Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Genetinių sekų vizualizavimo modeliai

Baigiamasis magistro studijų projektas

Kamilija Jablonskaitė

Projekto autorė

doc. dr. Tomas Ruzgas

Vadovas

Kaunas, 2023

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Genetinių sekų vizualizavimo modeliai

Baigiamasis magistro studijų projektas

Taikomoji matematika (6211AX006)

Kamilija Jablonskaitė

Projekto autorė

doc. dr. Tomas Ruzgas

Vadovas

doc. dr. Daiva Petkevičiūtė - Gerlach

Recenzentė

Kaunas, 2023

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Kamilija Jablonskaitė

Genetinių sekų vizualizavimo modeliai

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;

2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąraše;

3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs (-usi);

4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai nagrinėjant galimą akademinės etikos pažeidimą.

Kamilija Jablonskaitė

Patvirtinta elektroniniu būdu

Jablonskaitė, Kamilija. Genetinių sekų vizualizavimo modeliai. Magistro studijų baigiamasis projektas / vadovas doc. dr. Tomas Ruzgas; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika (Matematikos mokslai).

Reikšminiai žodžiai: Genetinių sekų vizualizavimas, matricų vaizdavimo metodas, chaoso žaidimo vaizdavimo metodas, chaoso žaidimo dažnių vaizdavimo metodas, fraktalinės struktūros.

Kaunas, 2023. 49 p.

Santrauka

Visų gyvybės formų genetinė informacija yra saugoma ilgose nukleotidų sekose, kurios sudaro individo dezoksiribonukleino rūgštį. Nuo organizmo sudėtingumo priklauso ir nukleotidų sekos ilgis. Vis didėjantis naujai nuskaitomų įvairių organizmų genomų skaičius atveria galimybes ir naujų tyrimų vystymui, tačiau dėl nevienodų sekų ilgių, didelio informacijos kiekio ir jos pateikimo būdo sudėtinga analizuoti genomo seką naudojant skaitinius algoritmus. Paprastesnis ir greitesnis sekų palyginimo ir analizės metodas galėtų būti genomo vaizdavimas plokštumoje. Sekos atvaizdavimo plokštumoje būdu visą sekos struktūrą, pavienių nukleotidų ir jų kombinacijų dažnį ir dėsningumus būtų galima pavaizduoti pakankamai glaustai – viename paveiksle.

Šiame darbe nagrinėjami chaoso žaidimo, chaoso žaidimo dažnių ir matricos vaizdavimo metodai ir jų parametrų įtaka gaunamam vaizdui. Darbe pateikiamos žmogaus chromosomų, atsitiktinai sugeneruotų ir daugianare logistine regresija gautų DNR sekų vizualizacijos. Vaizduose matomiems raštams pavaizduoti naudojamas vaizdo kontrasto didinimas. Vaizdų palyginimui pasirinkti struktūrinio panašumo indekso, Pirsono koreliacijos koeficiento parametrai ir vaizdų atimties metodas. Logistine regresija taip pat vertinta, ar priklausomybė tarp duomenų vienetų egzistuoja, ir nustatytas Markovo grandinių modelio tinkamumas sekoms nusakyti.

Atlikus tyrimus pastebėta, jog gaunamo vaizdo struktūra kai kurių sekų vaizdavimo atvejais turi fraktalo savybių; tai parodo ir atitinkamo genomo struktūros požymį turėti ar neturėti atitinkamų nukleotidų ir jų kombinacijų. Nustatyta, jog fraktalines struktūras vaizde galima pastebėti, kai vizualizavimo metodų pasirenkamo parametro (sekos fragmento ilgio) reikšmė yra didesnė už 3. Pastebėta, jog chaoso žaidimo metodas nėra tinkamas vizualizuoti ilgas sekas, jeigu siekiama pavaizduoti susidarančias fraktalines struktūras – tokiu atveju papildomai reikia skaičiuoti taškų dažnį. Palyginus atsitiktinai sugeneruotas, logistine regresija gautas ir realias DNR sekas nustatyta, jog DNR sekų struktūra yra unikali – nors dauguma atvejų įmanoma prognozuoti nukleotido bazę žinant jam iš šonų esančius kaimyninius nukleotidus, logistine regresija gautos sekos neatitinka realių DNR sekų vizualizacijų.

Jablonskaitė, Kamilija. Visualization methods of genetic sequences. Master's Final Degree Project / supervisor doc. dr. Tomas Ruzgas; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Applied Mathematics (Mathematical Sciences).

Keywords: Visualization of genetic sequences, matrix representation method, chaos game representation method, chaos game frequency representation method, fractal structures.

Kaunas, 2023. 49 p.

Summary

The genetic information of all life forms is stored in long sequences of nucleotides that make up an individuals deoxyribonucleic acid. The length of the nucleotide sequence also depends on the complexity of the organism. The ever–increasing number of newly read genomes of various organisms opens up opportunities for the development of new research, but due to the unequal lengths of the sequences, the large amount of information and the way it is presented, it is difficult to analyze the genome sequence using numerical algorithms. A simpler and faster method for sequence comparison and analysis could be to represent the genome in a two–dimensional surface. By mapping the sequence on a plane, it could be possible to visually see the whole structure of the sequence, the frequency and regularities of single nucleotides and their combinations in one picture.

This paper examines visualization methods of chaos game, chaos game frequencies and matrix representation and the influence of their parameters on the resulting image. The work presents visualizations of DNA sequences of human chromosomes, randomly generated and obtained by multinomial logistic regression. Image contrast adjustment is used to show patterns seen in images. Structural similarity index, Pearson correlation coefficient parameters and image subtraction method were selected for image comparison. Logistic regression was used to assess whether the dependence between data units exists and determine the suitability of the Markov chain model for predicting sequences.

The results have shown that the structure of the resulting image has fractal properties in some cases of sequence representation. It has been established that fractal structures can be observed in the image when the value of the selected parameter (sequence fragment length) of the visualization methods is greater than 3. It was observed that the chaos game method is not suitable for visualizing long sequences, if the goal is to depict the resulting fractal structures - in this case, point frequency must be calculated additionally. Comparing randomly generated, obtained by logistic regression and real DNA sequences revealed that the structure of DNA is unique - although in most cases it is possible to predict the base of a nucleotide by knowing its neighboring nucleotides, the sequences obtained by logistic regression do not match visualizations of real DNA sequences.

Turinys

Paveikslų sąrašas

Santrumpų sąrašas

Santrumpos:

DNR – dezoksiribonukleino rūgštis;

RNR – ribonukleino rūgštis;

SSIM – (angl. *structural similarity index measure*) struktūrinio panašumo indeksas;

NCBI – (angl. *National Center for Biotechnology Information*) Nacionalinis biotechnologijų informacijos centras;

RGB – (angl. *red, green, blue*) spalvų modelis, kuriuo naudojantis įvairios spalvos išreiškiamos pagrindinių trijų spalvų (raudonos, žalios, mėlynos) procentais.

GINS2 – proteiną koduojantis genas, esantis 16 – oje žmogaus chromosomoje. Koduoja replikacijoje naudojamą proteiną.

CTRL – proteiną koduojantis genas, esantis 16 – oje žmogaus chromosomoje. Koduoja virškinimo sistemoje naudojamą fermentą.

BCAR1 – (angl. *breast cancer anti-estrogen resistance protein 1*) proteiną koduojantis genas, esantis 16 – oje žmogaus chromosomoje.

MTHFR – (angl. *methylenetetrahydrofolate reductase*) proteiną koduojantis genas, esantis 1 – oje žmogaus chromosomoje.

MTOR – (angl. *mechanistic target of rapamycin kinase*) proteiną koduojantis genas, esantis 1 – oje žmogaus chromosomoje.

AGT – (angl. *angiotensinogen*) proteiną koduojantis genas, esantis 1 – oje žmogaus chromosomoje. Proteinas reguliuoja kraujospūdį ir skysčių bei druskų pusiausvyrą organizme.

VDR – (angl. *vitamin D receptor*) proteiną koduojantis genas, esantis 12 – oje žmogaus chromosomoje. Proteinas leidžia organizmui reaguoti į vitaminą D.

Įvadas

Visų gyvybės formų genetinė informacija yra saugoma ilgose nukleotidų sekose, kurios savo ruožtu sudaro individo dezoksiribonukleino rūgštį (toliau – DNR). Nuo organizmo sudėtingumo priklauso ir nukleotidų sekos (toliau – genomo) ilgis. Paprastuose mikrooorganizmuose, pavyzdžiui, virusuose, sekos ilgis gali siekti apie 300 tūkst. skaičių nukleotidų, o žmogaus genomo ilgis siekia net 3,5 milijardo nukleotidų. Vis didėjantis naujai nuskaitomų įvairių organizmų genomų skaičius atveria galimybes ir naujų tyrimų vystymui, tačiau dėl nevienodų sekų ilgių, didelio informacijos kiekio ir jos pateikimo būdo sudėtinga analizuoti genomo seką naudojant įmantrius skaitinius algoritmus. Paprastesnis ir greitesnis sekų palyginimo ir analizės metodas galėtų būti genomo vaizdavimas plokštumoje – tokiu būdu visą sekos struktūrą ir jos dėsningumus, pavienių nukleotidų ir jų kombinacijų dažnį būtų galima pavaizduoti pakankamai glaustai – viename paveiksle.

DNR dalyvauja įvairiuose procesuose, kurie lemia kiekvieno net ir tai pačiai rūšiai priklausančio individo skirtumus. Vieno svarbiausių procesų – replikacijos – metu DNR dalijasi į dvi lygias dalis, tokiu būdu leidžiant organizmui augti ir atnaujinti ląsteles. Tačiau replikacijos metu gali pasitaikyti klaidų (mutacijų), kurios gali daryti įtaką ne tik neteisingam genų nuskaitymui sekoje, tačiau ir įvairių ligų atsiradimui ir jų paveldimumui (pavyzdys – įvairios genetinės ligos, pvz. Hantingtono liga [1]). Sekos atvaizdavimo plokštumoje būdu atsirastų galimybė vizualiai pamatyti ligų požymius sekoje ar netgi surasti ligų, kurių priežastys dar nėra nustatytos, požymius.

Svarbu paminėti, jog žmogaus genomo seka pirmąkart nuskaityta tik 2000 metais – tačiau tuomet buvo nuskaityti tik ~92% visos sekos. Visa žmogaus genomo seka, neturinti jokių spragų, nuskaityta tik 2022 metų pradžioje. Turima informacija nuolat atnaujinama, ir jos vis daugėja – todėl sekoms nagrinėti vis dar kuriami nauji metodai, tačiau diduma jų gali efektyviai nagrinėti tik mažą dalį genomo. Rečiau gilinamasi į ilgesnės sekos, ne genų ar viso genomo struktūrą. Atsižvelgiant į nukleotidų biologines savybes ir trumpas, ir ilgas sekas galima pavaizduoti įvairiais metodais gaunant skirtingus vaizdus, o gaunamas rezultatas priklauso nuo vaizdavimui pasirinktų nukleotidų savybių. Gaunamo vaizdo raštas kai kurių sekų vaizdavimo atvejais taip pat turi fraktalinę struktūrą; tai parodo ir atitinkamo genomo požymį turėti ar neturėti atitinkamų nukleotidų ir jų kombinacijų, ir parodo unikalią ne tik trumpų, bet ir ilgų sekų struktūrą.

