos sekos (pvz., genai ar baltymai) dažnai lyginamos tarpusavyje, tačiau kada jų yra daug, net palyginimo rezultatai yra sunkiai aprėpiami [1].

1.1. „Genomo parašas“

Publikacijose apie priklausomybes nukleotidų sekose dažnai minimas fraktalinis jų pobūdis (mastelio simetrija, lėtai gęstanti koreliacinė funkcija) [2,3] ir šią jų savybę galima išnaudoti. Vienas iš būdų vizualizuoti ilgas genetines sekas – naudoti iteracinių funkcijų sistemas (IFS). Paprastai tokiu būdu gaunamas sekos vaizdas yra fraktalas.

Pagrindinė idėja paprasta: fraktalus generuojančiuose algoritmuose atsitiktinius dydžius pakeisti DNR sekos nukleotidais A, C, G, T. Gautas vaizdas ir yra DNR vizualizacija [4]. Tokie DNR generuoti fraktalai atrodo visiškai kitaip nei nepriklausomų atsitiktinių dydžių generuoti fraktalai. Užduotis – gautus vaizdus padaryti lengvai interpretuojamus, o skirtumus tarp jų informatyviais.

Atskiras IFS atvejis yra taip vadinamas „*chaoso žaidimas*“. Plokštumoje parenkami taisyklingo m-kampio viršūnes sudarantys taškai A_i, $i = \overline{1,m}$. Kiekvienai viršūnei priskirkime jos išrinkimo tikimybę π_i, $\sum_i \pi_i = 1$. Nagrinėjame atvejį, kada $\pi_i = 1/m$. Pradiniu tašku pasirenkama bet kuri daugiakampio viršūnė. Nuo šio taško einama link kitos, atsitiktinai su tikimybėmis π_i pasirinktos viršūnės, atkarpos viduryje dedamas taškas ir procesas kartojamas iš naujo. Paprasčiausiu „*chaoso žaidimo*“ atveju, kada $m = 3$ gauta taškų aibė vadinama *Sierpinskio nėriniu*. Tokį algoritmą galima užrašyti iteracinėmis funkcijomis, kuriomis perskaičiuojamos sekančio žingsnio taško koordinatės. Plokštumoje tai tiesinių lygčių pora:

$$x_{i+1} = ax_i + b, \quad y_{i+1} = cy_i + d, \quad a, b, c, d \in R. \tag{1}$$

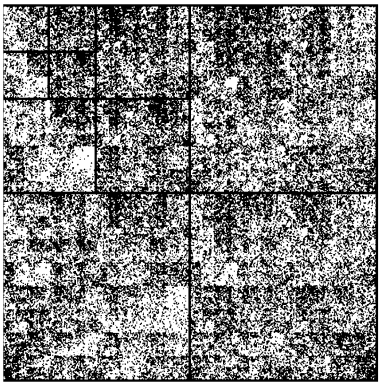
Keturių taškų „chaoso žaidimo“ atveju reikalingos keturios poros tokių funkcijų. Tegu tie taškai būna vienetinio kvadrato viršūnės A, G, C ir T su atitinkamomis koordinatėmis $(0, 0)$, $(1, 1)$, $(0, 1)$ ir $(1, 0)$. Lygčių koeficientus kompaktiškai galima užrašyti matriciniu pavidalu (1 lentelė).

Jei visų viršūnių išrinkimo tikimybės vienodos ir lygios $\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$, tada 4 viršūnių „chaoso žaidimas“ kvadratą padengia tolygiai. Kiekvienai kvadrato viršūnei priskirkime po atitinkamą nukleotidą. Tada kiekviename žingsnyje viršūnė parenkama pagal DNR sekoje išsidėsčiusių nukleotidų tvarką. Kadangi DNR grandinė nėra atsitiktinių ir tolygiai pasiskirsčiusių nukleotidų seka, kvadratas užpildomas netolygiai. Neatsitiktinis nukleotidų išsidėstymas DNR sekoje gali būti suprantamas kaip jos struktūra, kuri ir atsispindi kvadrato užpildyme. Plačiau toks nukleotidų sekų atvaizdavimo būdas aptariamas H.J. Jeffrey straipsnyje [5].