Taigi, šio **tyrimo tikslas** – genetinių algoritmų taikymo idėjos pagrindu įgyvendinti genetinės sekos išdėstymo (vizualizavimo) plokštumoje metodą.

Tikslui pasiekti buvo suformuluoti tokie **uždaviniai**:

1. apžvelgus mokslinę literatūrą parinkti skirtingus genetinių sekų vaizdavimo metodus bei vaizdų panašumui palyginti skirtus metodus;
2. pavaizduoti genetines sekas keliais skirtingais vaizdavimo metodais dvimatėje plokštumoje;
3. įvertinti sekų struktūros panašumus kelių skirtingų sekų ilgių atžvilgiu;
4. palyginti žmogaus chromosomos, atsitiktinai sugeneruotos ir logistine regresija gautų sekų vizualizacijas pasirinktais vaizdų palyginimo parametrais;
5. įvertinti, ar įmanoma prognozuoti nukleotido bazę žinant jam iš šonų esančius kaimyninius nukleotidus.

1. Literatūros apžvalga

Šiame skyriuje yra apžvelgiama DNR sandara ir egzistuojančios taisyklės, kuriomis remiantis sekoje koduojama informacija. Taip pat apžvelgiami įvairūs matematiniai DNR sekų tyrimai, kuriais pagrindžiama pasirinktos temos svarba bei pateikiami įvairūs sekų vaizdavimo metodai ir gautų vaizdų palyginimo parametrai.

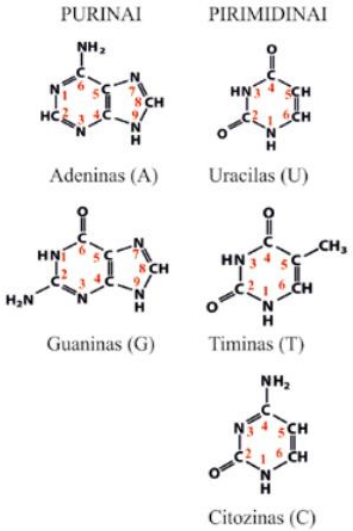
Remiantis atlikta analize, pagrindžiamas genomo vaizdavimo metodų pasirinkimas, pateiktas matematinių metodų, modelių ir programinės įrangos pasirinkimo pagrindimas bei patikslinti darbe sprendžiami uždaviniai.

1.1. DNR struktūra ir pagrindiniai procesai

DNR ir RNR molekulės yra nereguliarūs polimerai, sudaryti iš kovalentiškai sujungtų tarpusavyje struktūrinių monomerų – nukleotidų. Nukleotidai skiriasi azoto bazėmis: DNR turi 4 skirtingas pagrindines bazes – adenino (A), citozino (C), guanino (G) ir timino (T). RNR (ribonukleino rūgštyje) sekoje vietoj timino (T) heterociklinės bazės yra uracilo (U) bazė. Taip paaiškinama pirminė DNR ir RNR struktūra.

Nukleotidai gali būti skirstomi pagal:

1. heterociklines bazes į purinus (A ir G; turi purino žiedą) ir pirimidinus (C ir T (U, jeigu RNR); turi pirimidino žiedą) (žr. 1 pav.);;
2. vandenilinius ryšius (du tarp A ir T (U, jeigu RNR) ir trys tarp C ir G) (žr. 1 pav.);
3. bazių grupes į amino (A ir C) ir keto (G ir T (U, jeigu RNR)).



1 pav. Nukleorūgščių struktūriniai komponentai. Heterociklinių bazių, įeinančių į DNR ir RNR nukleotidų sudėtį, struktūra [1]

Egzistuoja įvairūs nukleotidų sekų užrašymo būdai. Pats paprasčiausias ir dažniausiai naudojamas DNR (ir RNR) nukleotidų užrašymo būdas yra sekos užrašymas vienaraidžiais simboliais, pavyzdžiui:

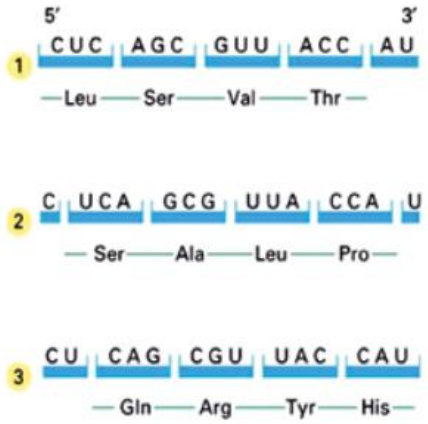
GCATATTGC

Šiuo būdu (vadinamu FASTA formatu) yra užrašomos organizmų genomų nukleotidų sekos, kurios yra talpinamos duomenų bazėse [1].

DNR grandinė sudaryta iš dviejų deoksiribonukleotidų (toliau – nukleotidų) grandinių, sujungtų jų poromis. Kiekviena grandinė turi du galus: vieną laisvą 5′ (fosforo rūgšties liekanos) galą, ir kitą laisvą 3′ (hidroksigrupės) galą. Dvi grandinės, sudarančios DNR molekulę, yra priešingų krypčių (antilygiagrečių - viena grandinė orientuota 5′→3′ kryptimi, o kita – priešinga, t. y. 3′→5′, kryptimi). Erdvinis DNR struktūros modelis – dviguba grandinių spiralė, susukta apie tariamą ašį – yra laikomas antrine DNR struktūra.

DNR molekulėse saugoma genetinė informacija yra užkoduota būsimų baltymų aminorūgščių sekos pavidalu. Vienoje grandinėje esantys nukleotidų tripletai (kitaip - kodonai) koduoja aminorūgštis, kurios yra sintetinamos į baltymus. Nukleotidų tripletai gali sudaryti 64 skirtingas kombinacijas (4×4×4), tačiau skirtingų aminorūgščių egzistuoja tik 20 – yra žinoma, jog kelios skirtingos kombinacijos koduoja tas pačias aminorūgštis. Tokios kombinacijos yra vadinamos sinonimais.

Kadangi aminorūgštis koduojantys tripletai nepersidengia, vienos DNR grandinės seka gali turėti tris skaitymo rėmelius (žr. 2 pav.).



2 pav. Trys galimi skaitymo rėmeliai RNR [1]

Baltymų biosintezės metu atrenkamos tik reikalingos DNR sekos dalys, kuriose yra užkoduotos aminorūgštys – šios dalys vadinamos atvirais skaitymo rėmeliais (angl. *open reading frame*, ORF). Atviri skaitymo rėmeliai yra apriboti specialiais kodonais, kurie nusako rėmelio pradžią ir pabaigą. Pradžios kodonas visuomet yra ATG (RNR atitikmuo – AUG), kuris aprašo aminorūgštį metioniną. Pabaigos kodonai gali būti trys – TAA/TAG/TGA (RNR atitikmuo – UAA/UAG/UGA). Šie kodonai beveik niekuomet nekoduoja jokių aminorūgščių, jų pagrindinė paskirtis – sustabdyti atviro skaitymo rėmelio nuskaitymą [1].

Tiesa, skirtinguose organizmuose dėl įvykusių mutacijų ar kitų faktorių gali egzistuoti keli papildomi skirtingi pradžios ir pabaigos kodonai – paminėti yra universalūs visiems organizmams [2]. Atviri skaitymo rėmeliai dažnai (įprastai – eukariotuose) sudaro tik mažą dalį visos DNR sekos (pvz. žmogaus genome – tik apie 1,5 – 2%), tačiau prokariotuose didesnė dalis (pvz.. bakterijos e.coli atveju – apie 80 – 90%) DNR sudaryta iš galimų genų [1]. Atviri skaitymo rėmeliai gali būti įvairaus ilgio ir tik dalis jų sudaro genus – paliekami tik tie rėmelių fragmentai, kurie yra reikalingi baltymų sintezei [1].

DNR dalyvauja replikacijos, transkripcijos ir reparacijos procesuose. Vykstant replikacijai dvi DNR molekulės šakos atsiskiria į dvi dalis ir prisijungia naujus komplementarius nukleotidus sudarydami jų poras (A jungiasi su T, C jungiasi su G). Tokiu būdu po replikacijos susidaro dvi DNR molekulės, kurios yra identiškos pradinei molekulei. Taip genetinė informacija yra išsaugoma ir perduodama palikuonims [1]. Reparacijos metu yra ištaisomos replikacijos metu atliktos klaidos – dėl to replikacijos metu padarytų klaidų skaičius sumažėja apie 1000 kartų. Transkripcijos proceso metu iš DNR molekulės nukleotidų sekos dalies sintetinama jai komplementari RNR molekulė. Transkripcijos metu gaunama RNR molekulės nukleotidų seka yra pirmas geno formavimo etapas [1].

Replikacijos proceso metu gali pasitaikyti klaidų – pavyzdžiui, gali susidaryti nekomplementari nukleotidų pora, nukleotidai gali neprisijungti prie poros arba jų prisijungti gali per daug. Nors replikacijos metu (ir po jos – reparacijos metu) klaidos yra taisomos, neatitikimų tarp pradinės genomo sekos ir replikuotos tikimybė gali išlikti nuo 10^{-10} iki 10^{-5}. Tai reiškia, jog vienos replikacijos metu gali pasitaikyti nuo vienos (arba nei vienos) klaidos iki 60000 klaidų. Neištaisyti netikslumai yra vadinami mutacijomis – jos lemia ir tos pačios rūšies organizmų individualumą, ir rūšių tarpusavio skirtumus. Sisteminga genetinio kodo mutacija, vykstančia dėl gamtinės atrankos yra grindžiama ir evoliucijos teorija [3].

1.2. DNR sekų tyrimai

Nuolat didėjantis įvairių organizmų nuskaitytų genomų skaičius leidžia patikrinti teorinių DNR sekų struktūros prielaidų teisingumą praktiškai – realių duomenų pagalba. Vieni paprasčiausių naudojamų metodų, taikytų prognozuoti nukleotidų išsidėstymą sekoje, iki šiol išlieka Markovo grandinių modeliai. Markovo metodų tikslumą galima patikrinti naudojant logistinę regresiją [4] [5]. Daugianarė logistinė regresija nusakoma formule (1).

$$p = \frac{\exp\left(a+b_1X_1+b_2X_2+b_3X_3+\cdots\right)}{1+\exp\left(b_1X_1+b_2X_2+b_3X_3+\cdots\right)}. \quad \textbf{(1)}$$

čia p – tikimybė, jog atvejis priklauso tam tikrai kategorijai, a – lygties konstanta, b – prognozuojamojo arba nepriklausomo kintamojo koeficientas.

Dėl genomų sekų nuskaitymo klaidų duomenyse gali egzistuoti tarpai – nežinomos trumpesnės sekos dalys, dažnai žymimos N raidėmis sekoje. Trūkstamoms dalims prognozuoti sukurta daug skirtingų modelių, tačiau beveik visi jie paremti Markovo grandinių modeliais [6]. Markovo grandinėmis grįsti modeliai taikomi ne tik siekiant prognozuoti sekos struktūrą, bet ir mėginant aptikti genomo sekos mutacijas – pavyzdžiui, vėžinius DNR sekos fragmentus genuose [7].

Markovo grandinių modelis yra paprastas, tačiau jo praktinis taikymas DNR sekoms turi ir trūkumų. Žemesnės eilės Markovo grandinių modeliai dažnai nėra tinkami ilgoms nukleotidų sekoms nusakyti [4] [5] [8]. Aukštesnės eilės modelius galima taikyti sekoms, tačiau sudėtinga statistiškai nustatyti modelio tinkamumą [5] ir suvaldyti eksponentiškai eilės numeriui didėjantį parametrų skaičių [8]. Tiesa, visuose pateiktuose pavyzdžiuose Markovo grandinės taikomos trumpoms sekos dalims.

Nors Markovo grandinėmis paremti modeliai išlieka vienais populiariausių metodų genomo struktūrai prognozuoti, vis dažniau sekų modeliavimui yra pasirenkami ir kiti metodai, iš kurių dažniausiai taikomi yra neuroniniai tinklai. Įvairūs neuroninių tinklų modeliai taikomi ne tik siekiant

užpildyti nežinomus sekos fragmentus [9] [10], bet ir bandant klasifikuoti sekoje esančius genus [11] ar mažesnius sekos fragmentus, reguliuojančius genų išraišką [12].

Neuroninių tinklų modeliais taip pat bandoma ir identifikuoti genetines ligas, kurias galima laikyti unikaliais organizmo požymiais, paveldimais iš tiesioginių palikuonių. Pavyzdžiui, didelė dalis neurodegeneracinių ligų (pvz. Alzheimeris, išsėtinė sklerozė ir t.t.) nėra išgydomos, tačiau yra žinoma, jog ~10% šių ligų atvejų atsiranda ligą paveldėjus iš biologinių tėvų [13]. Viena iš išimčių galima laikyti Hantingtono ligą, kurios paveldėjimo iš biologinių tėvų tikimybė siekia net 90% [14]. Hantingtono liga yra siejama su nukleotidų pasikartojimu – liga susergama, kai sekos fragmento CAG pasikartojimų skaičius IT15 gene viršija 36 [1]. Tiesa, dėl genetinio kodo pokyčių susergama ne tik neurodegeneracinėmis ligomis. Vienu tokių pavyzdžių galėtų būti vėžys. Ši liga nėra laikoma paveldima, tačiau dėl genomo mutacijų, kuriuos įmanoma perduoti tiesioginiams palikuonims, jiems rizika susirgti yra didesnė. Dėl šių ir kitų ligų yra atliekama DNR analizė, kurios metu yra ieškoma galimų ligų požymių, kartu nagrinėjant ir žmonijos populiacijos istoriją. Analizei naudojami įrankiai dažnai kuriami naudojant neuroninius tinklus [15] [16].

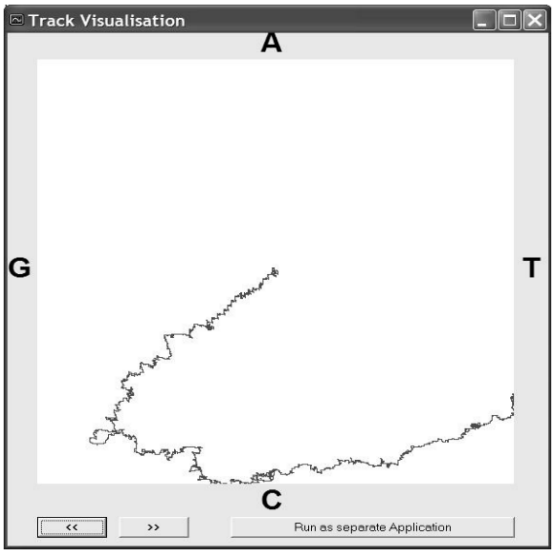
DNR nagrinėjama įvairiais tikslais, tačiau vienas svarbiausių uždavinių sekos ir jos procesų analizėje vis dar išlieka genomo mutacijų prognozavimas. Pavyzdžiui, dėl nuolat vykstančių genomo pokyčių virusai išvysto atsparumą vaistams. Neuroniniais tinklais mėginama prognozuoti tokias mutacijas virusuose ir surasti tikslius kintančius genomo sekos fragmentus, siekiant sukurti efektyvesnius priešvirusinius vaistus [17] [18]. DNR mutacijoms prognozuoti taip pat gali būti taikomas ir rečiau naudojamas ląstelinio automato (angl. *cellular automaton*) modelis [19]. Šis modelis gali būti naudojamas ir skirtumams tarp individų genų vizualizuoti [20].

Tiesa, su DNR susiję tyrimai (ypač tie, kurie tiria ligų vystymąsi ar priklausomybių daromą įtaką genetiniam kodui) dažnai gali trukti kelerius metus – dažniausiai aptinkami duomenys užfiksuoti tik tam tikru vienu laiko momentu, o vėlesniais laiko momentais užfiksuotų tų pačių organizmų DNR sekų aptikimas yra retas. Kadangi tyrimams reikalingi duomenys yra beveik neegzistuojantys, tirti DNR pokyčius skirtingais laiko momentais ar prognozuoti genomo mutacijas teisingai yra sudėtinga. Tačiau DNR nagrinėjimas vis dar yra aktualus – tyrimai padeda suprasti, kaip genetinis kodas nusako ligų ir viso organizmo vystymąsi.

1.3. DNR sekų vizualizavimo metodai

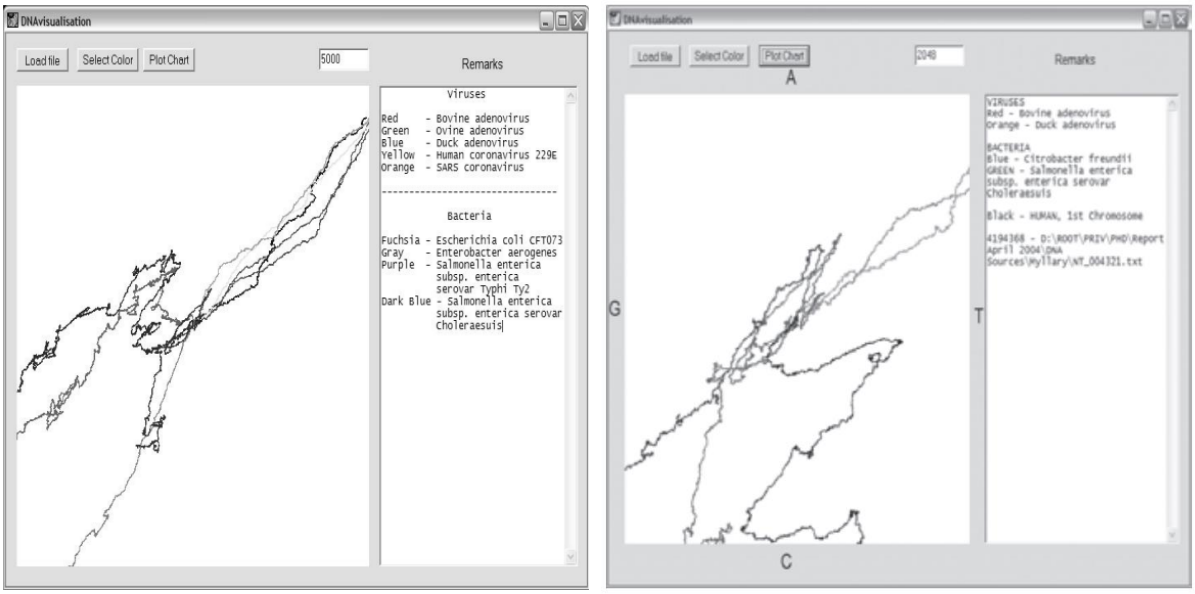
Dėl didelės duomenų apimties sudėtinga pastebėti DNR sekoje esančius sandaros ypatumus be papildomų įrankių. Genomo vizualizavimas plokštumoje yra vienas efektyviausių būdų ne tik pavaizduoti didelį duomenų kiekį, bet ir rasti sekos struktūros ypatumus, kurias pastebėti naudojant skaitinius metodus būtų sudėtinga. Per paskutinius dvidešimt metų nuo pirmo genomo nuskaitymo buvo sukurta daug įvairių įrankių, skirtingais metodais vizualizuojančių sekas, tačiau dažniausiai minimi yra kelio (angl. *track visualization*), arba kitaip – atsitiktinio ėjimo (angl. *random walk*), fraktalais grįsti ir chaoso žaidimo (angl. *chaos game*) vizualizavimo metodai [21] [22].

Atsitiktinio ėjimo vizualizavimo algoritmą tiksliai apibūdina jo pavadinimas. Algoritmui realizuoti reikalingas pradžios taškas – dažniausiai naudojamas koordinačių pradžios taškas (0;0). Keturias skirtingas nukleotidų rūšis atitinka keturios skirtingos linijos brėžimo kryptys. Kryptis galima pasirinkti įvairiai, pavyzdžiui, 3 paveiksle algoritmo žingsnis į viršų nusako adenino nukleotidą, į kairę – o į dešinę – timino, o žemyn – citozino. Nuo pasirinktų krypčių priklauso ir galutinis sekos vizualizacijos vaizdas.



3 pav. Kelio (atsitiktinio ėjimo) algoritmu vizualizuotos sekos pavyzdys [23]

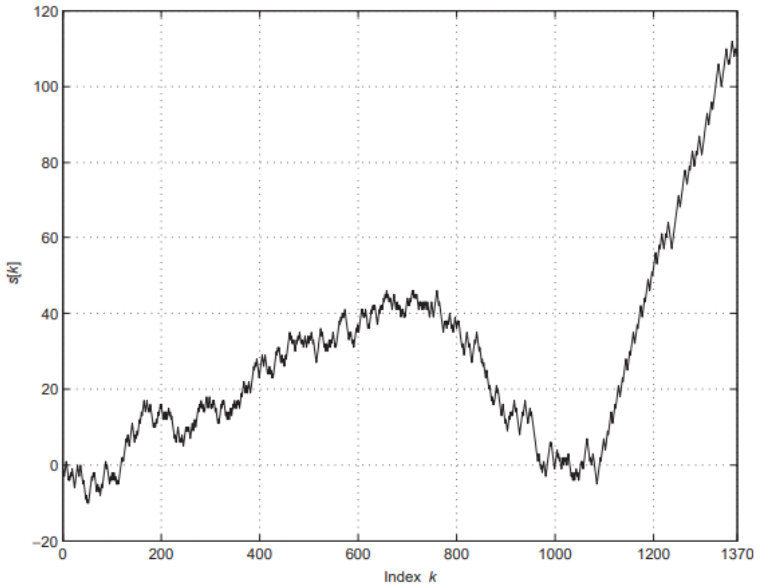
Turint pradino taško reikšmę ir pasirinktas krypčių reikšmes, vaizduojama nukleotido seka. Kiekvienas nukleotidas sekoje vaizduojamas iš eilės, nuo pradžios taško kryptingai brėžiant liniją, atitinkančią nukleotido rūšiai priskirtą kryptį. Šiuo vaizdavimo metodu lengva pastebėti nukleotidų tvarkos dėsningumus – jeigu visų nukleotidų tipų yra kiekis yra panašus, linija niekada daug nenutols nuo centro, o nukleotidų rūšies trūkumą sekoje parodys linijos kryptis. Pavyzdžiui, 3 paveiksle pateiktos sekos pavyzdyje matomas didelis adenino nukleotidų trūkumas. Sekos pradžioje vyrauja guanino ir citozino nukleotidai, o gale – timino ir adenino. Dar vienas pavyzdys pateiktas 4 paveiksle – čia pavaizduotos kelios skirtingos DNR sekos, kurias vaizduojančių linijų kryptingumas tarpusavyje yra labai panašus – sekose vyrauja arba adenino ir timino nukleotidų kombinacija, arba citozino ir guanino [23].