Dar 1976–1977 m. (G. J. Russell) biocheminiais tyrimais buvo parodyta, kad kaimyninių nukleotidų porų (AA, AC,..., TT) santykinių dažnių rinkinys tiek visame genome, tiek skirtingose jo dalyse išlieka labai panašus. Tokia genomo savybė pavadinta „genomo parašu“ yra pakankamai stabili ir tinkama DNR sekų identifikavimui. DNR generuotas „chaoso žaidimas“ yra to paties „genomo parašo“ realizacija ir būdas grafiškai atvaizduoti nukleotidų sekos trumpų „žodžių“ dažnių rinkinį į plokštumą. Gautą kvadratą padalinus į 4, 16, 64 ir bendru atveju į 4^n vienodo ploto mažesnių kvadratų, taškų skaičius kiekviename iš jų atitiks konkretaus n nukleotidų

1 lentelė

Viršūnė	a	b	c	d	Tikimybė π_i
A	0,5	0	0,5	0	π_A
G	0,5	0,5	0,5	0,5	π_G
C	0,5	0	0,5	0,5	π_C
T	0,5	0,5	0,5	0	π_T



1 pav. Bakterijos „genomo parašas“. Taškų skaičius $\frac{1}{4^n}$ ploto kvadrate atitinka n ilgio oligonukleotidų dažnius visoje nukleotidų sekoje.

ilgio „žodžio“ (oligonukleotido) dažnį visoje DNR sekoje. Kuo ilgesnis „žodis“, tuo mažesnio ploto kvadratas „genomo paraše“ jį atitinka. Galų gale visą nukleotidų seką atitinka vienas, paskutinis taškas. Taigi, turint kelis tokius skirtingų organizmų DNR generuotus 4 viršūnių „chaoso žaidimo“ fraktalus galima pasakyti, kad jų vizualiniai skirtumai yra ne kas kita, kaip trumpų nukleotidų sekų dažnių skirtumai.

1.2. Atstumas tarp genetinių sekų

Panaudojant IFS gaunamas kompaktiškas visos sekos atvaizdavimas. Kiekvienas „genomo parašo“ taškas atitinka konkretų (priklausomai nuo pasirinkto tikslumo) oligonukleotidą, tačiau atstumas tarp dviejų tokių taškų yra sunkiai interpretuojamas. Vienetiniame „genomo parašo“ kvadrate tuo pačiu atstumu nutolusius taškus atitinkantys oligonukleotidai pagal savo prasmę gali būti ir artimi, ir labai nutolę, ypač jei jie yra skirtinguose mažesnio ploto kvadratuose. Šiuo atveju gerai suvokiamas euklidinis atstumas nėra informatyvus matas atstumui tarp oligonukleoditų matuoti. Skaičiuojant atstumą tarp dviejų skirtingo ilgio simbolių sekų (pvz., skirtingų rūšių genomų, genų ar baltymų sekų) natūralus yra Levenšteino atstumas – minimalus operacijų skaičius, kurias reikia atlikti norint iš vienos sekos gauti kitą. Galimos operacijos: simbolio įterpimas, ištrynimas arba pakeitimas kitu simboliu. Imant DNR sekas šios operacijos atitinka nukleotidų mutacijas. Jei sekos yra vienodo ilgio, Levenšteino atstumas sutampa su Hamingo atstumu. Tačiau ir šie atstumai turi savo trūkumų. Įvertindami skirtumus tarp sekų, jie neatsižvelgia į tų sekų struktūrą ar biologinę prasmę.

Bioinformatikoje įvairios genetinės sekos tarpusavyje lyginamos ir atstumai tarp jų nustatinėjami labai dažnai (Smith, Waterman, 1981; Thompson, Higgins, Gibson, 1994). Taip nustatomas genų gimininguma, baltymų struktūros ar atliekamų funkcijų panašumas ir t.t. Dviejų organizmų gimininguma lyginant jų DNR taip pat susiveda į atstumo tarp dviejų genomų skaičiavimą. Turint tokius atstumus tarp rūšių, galima rekonstruoti filogenetinius medžius, tirti rūšių kilmę. Toliau suformuluosime atskiro sekų atvejo – binarinių sekų išrikiavimo ir atstumų tarp jų radimo uždavinį.

2. Binarinių sekų išrikiavimo uždavinys

Tarkime tarp n ilgio simbolių sekų x ir z erdvėje R^n įvestas atstumas $d(x, z)$. Panaudojant tiesinę ar netiesinę transformaciją šias sekas, maksimaliai išlaikant atstumus tarp jų, reikia atvaizduoti erdvėje R^k. Norint sekas išrikiuoti realioje tiesėje $k = 1$. Kiekvienas nukleotidų sekos elementas pasižymi dviem savybėmis ir yra iš aibės $\{A, C, G, T\} \sim \mathcal{A} \times \mathcal{A}$, $\mathcal{A} = \{0, 1\}$. Toliau tokią seką nagrinėsime tik pagal vieną jos savybę, todėl x yra n ilgio nulių ir vienetų (binarinė) seka:

$$x = x_1 x_2 x_3 \dots x_n = \{x_i \in \mathcal{A}, i = 1, \dots, n\}, \quad \mathcal{A} = \{0, 1\}. \tag{2}$$