4 pav. Kelio (atsitiktinio ėjimo) algoritmu atliktų kelių sekų vizualizacijų pavyzdžiai [23, 24]

Aptartas atsitiktinio ėjimo algoritmas nukleotidus skirsto pagal jų bazes į adeniną, citoziną, guaniną arba timiną, tačiau šis skirstymas yra tik vienas iš galimų simbolių kombinacijos variantų (toliau –

abėcėlių). Atsižvelgiant į DNR sekų biologines savybes, galima išskirti ir daugiau abėcėlių. Vienas iš galimų variantų yra skirstyti nukleotidus į dvi grupes, taip sudarant abėcėles, kurios gali būti nusakomos dvejais simboliais, pavyzdžiui: purinai ir pirimidinai, stiprų ir silpną ryšį tarp nukleotidų turintys arba iš keto ir amino bazių grupių sudaryti nukleotidai. Tokiu atveju vieną grupę atitinkančius nukleotidus galima žymėti įvairiai: pavyzdžiui, purinus – adeniną ir guaniną – žymėti vienu simboliu



5 pav. Vienmatis DNR sekos savybių vaizdavimas dvimatėje plokštumoje. X ašis nusako sekos nukleotido poziciją (iš viso – 1370 nukleotidų), o y ašyje nurodytas sekos biologinės savybės pasiskirstymas [25]

ar skaičiumi, o kitą grupę, pirimidinus – citoziną ir timiną – žymėti kitu simboliu. Taip pat galima išskirti abėcėles, kurios vietoje nukleotidų nusako aminorūgščių savybes. Kadangi aminorūgštys aprašomos trijų iš eilės esančių nukleotidų junginiu, vietoje kiekvieno sekos nukleotido savybės vaizdavimo algoritmas parodytų aminorūgšties, kuriai nukleotidas priklauso, savybę. Žinoma, jog nukleotidas gali priklausyti trims skirtingoms aminorūgštims priklausomai nuo to, kokiam skaitymo rėmeliui priklauso – todėl renkantis aminorūgštis nusakančias abėcėles reikėtų nagrinėti visus tris galimus skaitymo rėmelius. Detalesnis galimų abėcėlių skirstymas aprašytas [25] straipsnyje.

DNR sekos vizualizuojamos ir skirtingose dimensijose, taip parodant unikalias biologine]s sekos savybes. Pavyzdžiui, 5 paveiksle pavaizduotas purinų ir pirimidinų išsidėstymas sekoje. Kreivė gauta naudojant (2) formulę.

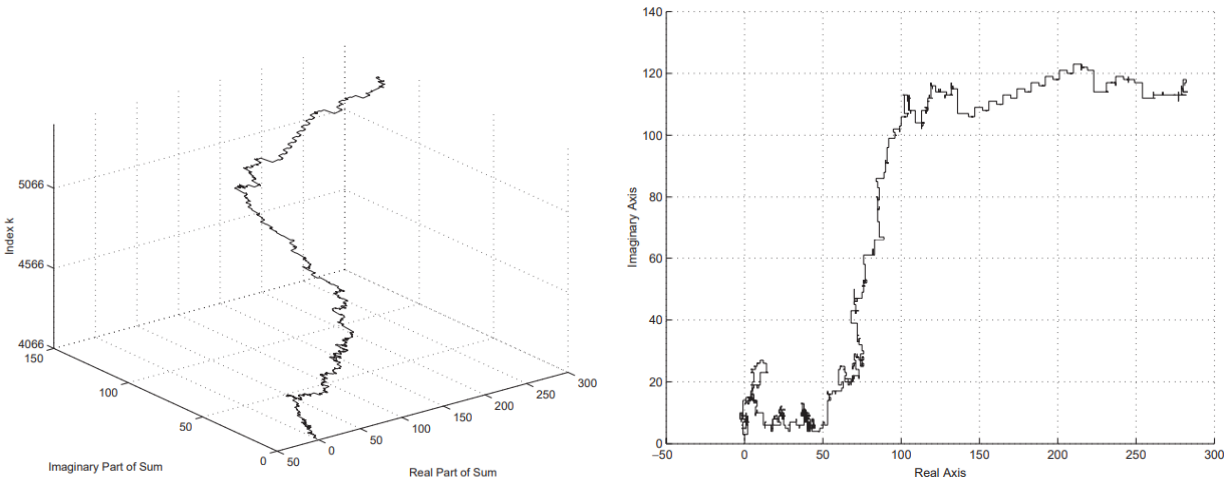
$$s[k] = \sum_{i=1}^{k} x[i], \quad \textbf{(2)}$$

čia x nusakoma nukleotido savybė: jeigu nukleotidas yra purinas, x reikšmė yra 1, o jeigu pirimidinas, x reikšmė yra –1.

Sekos vizualizavimui pasirinkus vienmatį vaizdavimo būdą (5 paveikslas) apsiribojama tik tomis abėcėlėmis, kurioms nusakyti reikalingi du simboliai (t.y. binarinis klasifikavimas). Jeigu pasirenkama abėcėlė, kuri nukleotidus skirsto į daugiau nei dvi grupes, seka vaizduojama dvimatėje arba trimatėje erdvėje.

Dvimačio sekos vaizdavimo pavyzdys pateiktas 6 paveiksle. Kreivei gauti naudojama (1) formulė, tačiau šiuo atveju x žymi nukleotidus, jiems priskyrus atitinkamas reikšmes iš kompleksinių skaičių

aibės [25]. Dvimatis vaizdavimas yra informatyvesnis – kadangi kiekvienas nukleotidas turi jam priskirtą unikalią skaitinę vertę ir nėra priskirtas jokiai bendrai grupei, vizualizacija parodo konkrečių nukleotidų dėsningumus sekoje. Pavyzdys parodo DNR sekos struktūros unikalumą – vizualizacijoje galima pastebėti ir pasikartojantį kreivės išlinkimą, primenantį laiptus, ir išskirtinę, daugiau

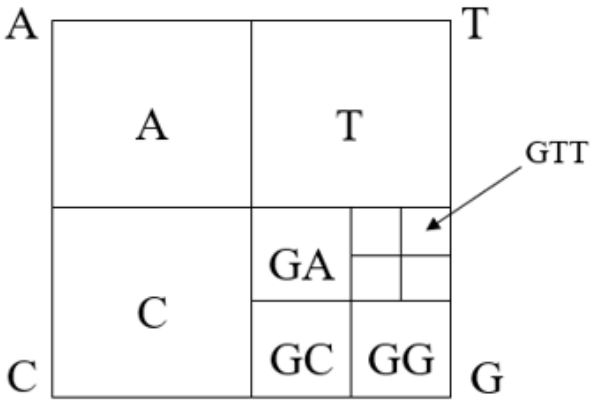


6 pav. Dvimatis DNR sekos savybių vaizdavimas trimatėje plokštumoje (kairėje) ir dvimačio vaizdavimo projekcija į dvimatę plokštumą (dešinėje) [25].

nepasikartojančią kreivės dalį.

Svarbu pažymėti, jog atsitiktinio ėjimo algoritmas naudojamas trumpoms sekoms ar trumpiems sekų fragmentams pavaizduoti. Vaizduojant ilgą seką neįmanoma pastebėti visų kreivės linkių, žyminčių išskirtinę DNR struktūrą, todėl vaizdavimo metodas tampa neefektyviu. Ilgoms nukleotidų sekoms ar jų fragmentams pavaizduoti tikslingiau naudoti fraktalais grįstą (angl. *fractal–based*) vizualizavimo metodą.

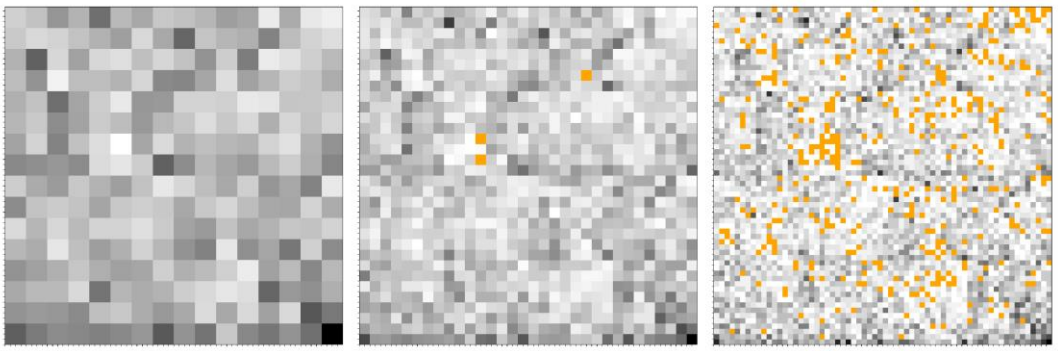
Fraktalais grįstas vizualizavimo algoritmas naudojamas trumpesnių sekos fragmentų dažniui bei trūkstamiems fragmentams sekoje vizualizuoti. Vaizduojant sekos fragmentus metodas neatsižvelgia į visos sekos nukleotidų eiliškumą.



7 pav. Fraktalais grįsto vizualizavimo algoritmo schema

Algoritmo veikimo principas parodytas 7 paveiksle. Kvadratinės formos plokštuma dalinama į keturias lygias dalis, kurios atitinka keturis galimus nukleotidų tipus (adeniną, guaniną, timiną, citoziną). Tuomet kiekvienas mažesnis kvadratas taip pat dalinamas į keturias dalis – nauji mažesni

kvadratai taip pat atitinka keturias galimas bazes. Visi kvadratai dalinami tiek kartų, kiek užtenka pažymėti trumpesniam sekos fragmentui – pavyzdžiui, norint pažymėti aštuonių nukleotidų seką, kvadratinė plokštuma bus dalinama aštuonis kartus, o norint pažymėti trijų nukleotidų ilgio seką,

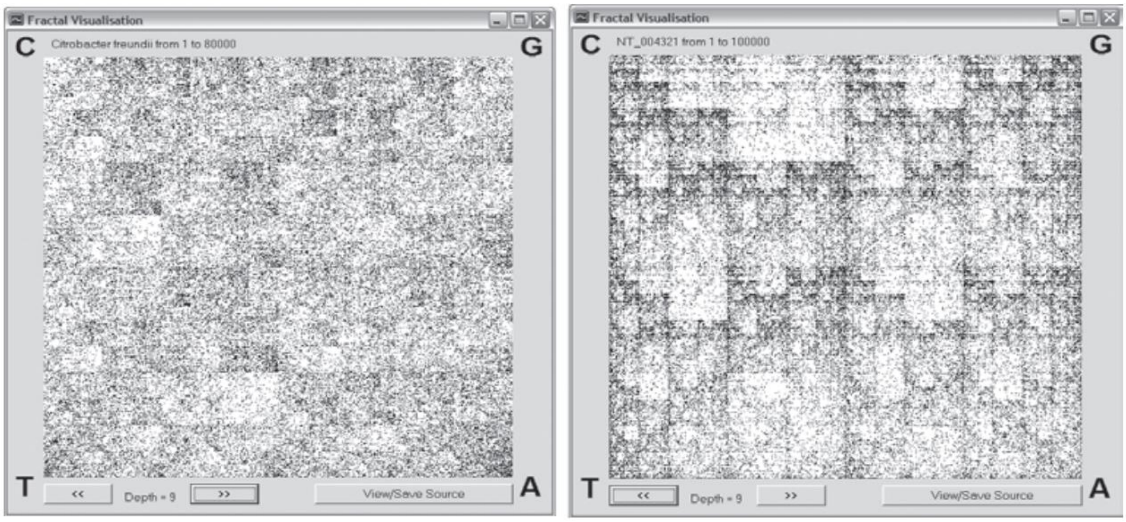


8 pav. E.coli bakterijos DNR sekos vizualizacija fraktalais grįstu algoritmu pasirinkus skirtingą k (sekos fragmento ilgį). Kairysis vaizdas gaunamas kai k = 4, vidurinysis – kai k =5, o dešinysis gaunamas, kai k = 6. Oranžinės spalvos plotai žymi atitinkamo fragmento trūkumą sekoje [26].

plokštumą reikės dalinti tris kartus (kaip parodyta 7 paveiksle) [23].

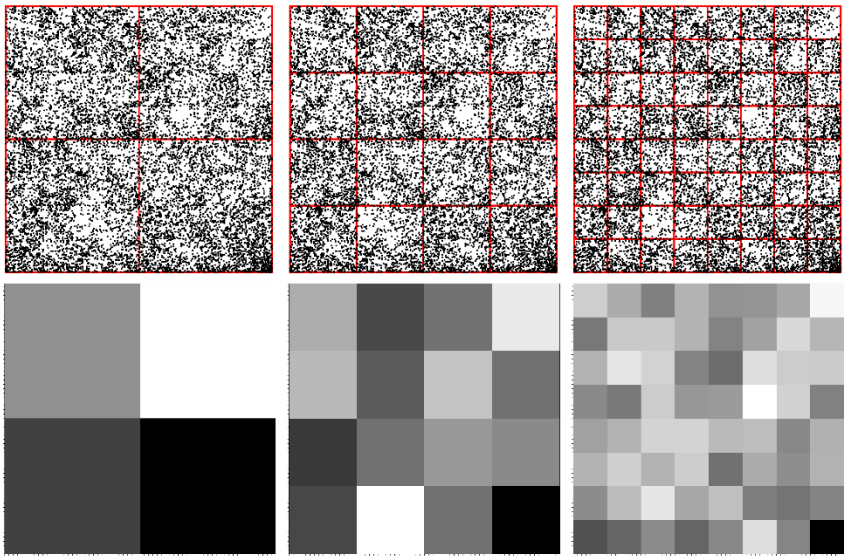
Ilgai sekai pavaizduoti algoritmas naudoja trumpesnius sekos fragmentus. Kai pasirinkto ilgio fragmentas yra pavaizduojamas, pirmas ilgos sekos nukleotidas yra praleidžiamas ir vaizduojamas naujas to paties ilgio fragmentas. Gaunamas rezultatas vizualizuoja ne tik trūkstamus fragmentus, bet ir sekoje esančių fragmentų dažnį (kuo ploto spalva tamsesnė, tuo dažniau fragmentas pasikartoja sekoje).

Kadangi sekai pavaizduoti algoritmas naudoja trumpesnius sekos fragmentus, gaunamo rezultato detalumas priklauso nuo pasirenkamo fragmento ilgio *k*, t.y. kuo ilgesni fragmentai yra pasirenkami, tuo gaunamas vaizdas yra detalesnis. Pavyzdžiui, pasirinkus seką nagrinėti trijų nukleotidų ilgio fragmentais, galimų skirtingų nukleotidų kombinacijų gali būti šešiasdešimt keturios ($4^3 = 64$) – todėl ir vaizdas užims šešiasdešimt keturių kvadratų plotą. Jeigu nagrinėjami fragmentai bus aštuonių nukleotidų ilgio, galimų unikalių nukleotidų kombinacijų skaičius padidėja iki 65536 – tiek kvadratų



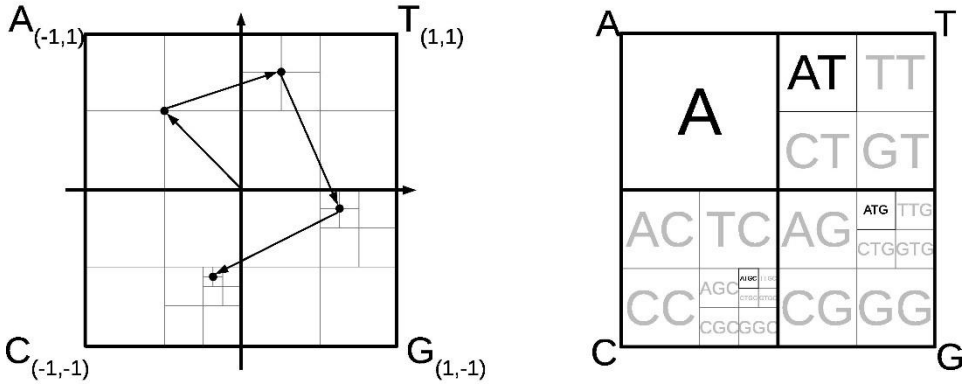
9 pav. Fraktalais grįstu algoritmu gauti sekų vizualizacijų pavyzdžiai. Kairėje pateikiama bakterijos *Citrobacter freundii* vizualizacija, dešinėje – žmogaus chromosomos dalies (NT 004321) vizualizacija, kurioje galima pastebėti fraktalines struktūras [24].

sudarys galutinį vaizdą, kuris bus detalesnis už trijų nukleotidų fragmentų metodu sudarytą vaizdą 1024 kartus [23]. Skirtingo detalumo vaizdų pavyzdžiai pavaizduoti 8 paveiksle.



10 pav. *E.coli* bakterijos sekos vizualizacija. Viršutinėje eilutėje matomas raudonos spalvos tinklelis, žymintis nukleotidų pozicijas. Apatinėje eilutėje matomas nukleotidų skaičiaus kiekviename kvadrate dažnių palyginimas – juoda spalva žymi didžiausią skaičių palyginus su kitais kvadratais, balta – mažiausią [26].

Vaizdus, gaunamus fraktalais grįstu algoritmu galima skirstyti į dvi grupes: vienų vaizdų taškų pasiskirstymas gali atrodyti atsitiktinis, tačiau kai kuriais atvejais pavaizdavus seką rezultate galima pastebėti fraktalines struktūras (t.y. atskiri vaizdo fragmentai primena visą gautą vaizdą). Šie du pavyzdžiai pateikti 9 paveiksle.



11 pav. Chaoso žaidimo algoritmo schema [26].

Turint detalų vaizdą, kuris buvo gautas vaizdavimui pasirinkus didelį sekos fragmento ilgį *n*, rezultatą galima nagrinėti ir mažesnių fragmentų atžvilgiu. Tokio nagrinėjimo pavyzdys pateiktas 10 paveiksle. Viršutinėje eilutėje pateikta *E.coli* bakterijos sekos vizualizacija. Visuose viršutinės eilės paveiksluose matomas skirtingo mastelio raudonas tinklelis, kuris žymi trumpesnius sekos fragmentus. Kadangi vaizde sunku pastebėti taškų pasiskirstymo plokšumoje dėsningumus ir trūkstamus fragmentus bei jų kiekį, vaizdas suskirstomas į mažesnius vienodo dydžio kvadratus *k* (tačiau $k < n$), o kiekvienaname kvadrate esantis taškų kiekis suskaičiuojamas. Didžiausią kiekį taškų

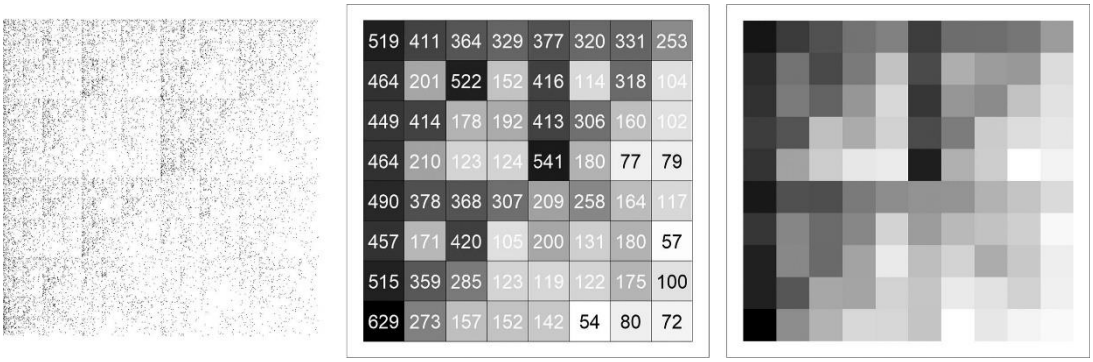
turinčiam kvadratui priskiriama juoda spalva, o visiems kitiems – pilkos spalvos atspalviai, šviesėjantys mažėjant taškų skaičiui kvadrate [26].

Fraktalais grįstas algoritmas dažnai yra siejamas ir yra pagrįstas kitu algoritmu, vadinamu chaoso žaidimo (angl. *chaos game representation*) vizualizavimo metodu. Chaoso žaidimo algoritmo schema (pavaizduota 11 paveiksle) primena fraktalais grįsto algoritmo schemą: turima kvadratinė plokštuma, kuri suskirstoma į mažesnius vienodo dydžio kvadratus, reprezentuojančius unikalų sekos fragmentą. Tiesa, algoritmai nėra identiški – chaoso žaidimo algoritmas nevizualizuoja sekos ją dalindamas į trumpesnius fragmentus. Šiuo atveju yra vaizduojama visa seka nenutrūkstamai, o kiekviena taško pozicija, reprezentuojanti nukleotidą, priklauso nuo prieš tai buvusio taško pozicijos. Tokiu būdu kiekvienas taškas reprezentuoja trumpesnę sekos dalį, kuri visuomet prasideda nuo sekos pradžios.

Taško poziciją koordinatėse galima nusakyti lygtimi (3).

$$P_i^j = P_{i-1}^j + sf\left(V_{i-1}^j - P_{i-1}^j\right), \quad \textbf{(3)}$$

čia P_0^j – pradžios taškas, kuris yra arba parenkamas atsitiktinai, arba nustatomas iš anksto; j – algoritmo dimensija, DNR atveju 2; i – sekos S nukleotido pozicija; sf – mastelio koeficientas, DNR atveju 0,5. V žymi viršūnių koordinates: $V_i^0 = 1$ jeigu S_i yra T arba G, kitu atveju $V_i^0 = -1$; $V_i^1 = 1$ jeigu S_i yra A arba T, kitu atveju $V_i^1 = -1$ [27].



12 pav. Chaoso žaidimo dažnių matricos vizualizacijos metodo pavyzdys. Kairėje vaizduojamas chaoso žaidimo metodu gaunamas vaizdas, viduryje – kvadratuose esančių taškų dažniai, dešinėje – galutinis vaizdas [27].