Visų tokių 2^n sekų aibė $\mathcal{M}_n$ yra izomorfiška n-mačio kubo viršūnių aibei. Funkcijos $x \in \mathcal{M}_n$ sudėtingumui tirti naudojamas tiesinis operatorius $\mathcal{B}$

$$\mathcal{B}\colon \mathcal{M}_n \to \mathcal{M}_n, \quad y = \mathcal{B}x. \tag{3}$$

Operatorius $\mathcal{B}$ apibrėžtas formule $y_i = (x_{i+1} - x_i) \bmod (2)$. Tokiu būdu funkcija y yra funkcijos x pirmos eilės skirtumai. Kad skirtumų būtų n, funkcijos x reikšmė x_{n+1} prilyginama x_1, t.y. laikoma, jog funkcija x su reikšmėmis x_i taškuose i yra periodinė su periodu n. Baigtinės aibės $\mathcal{M}_n$ atvaizdavimas į save pačią užduodamas grafu su 2^n viršūnių. Iš kiekvienos viršūnės x išeina viena briauna ir sujungia ją su viršūne $\mathcal{B}x$. Gautas grafas gali turėti kelias jungiąsias komponentes, kurios visos turi po vieną ciklą, o ciklo viršūnės yra binarinio medžio kamienas. Tokiu būdu galima įvesti tam tikrą sekų x tvarką, kurios kriterijus – funkcijos sudėtingumas. Jei funkcija yra konstanta, pirmi skirtumai bus nuliai. Jei pirmi skirtumai konstanta, tai funkcija bus pirmo laipsnio daugianaris. Jei antri skirtumai yra konstanta, tai funkcija yra ne daugiau kaip antro laipsnio daugianaris ir t.t. Pagal šią schemą funkcija x laikoma tuo sudėt ingesnė, kuo ilgesnio ciklo komponentėje ji yra, o komponentės ribose x tuo sudėtingesnė, kuo ji toliau nuo ciklo viršūnės (plačiau apie binarinių sekų sudėtingumą V.I. Arnold publikacijoje [6]). Tačiau tokiu būdu gautas funkcijų x išdėstymas yra hierarchinis (o prie $n = 2^k$ tiksliai atitinka binarinį medį); ciklo viršūnės yra tokio paties sudėtingumo funkcijos, taip pat lygiavertės yra vienodai nuo ciklo viršūnių ar binarinio medžio kamieno nutolusios funkcijos, todėl sekų x negalima vienareikšmiškai išdėstyti ant tiesės, jų išrikiuoti.

Binarinių sekų išrikiavimo uždaviniui spręsti pritaikysime V.I. Arnold pasiūlytą idėją išnaudoti informaciją apie funkcijos skirtumus. Seką x atvaizduosime didesnio matavimo erdvėje išnaudodami informaciją apie jos pirmos, antros ir t.t. iki $(n-1)$-tos eilės skirtumus. Čia laikome, kad seka x nėra periodinė, todėl aukštesnės eilės skirtumų bus vis mažiau: pirmos eilės skirtumų bus $(n-1)$, antros $(n-2)$ ir t.t. iki vieno.

Apibrėšime operatorių $\mathcal{B}$. Simbolių seką x sutapatinsime su vektoriumi $x = (x_1, \dots, x_n)$ erdvėje R^n.

$$\mathcal{M}_n \overset{\mathcal{B}_1^{(n)}}{\to} \mathcal{M}_{n-1} \overset{\mathcal{B}_1^{(n-1)}}{\to} \mathcal{M}_{n-2} \overset{\mathcal{B}_1^{(n-2)}}{\to} \dots \overset{\mathcal{B}_1^{(3)}}{\to} \mathcal{M}_2 \overset{\mathcal{B}_1^{(2)}}{\to} \mathcal{M}_1, \quad \mathcal{M}_i \subset R^i, \ \ i = \overline{1, n}. \tag{4}$$

Čia operatorius $\mathcal{B}_1^{(k)}$ $(k \in \{2, \dots, n\})$ išreiškiamas formule

$$\mathcal{B}_1^{(k)} x = \left\{(x_{i+1} - x_i)/2, \ i = \overline{1, k-1}\right\}, \tag{5}$$

o operatoriai $\mathcal{B}_l^{(n)}$: $\mathcal{M}_n \to \mathcal{M}_{n-l}$ gaunami rekuretiškai

$$\mathcal{B}_l^{(n)} = \mathcal{B}_1^{(n-l+1)} \mathcal{B}_{l-1}^{(n)}, \quad l = \overline{2, n-1}. \tag{6}$$

Operatoriaus $\mathcal{B}_l^{(n)}$ viršutinis indeksas n žymi, kokio matavimo erdvėje jis veikia, o apatinis indeksas l parodo, kiek mažesnis yra jo vaizdo matavimas. Duotam $x \in \mathcal{M}_n \subset R^n$, jį atitinkančio taško $y = \mathcal{B}x$ koordinatės erdvėje $R^{n(n+1)/2}$ išreiškiamos formule

$$y = y(x) = \mathcal{B}x = \big(x,\, \mathcal{B}_1^{(n)} x,\, \mathcal{B}_2^{(n)} x, \ldots, \mathcal{B}_{n-1}^{(n)} x\big). \tag{7}$$

Tegul atstumas tarp dviejų binarinių sekų x ir z yra vektorių $y(x)$ ir $y(z)$ skirtumo euklidinė norma

$$d(x, z) = \big\| y(x) - y(z) \big\|, \quad x, z \in \mathcal{M}_n, \quad y(x), y(z) \in \mathcal{M} \subset R^{n(n+1)/2}. \tag{8}$$

Kiek galima mažiau iškraipant sekų x is $\mathcal{M}_n$ tarpusavio išsidėstymą naujoje erdvėje, t.y. atsižvelgiant į jų tarpusavio atstumus $d(x, z)$, $x, z \in \mathcal{M}_n$, jas reikia išrikiuoti ant tiesės. Tai yra klasikinis daugiamačio mastelio parinkimo (*multidimensional scaling*) uždavinys: turint atstumus $d(x, z)$ erdvėje $R^{n(n+1)/2}$, reikia rasti tokius $w(x)$ ir $w(z)$ erdvėje R^k, $k \ll n(n+1)/2$, kad $\|w(x) - w(z)\| \approx d(x, z)$ su kuo mažesne paklaida. Sekų x koordinatės $w(x)$ gali būti randamos kaip ortogonali $y(x)$ projekcija į k matavimų tiesinį poerdvį. Kai $k = 1$, $w(x)$ atitiks binarinių sekų x koordinates ant realios tiesės. Šis uždavinys taip pat ekvivalentus objektų x atvaizdavimui naudojant pagrindinių komponenčių metodą [7].

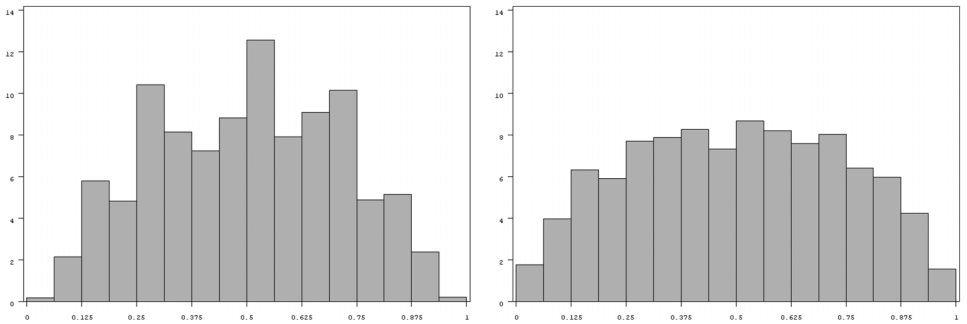
3. Rezultatai ir jų pritaikymas

Koordinatėms $w(x)$ apskaičiuoti, turint sekų x tarpusavio atstumų matricą $\mathbf{D} = \{d(x, z)\}$, $\forall x, z \in \mathcal{M}_n$, naudojame statistinio paketo *SAS* procedūrą MDS (Young, Lewyckyj, Takane, 1986). Kadangi pradinių koordinačių $y(x)$ pasukimas ir projektavimas nekeičia sekų x tarpusavio atstumų, tokio uždavinio sprendinys $w(x)$ gali būti nevienintelis. Gautos vienmatės binarinių kodų x koordinatės $w(x)$ išlaikant jų tvarką tiesine transformacija normuojamos taip, kad tilptų į vienetinį intervalą, $w(x) \in [0, 1]$.

Galima pastebėti, kad tokiu būdu gauta kodų tvarka pasižymi tam tikra išdėstymo simetrija. Be to, išdėstymo galuose visada yra mažiausio periodo funkcijos, o viduryje yra konstantos, t.y. sekos vien tik iš nulių arba vien tik iš vienetų.