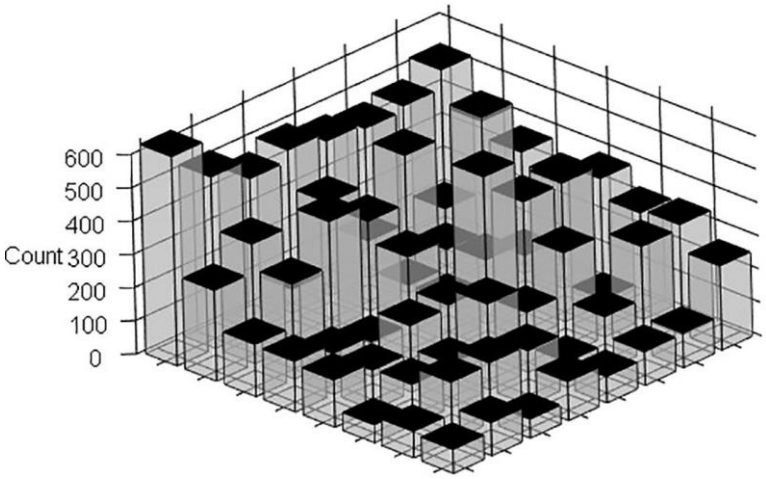
Svarbu pažymėti, jog pradinio taško koordinatės didelės įtakos vaizdo raštui nedaro – nesvarbu, nuo kurio taško yra pradedamas algoritmas, raštas visuomet bus identiškas, o skirtumą tarp vaizdų su skirtingomis pradinėmis taško koordinatėmis galima pastebėti tik vaizdus palyginus specialiais vaizdų palyginimo metodais [27].

Priešingai nei fraktalais grįstame algoritme, chaoso žaidimo algoritmas kiekvieną nukleotidą sekoje pažymi tašku, neleidžiant pasirinkti vaizdo detalumo. Norint patikrinti sekos fragmentų dažnį, vaizdas padalinamas į lygaus ploto kvadratus, kurių skaičius yra tiesiogiai proporcingas fragmentų ilgiui. Pavyzdys su trijų nukleotidų ilgio fragmentais pavaizduotas 12 paveiksle. Vaizdas suskirstomas į šešiolika lygių kvadratų, o kiekviename kvadrate esantis taškų skaičius yra sumuojamas. Skirtingomis spalvomis pateikiami skirtingus dažnius vaizduojantys kvadratai (tamsiausia spalva pateikiamas daugiausiai taškų turintis, šviesiausia – mažiausiai taškų turintis kvadratas). Šis DNR vizualizavimo metodas vadinamas chaoso žaidimo dažnių matrica (angl. *frequency matrix chaos game representation*) [27].

Chaoso žaidimo dažnių vizualizavimo metodą galima nusakyti lygtimis (4).

$$F = \left(a_{i,j}\right), 1 \le i,j \le 2^k, i,j \epsilon \mathbb{N},$$
$$i = 2^k - \lceil 2^{k-1}(x+1) \rceil + 1,$$
$$j = \lceil 2^{k-1}(y+1) \rceil; \quad \textbf{(4)}$$

čia F yra dažnių matrica, $a_{i,j}$ – matricos elemento pozicija, i ir j – elemento poziciją matricoje nusakantys skaičiai, x ir y – chaoso žaidimo metodu gauto taško koordinatės, k – pasirinktas skaičius, nusakantis gaunamo vaizdo dimensiją ($2^k \times 2^k$).



13 pav. Chaoso žaidimo dažnių matricos vizualizacijos metodo pavyzdys trimatėje erdvėje [27].

Tą patį vaizdą, gaunamą dvimatėje plokštumoje naudojant chaoso žaidimo dažnių matricos algoritmą, galima pavaizduoti ir trimatėje erdvėje. Sekos, pavaizduotos 12 paveiksle, trimatis vaizdas histogramų pavidalu pateiktas 13 paveiksle.

	0	1
1	A (0,1)	G (1,1)
0	C (0,0)	T (1,0)

	00 (0)	01 (1)	10 (2)	11 (3)
11 (3)	AA (00,11)	AG (01,11)	GA(10,11)	GG (11,11)
10 (2)	AC (00,10)	AT (01,10)	GC (10,10)	GT (11,10)
01 (1)	CA (00,01)	CG (01,01)	TA (10,01)	TG (11,01)
00 (0)	CC (00,00)	CT (01,00)	TC (10,00)	TT (11,00)

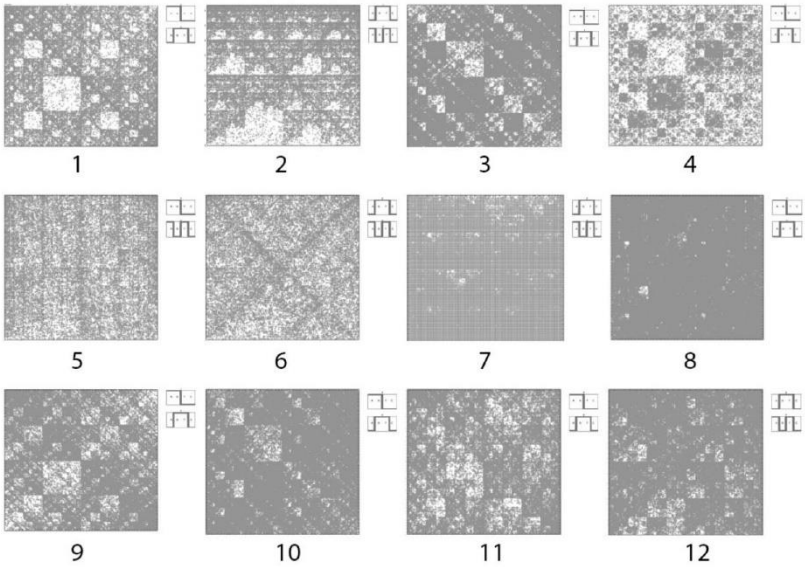
	000 (0)	001 (1)	010 (2)	011 (3)	100 (4)	101 (5)	110 (6)	111 (7)
111 (7)	AAA (000,111)	AAG (001,111)	AGA (010,111)	AGG (011,111)	GAA (100,111)	GAG (101,111)	GGA (110,111)	GGG (111,111)
110 (6)	AAC (000,110)	AAT (001,110)	AGC (010,110)	AGT (011,110)	GAC (100,110)	GAT (101,110)	GGC (110,110)	GGT (111,110)
101 (5)	ACA (000,101)	ACG (001,101)	ATA (010,101)	ATG (011,101)	GCA (100,101)	GCG (101,101)	GTA (110,101)	GTG (111,101)
100 (4)	ACC (000,100)	ACT (001,100)	ATC (010,100)	ATT (011,100)	GCC (100,100)	GCT (101,100)	GTC (110,100)	GTT (111,100)
011 (3)	CAA (000,011)	CAG (001,011)	CGA (010,011)	CGG (011,011)	TAA (100,011)	TAG (101,011)	TGA (110,011)	TGG (111,011)
010 (2)	CAC (000,010)	CAT (001,010)	CGC (010,010)	CGT (011,010)	TAC (100,010)	TAT (101,010)	TGC (110,010)	TGT (111,010)
001 (1)	CCA (000,001)	CCG (001,001)	CTA (010,001)	CTG (011,001)	TCA (100,001)	TCG (101,001)	TTA (110,001)	TTG (111,001)
000 (0)	CCC (000,000)	CCT (001,000)	CTC (010,000)	CTT (011,000)	TCC (100,000)	TCT (101,000)	TTC (110,000)	TTT (111,000)

14 pav. Matricos vaizdavimo metodo pavyzdys. Parodyti trys galimi matricų variantai pasirinkus skirtingą sekos fragmento ilgį (seka nagrinėjama po vieną fragmentą, du ir tris). Raudonai pažymėti skaičiai skliaustuose žymi binarinių skaičių atitikmenį dešimtainiais skaičiais [28].

Dar vienas algoritmas, panašus į fraktalais grįstą algoritmą, vadinamas matricos vaizdavimo metodu. Metodas seką taip pat vaizduoja kvadratinėje plokštumoje, tačiau gaunamas rezultatas priklauso nuo vaizdavimui pasirinktų biologinių sekos savybių. Turima DNR seka paverčiama į binarinį skaičių kiekvieną nukleotidą pakeičiant nuliu arba vienetu – kiekvienu atveju skaitmuo žymi nukleotido biologinę sandarą, pavyzdžiui, ar nukleotidas yra pirimidinas ar purinas, turi stiprų ar silpną ryšį, turi keto ar amino bazę. Norint nusakyti visas tris savybes, reikalingos trys binarinių skaičių sekos. Kiekviena seka paverčiama į taško koordinačių seką, jos fragmentus iš binarinių skaičių paverčiant dešimtainiais. Algoritmo schema pavaizduota 14 paveiksle [28].

Matricos vaizdavimo metodu gauti vaizdai taip pat gali turėti skirtingą detalumą. Vaizdo detalumas priklauso nuo pasirenkamo sekos fragmento ilgio – kuo seka ilgesnė, tuo detalumas yra didesnis. Tačiau didinant sekos fragmento ilgį taškų skaičius mažėja, o vaizdo dimensijos didėja (fragmento ilgį padidinus vienetu, dimensijos padidėja dvigubai).

Naudojant matricų vaizdavimo algoritmą tą pačią seką galima pavaizduoti trimis skirtingais vaizdais dvimatėje plokštumoje (neįskaitant galimų ašių apsukimų ir sukeitimų). Kiekvienas gaunamas vaizdas yra skirtingas nepaisant to, jog vaizduojama DNR seka yra tik viena. Įvairūs matricų metodu vizualizuotų sekų pavyzdžiai pavaizduoti 15 paveiksle. Nors visi vaizdai paveiksle yra skirtingi, kai kurie jų vaizduoja tas pačias sekas. Pavyzdžiui, pirmas ir antras vaizdai 15 paveiksle pateikia vizualizuotą žmogaus proteino geną (CNTNAP2). Galima pastebėti, jog vaizdai skiriasi, tačiau abiejuose pavyzdžiuose galima pastebėti fraktalines struktūras.



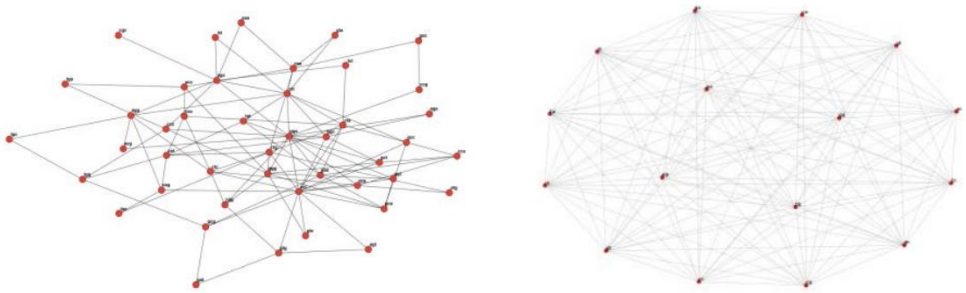
15 pav. Matricos vaizdavimo algoritmu gautų vaizdų pavyzdžiai. 1 ir 2, 5 ir 6, 11 ir 12 vaizdai gauti naudojant vienodas DNR sekas, tik skirtingai išreikštas binariniais skaičiais [28].