2 lentelė. $n = 5$ ilgio binarinių kodų x išrikiavimo lentelė ir jų koordinatės $w(x)$

nr.	kodas	taškas	nr.	kodas	taškas	nr.	kodas	taškas	nr.	kodas	taškas
1	10101	0,0000	9	10110	0,2588	17	00000	0,5083	25	10010	0,7755
2	00101	0,0403	10	11100	0,2969	18	01111	0,5523	26	11000	0,7799
3	10100	0,0623	11	10001	0,3307	19	11110	0,5743	27	00010	0,8224
4	00100	0,1033	12	00110	0,3550	20	11001	0,6193	28	01000	0,8481
5	11101	0,1519	13	01100	0,3807	21	10011	0,6450	29	11011	0,8967
6	10111	0,1776	14	00001	0,4257	22	01110	0,6693	30	01011	0,9377
7	01101	0,1988	15	10000	0,4477	23	01001	0,7155	31	11010	0,9597
8	00111	0,2458	16	11111	0,4917	24	00011	0,7288	32	01010	1,0000



2 pav. $n = 10$ ilgio binarinių kodų pasiskirstymas bakterijų *Bordetella bronchiseptica* (kairėje) ir *Escherichia coli* nekoduojančios genomo dalies $m = 10^4$ nukleotidų ilgio sekoje.

Kadangi kiekvienas iš nukleotidų pasižymi dviem savybėmis (vandenilinių jungčių skaičius ir molekulės tipas), nukleotidų seką $s = s_1 s_2 s_3 \dots s_m = \{s_i \in \mathcal{A}, i = \overline{1, m}\}$, kur $\mathcal{A} = \{A, C, G, T\}$ galima užrašyti atitinkama dvimate binarine seka. Slenkančiai n ilgio vienmatei binarinei sekai $x = x(j)$, kur $x(j) = s_j s_{j+1} \dots s_{j+n},\ j = \overline{1, m-n+1}$ priskyrus atitinkamo kodo koordinatę $w(x)$, galima gauti modifikuoto „genomo parašo“ koordinates. Imant tik vieną „genomo parašo“ koordinatę, galima nagrinėti sekų x pagal vieną iš nukleotidų savybių pasiskirstymą sekoje.

2 pav. pateikiamas dviejų skirtingų bakterijų (*Bordetella bronchiseptica* ir *Escherichia coli*) nekoduojančios genomo dalies $n = 10$ ilgio binarinių sekų x pasiskirstymas. Lyginant du organizmus, kuo didesnis n, tuo aukštesnio lygio genomo struktūros skirtumus galima įvertinti. Tai susiję su tuo, jog skiriasi ne tik dviejų genomų nukleotidinė sudėtis, bet ir gretimų nukleotidų porų bei ilgesnių oligonukleotidų dažniai. Nors paimtos dvi sekos s ir nekoduojančios, tačiau galima pastebėti, jog skirtingų organizmų kodų histogramos skiriasi. Nukleotidų seka čia perkoduojama į binarinę pagal taisyklę $\{C, G\} \to \{1\}$ ir $\{A, T\} \to \{0\}$.

Panašiu metodu gavus ilgų, $n > 100$ ilgio sekas atitinkančių kodų išdėstymus, galima būtų tirti nukleotidų sekų evoliucijas, vizualizuoti jų trajektorijas, skaičiuoti evoliucijų statistikas.

Literatūra

1. E. Huai-hsin Chi, J. Riedl, E. Shoop, J.V. Carlis, E. Retzel, P. Barry, Flexible information visualization of multivariate data from biological sequence similarity searches, in: *IEEE Visualization '96* (1996), pp. 133–140.
2. 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).
3. 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).
4. D. Ashlock, J. Golden, Ch.11, *Evolutionary Computation and Fractal Visualization of Sequence Data*, Morgan Kaufmann (2002).
5. H.J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Res.*, **18**, 2163–2170 (1990).
6. V.I. Arnold, Slozhnost konechnyh posledovatelnostei nulei i edinic i geometrija konechnyh funkcionalnyh prostranstv (2005)

http://mms.math-net.ru/meetings/2005
/arnold.pdf.
7. N.H. Timm, *Applied Multivariate Analysis*, Springer, New York (2002).

SUMMARY

T. Rekašius. Visualisation of nucleotide sequences

The paper reviews the visualisation problems of genetic sequences and of their analysis results as well as other related problems. A possible formulation of the problem based on similarity, complexity measure (distance) between DNA sequences is proposed. It is solved by making use of multidimensional scaling (MDS) of principal component analysis (PCA) methods.

Keywords: DNA sequences, complexity measure, multidimensional scaling.