Vizualizavus genetines sekas vienmatėje, dvimatėje ar trimatėje erdvėje gautuose rezultatuose galima pastebėti pasikartojančius raštus. Tokie raštai dažnai vadinami fraktalinėmis mozaikomis, fraktalinėmis gardelėmis (angl. *fractal lattices*) arba fraktalus primenančiais raštais [28]. Dvimatėje plokštumoje gauti rezultatai yra siejami su Kroneckerio matricų daugybos operacija – kiekvieną kartą matricą pakėlus laipsniu ir rezultatą vizualizavus, gaunamas vaizdas raštu primena DNR sekos vizualizaciją [28]. Tokiu būdu parodoma, jog vaizdo raštą, primenantį fraktalines struktūras, lemia pasikartojantis mažesnis tam tikrų sekos fragmentų kiekis, o ryšys tarp fragmentų yra pastebimas tik

seką pavaizdavus plokštumoje. Matricų vizualizavimo metodu genetinę seką pavaizdavus trimatėje erdvėje gaunama figūra primena Sierpinskio trikampį [28], tačiau

1.4. Rečiau naudojami vaizdų vizualizavimo metodai

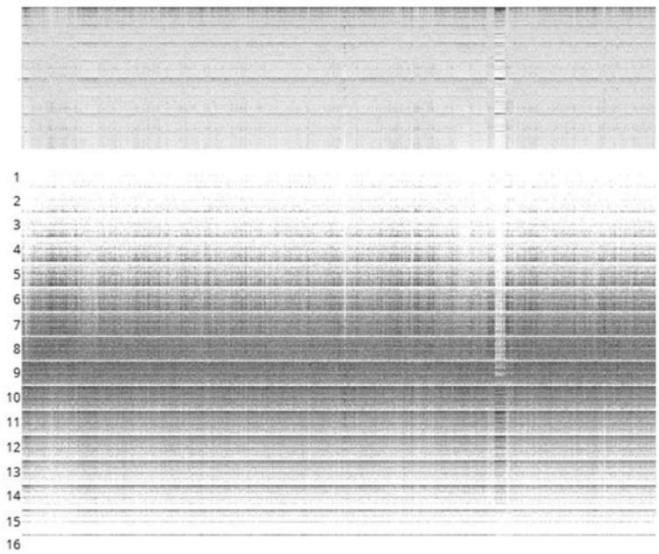
Siekiant pavaizduoti seką plokštumoje dažniausiai naudojami matricų, chaoso žaidimo ir fraktalais grįsti vizualizavimo algoritmai. Tačiau yra ir kitų, ne tokių populiarių algoritmų, kurie neturi jokių panašumų su anksčiau aptartais metodais. Vienas iš tokių metodų yra sekos vaizdavimas grafu [29]. Grafas braižomas pagal nustatytas taisykles: jo viršūnės nusako skirtingus vienodo ilgio sekos fragmentus, o briauna tarp viršūnių nusako atitinkamų fragmentų gretimą poziciją sekoje. Briaunos gali būti kryptingos siekiant parodyti, šalia esančių fragmentų poziciją sekoje. Grafais pavaizduotų sekų pavyzdžiai pateikti 16 paveiksle.



16 pav. Insulino geno vizualizavimas grafais. Kairėje pateikiamas grafas, kurio viršūnės nusako 3 nukleotidų ilgio fragmentus, o dešinėje – dviejų nukleotidų ilgio fragmentus [29].

Nors dažniausiai pasirenkama DNR seką vaizduoti dvimatėje plokštumoje, sekos sandaros ypatumus galima pastebėti ir ją pavaizdavus vienmatėje ar trimatėje erdvėje. Vaizdavimo vienmatėje erdvėje pavyzdžiu galima laikyti spektrinį DNR sekos skaidymą (angl. *spectral decomposition*), kurio vaizdavimo pavyzdys pateikiamas 17 paveiksle [30].

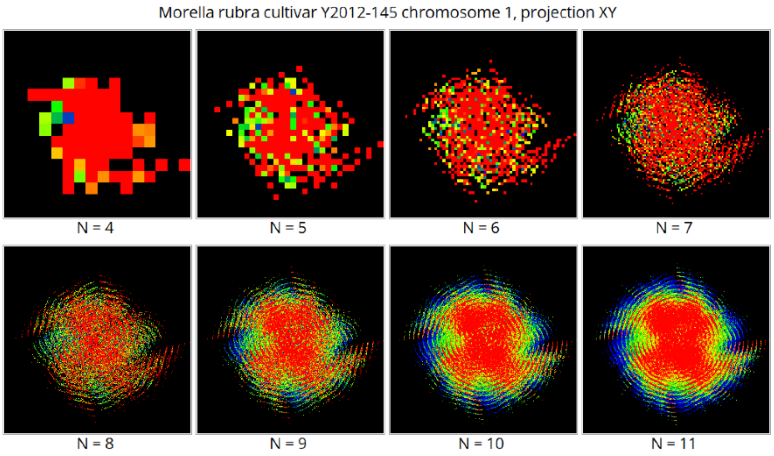
Spektrinio skaidymo metodu pavaizduota seka yra padalijama į sritis, lygias pasirinktam parametrui N. Srityse pažymimi tik tie taškai, kurių binarinio kodo atitikmens vienetų sumos reikšmė yra lygi N



17 pav. DNR sekos vizualizavimas spektrinio skaidymo metodu [30].

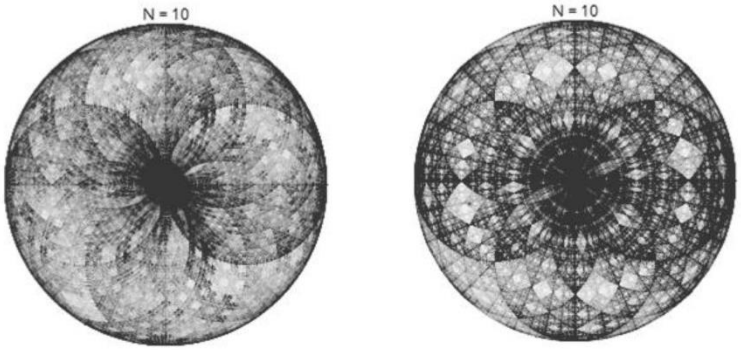
reikšmei [31]. Metodas parodo sekos nukleotidų išsidėstymo ypatumus pagal vaizdavimui pasirinktą abėcėlę [30]. Vienmatėje plokštumoje sekas galima vaizduoti ir daugialypės kompozicijos (angl. *multiscale composition*) metodu, kuris parodo genomo fraktalinę struktūrą vaizdavimui naudojant skirtingus mastelius [30].

DNR sekas galima vaizduoti ir polinėje koordinačių sistemoje. Tokiu būdu galima pavaizduoti vieną sekos biologinę savybę, išreikštą binariniais skaičiais. Kiekvienas vektorius polinėse koordinačių sistemose nusako skirtingą sekos fragmentą, išreikštą binariniu skaičiumi, o vektoriaus ilgis lygus vienetų skaičiui fragmente [30]. Sekos vaizdavimo polinėje koordinačių sistemoje pavyzdys pateiktas 18 paveiksle.



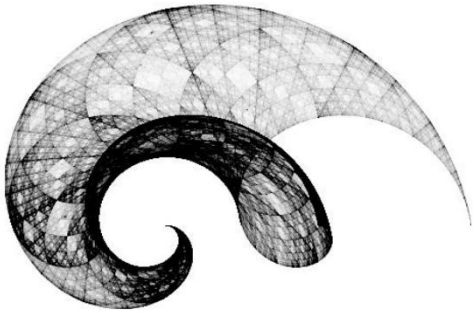
18 pav. DNR sekos vizualizavimas polinėje koordinačių sistemoje [30].

Sekos taip pat gali būti vaizduojamos žiediniais vaizdavimais (angl. *ring projections*), kurie gaunami kiekvieną binarinį vektorių vaizduojant apskritimu. Keičiant metodų parametrų reikšmes ir vaizdavimui pasirinktų abėcėlių kiekį, galima gauti įvairius skirtingus sekos vaizdavimus, pavaizduotus 19 ir 20 paveiksluose.



19 pav. Žiedinis DNR sekos vaizdavimas [31].

Nors DNR sekų vaizdavimų būdų yra įvairių (žiediniai vaizdavimai galimi ir trimatėje plokštumoje), ne visi jie yra vienodai informatyvūs. Nors spektrinis sekos skaidymas ir įvairūs žiediniai vaizdavimai taip pat vizualizuoja fraktalines struktūras, žiediniai vaizdavimai neparodo konkrečių sekos fragmentų trūkumo ir nevizualizuoja sekos eilės tvarka.



20 pav. Žiedinis DNR sekos vaizdavimas, kai vektoriaus ilgis lygus sekos fragmento ilgiui [31].

1.5. Vaizdų palyginimo metodai

Siekiant palyginti vaizdavimo metodus yra pasirenkami įvairūs metodais gautų vaizdų palyginimo parametrai. Vienas iš jų yra struktūrinio panašumo indeksas (kitaip – SSIM). Parametras vertina tris skirtingus vaizdų aspektus: vaizdo intensyvumą, struktūrą ir kontrastą. Kiekvieną iš šių parametrų galima nusakyti funkcijomis (5) [32].

$$l(x,y) = \frac{2\bar{x}\bar{y}+c_1}{\bar{x}^2+\bar{y}^2++c_1},$$

$$c(x,y) = \frac{2s_x s_y+c_2}{s_x^2+s_y^2+c_2},$$

$$s(x,y) = \frac{s_{x,y}+c_3}{s_x s_y+c_3}; \quad \textbf{(5)}$$

čia l – funkcija, lyginanti dviejų vaizdų intensyvumą, c – funkcija, lyginanti vaizdų kontrastą, s – funkcija, lyginanti vaizdų struktūrą, x ir y – lyginami vaizdai, $\bar{x}$ ir $\bar{y}$ – x ir y vidurkiai, s_x^2 ir s_y^2 – x ir y dispersijos, $s_{x,y}$ – kovariacija tarp x ir y. Mažos konstantos c_1, c_2 ir c_3 universaliu atveju yra lygios nuliui.

Bendru atveju struktūrinį panašumą galima nusakyti (6) formule. Kuo gaunama reikšmė yra didesnė, tuo panašesni yra lyginami vaizdai.

$$SSIM(x,y) = l(x,y)c(x,y)s(x,y); \quad \textbf{(6)}$$

čia SSIM – struktūrinio panašumo indeksas, l – funkcija, lyginanti dviejų vaizdų intensyvumą, c – funkcija, lyginanti vaizdų kontrastą, s – funkcija, lyginanti vaizdų struktūrą.

Vienu paprasčiausių vaizdų palyginimo metodų išlieka vaizdų atimtis. Metodas įvykdomas iš vieno vaizdo pikselių atimant kito vaizdo pikselius – todėl svarbu, jog vaizdų dimensijos būtų vienodos. Vaizdų atimties metodą galima apibendrinti (7) formule.

$$b(m,n) = f_1(m,n) - f_2(m,n); \quad \textbf{(7)}$$

čia b – vaizdų atimties rezultatas, f_1 ir f_2 – lyginami vaizdai, o m ir n – pikselio koordinatė vaizde (eilutė ir stulpelis) [33].

Pirsono koreliacijos koeficientas taip pat naudojamas vaizdų skirtumams palyginti (8). Kuo gaunamas koeficientas yra mažesnis, tuo labiau vaizdai skiriasi vienas nuo kito [34].

$$r = \frac{\sum_i (x_i - x_m)(y_i - y_m)}{\sqrt{\sum_i (x_i - x_m)^2} \sqrt{\sum_i (y_i - y_m)^2}}; \quad \textbf{(8)}$$

čia r – koreliacijos koeficientas, x_i – i – tojo pikselio intensyvumas pirmame vaizde, y_i – i – tojo pikselio intensyvumas antrame vaizde, x_m – pirmojo vaizdo pikselių intensyvumo vidurkis, y_m – antrojo vaizdo pikselių intensyvumo vidurkis.

1.6. Aktualumas

Remiantis išnagrinėta moksline literatūra, akivaizdu, jog DNR sekų analizė vis dar yra aktuali tyrimų sritis. Dėl didelio duomenų kiekio nespėjama tirti visų DNR sekų, sunku rasti ir nagrinėti sekose esančius struktūros skirtumus – mutacijas, kurios ilgainiui gali sukelti įvairias ligas. Dėl šių priežasčių vaizdų vizualizavimo metodai vis dar yra aktualūs – jie leidžia net ir ilgai DNR sekas greitai ir nesudėtingai pavaizduoti vienmatėje, dvimatėje ar trimatėje plokštumoje, kartu vizualizuojant ir sekos struktūros savybes, kurias skaitiniais metodais gali būti sunku pastebėti.

Analizuojant literatūrą buvo aptarti keli skirtingi metodai sekų vizualizavimui, tačiau ne visi jie yra vienodai informatyvūs. Daugiausiai informacijos pateikiama seką vizualizavus chaoso žaidimo, chaoso žaidimo dažnių ir matricų vizualizavimo metodais. Chaoso žaidimas gali vizualizuoti ilgas sekas atsižvelgiant į nukleotidų išsidėstymo tvarką sekoje – kiti metodai dažnai į sekos eiliškumą neatsižvelgia.

Dažniausiai pasirenkama vizualizuoti virusų, bakterijų ar žmogaus genų DNR sekas, kurios nėra ilgos ir paprastai neviršija 10 mln. nukleotidų ilgio. Ilgesnės, atsitiktinai sugeneruotos ir logistine regresija gautos sekos vizualizuojamos retai, o gaunami rezultatai tarpusavyje palyginti dar nebuvo. Nors fraktalinės struktūros susidarančios vaizde jau buvo pastebėtos, rašto pasikeitimai atsižvelgiant į vaizduojamo fragmento ilgį nagrinėti nebuvo.

Todėl šiame magistriniame darbe bus detaliau tiriami trys anksčiau paminėti vizualizavimo metodai – chaoso žaidimo, chaoso žaidimo dažnių ir matricų metodai. Metodais gaunamus vaizdus pasirinkta palyginti keliais skirtingais vaizdų palyginimo parametrais: struktūrinio panašumo indeksu, Pirsono koreliacijos koeficientu ir vaizdų atėmimo metodu. Taip pat pasirinkta įvertinti genetinės sekos, atsitiktinai sugeneruotos sekos ir logistine regresija gautos sekos vizualizavimo rezultatus juos palyginant tarpusavyje. Vizualizavimui pasirinktos ir ilgos, ir trumpos DNR sekos, siekiant įvertinti metodų tinkamumą ilgoms ir trumpoms sekoms.

2. Duomenys ir tyrimo metodai

Skyriuje pateikiama informacija apie pasirinktus nagrinėjamus duomenis; taip pat aprašoma metodika duomenų apdorojimui ir jų nagrinėjimui, bei nurodomi matematiniai metodai analizei atlikti.

2.1. Duomenys

Duomenys analizei pasirinkti iš nukleotidų duomenų bazės NCBI. Nagrinėtos kelios žmogaus chromosomų nukleotidų sekos, iš kurių buvo išskirti atskiri fragmentai – koduojančios ir nekoduojančios genų dalys ir keli skirtingi žmogaus genai.

Duomenys pateikti FASTA formatu. Tai – specialus tekstinis formatas, kuriuo pateikiamos nukleotidų arba aminorūgščių (t.y. proteinų) sekos. Kiekvienas simbolis sekoje žymi nukleotidą arba aminorūgštį.

Svarbu pažymėti, jog DNR sekos nuskaitymas ir duomenys, gaunami nuskaitymo metu nėra visiškai tikslūs. Norint nuskaityti visą organizmo genomą, seka yra nuskaitoma mažomis dalimis, kurias sujungus gaunama visa genomo seka. Deja, šis metodas negarantuoja viso genomo nuskaitymo – sekoje dažnai atsiranda tuščių vietų, kurių nuskaityti nepavyksta (kurios įprastai duomenų failuose žymimos „N“ raidėmis). Dėl šios priežasties nuskaitytas sekas reikia pertvarkyti.

Kiekvienas duomenų failas yra pertvarkomas: failo pradžioje esanti informacija apie seką yra ištrinama, paliekama tik organizmo nukleotidų seka. Kiekviena seka nuskaitoma, performuojama į vieną eilutę (originaliame faile vietoje eilutėje yra ~70 simbolių), pašalinami nereikalingi tarpai atsiradę po nuskaitytos matricos pertvarkymo ir pašalinamos „N“ raidės, žyminčios trūkstamus (nenuskaitytus) sekos nukleotidus. Lentelėje pateikti sekos ilgiai gauti po duomenų sutvarkymo.

Gautas rezultatas yra viena duomenų eilutė, kurios ilgis atitinka turimos nukleotidų sekos ilgį. Toliau sutvarkyti duomenys naudojami kiekviename metode individualiai – arba naudojami tokie, kokie yra, arba dar papildomai pertvarkomi (matricų vaizdavimo metodo atveju).

2.1.1. Duomenų skirstymas

Ta pati DNR seka nagrinėta keliais skirtingais būdais. Seka gali būti nagrinėjama visa, arba išskirta į regionus, kurie koduoja ar nekoduoja geno. Norint iš sekos išskirti koduojančius regionus, duomenis reikia pertvarkyti papildomai.

Koduojančios DNR sekos sritys yra išskirtos specialiomis sekos fragmentų kombinacijomis. Geną koduojantis fragmentas visuomet pradedamas ATG nukleotidų kombinacija (kodonu), o pabaigos – TAA/TAG/TGA kombinacijomis. Visos kombinacijos sekoje yra surandamos ir dar kartą pertvarkomos – kadangi po pradžios kodono sekoje turi eiti pabaigos kodonas, o po jo – vėl pradžios, eilės tvarkos neatitinkantys kodonai geno neišskiria ir yra ištrinami. Taip pat patikrinamas pirmas sekoje surastas kodonas, kuris yra išmetamas, jeigu nenusako koduojančios geną pradžios. Tokiu būdu gaunamas vektorius, kuriame yra sekoje esančių pradžios ir pabaigos kodonų koordinatės. Pagal gautą vektorių seka yra atskiriama į dvi dalis – koduojančią geną, ir nekoduojančią.

Svarbu paminėti, jog seka turi tris skirtingus skaitymo rėmelius – todėl reikia tris kartus išskirti sekoje esančias dalis kiekvieno skaitymo rėmelio atžvilgiu.

Taip pat iš sekos galima išskirti fragmentus, nusakančius genus. Genams išskirti iš sekos yra sukurti specialūs programų metodai, tačiau duomenis taip pat galima rasti ir nukleotidų duomenų bazėse.

Darbe naudoti duomenys pateikiami 1 lentelėje.

1 lentelė. Naudotos sekos ir jų ilgiai

Žmogaus genomo dalis	**Sekos ilgis**
1 – a chromosoma	230481012
Pirmo 1 – os chromosomos rėmelio nekoduojanti dalis	156357063
Antro 1 – os chromosomos rėmelio nekoduojanti dalis	155247181
Trečio 1 –os chromosomos rėmelio nekoduojanti dalis	169348387
Pirmo 1 – os chromosomos rėmelio koduojanti dalis	74123949
Antro 1 – os chromosomos rėmelio koduojanti dalis	75233831
Trečio 1 –os chromosomos rėmelio koduojanti dalis	61132625
1 – os chromosomos genas MTHFR	86126
1 – os chromosomos genas MTOR	312054
1 – os chromosomos genas AGT	40487
3 – ia chromosoma	198100135
16 – a chromosoma	81805943
Pirmo 16 – os chromosomos rėmelio nekoduojanti dalis	55120764
Antro 16 – os chromosomos rėmelio nekoduojanti dalis	54563590
Trečio 16 –os chromosomos rėmelio nekoduojanti dalis	59811363
Pirmo 16 – os chromosomos rėmelio koduojanti dalis	26685179
Antro 16 – os chromosomos rėmelio koduojanti dalis	27242353
Trečio 16 –os chromosomos rėmelio koduojanti dalis	21994580
16 – os chromosomos genas GINS2	25514
16 – os chromosomos genas CTRL	4578
16 – os chromosomos genas BCAR1	99630
Atsitiktinė duomenų seka	4440296
VDR genas (12 chromosoma)	126937

Duomenų palyginimui buvo sukurta ir pastovi atsitiktinė duomenų seka. Atsitiktinei DNR sekai sukurti buvo pasinaudota MATLAB programos funkcija *randseq*. Funkcija automatiškai sugeneruoja atsitiktinę nurodyto ilgio DNR seką, sudarytą iš keturių simbolių, nusakančių nukleotidus – A, C, T ir G.

Be žmogaus chromosomų sekų ir atsitiktinai sugeneruotos sekos buvo nagrinėjama ir daugianare logistine regresija gauta seka. Šiai sekai gauti turimi duomenys buvo papildomai performuoti:

1. Seka dubliuojama du kartus;
2. Viena kopija perrašoma taip, kad nusakytų, ar nukleotidas yra pirimidinas ar purinas;
3. Kita kopija perrašoma taip, kad nusakytų, kelis vandenilinius ryšius turi nukleotidas (2 arba 3);
4. Kopijos „pastumiamos“ per vieną vietą į kairę – jos nusakys centrinio nagrinėjamo nukleotido kaimynus iš dešinės;
5. 1 – 4 žingsniai pakartojami dar kartą, tik 4 žingsnyje kopijos pastumiamos į dešinę, kad nusakytų kaimynus iš kairės.

Tokiu būdu pertvarkyti duomenys naudojami logistinės regresijos modelyje trims skirtingiems atvejams – nagrinėjama centrinio nukleotido priklausomybė nuo tik iš kairės, tik iš dešinės, ir iš abiejų pusių esančių kaimynų. Kiekvienas sukurtas daugianarės logistinės regresijos modelis lygintas su nuliniu modeliu (kuriame visos reikšmės turi vienodą tikėtinumą) norint nustatyti daugianarės logistinės regresijos modelio reikšmingumą. Nulinė hipotezė šiuo atveju – jog nulinis modelis ir daugianarės logistinės regresijos modelis yra vienodai statistiškai reikšmingi (modeliai nesiskiria). Jeigu reikšmingumo lygmuo (angl. *p-value*) < 0,05, nulinė hipotezė atmetama ir priimama alternatyvi – jog daugianarės logistinės regresijos modelis geriau nusako duomenis, nei nulinis. Daugianarės logistinės regresijos modelio tinkamumas buvo įvertintas Chi – kvadrato patikimumo kriterijumi. Bendrinė Chi – kvadrato patikimumo kriterijaus formulė nusakyta (9).

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}. \quad \textbf{(9)}$$

čia O_i – stebėta vertė (tikroji vertė), E_i – tikėtina vertė [35].

Gaunamos sekos pavaizduojamos trimis vizualizavimo metodais, o vaizdai yra palyginami rezultatų skiltyje.

2.2. Tyrimo metodai

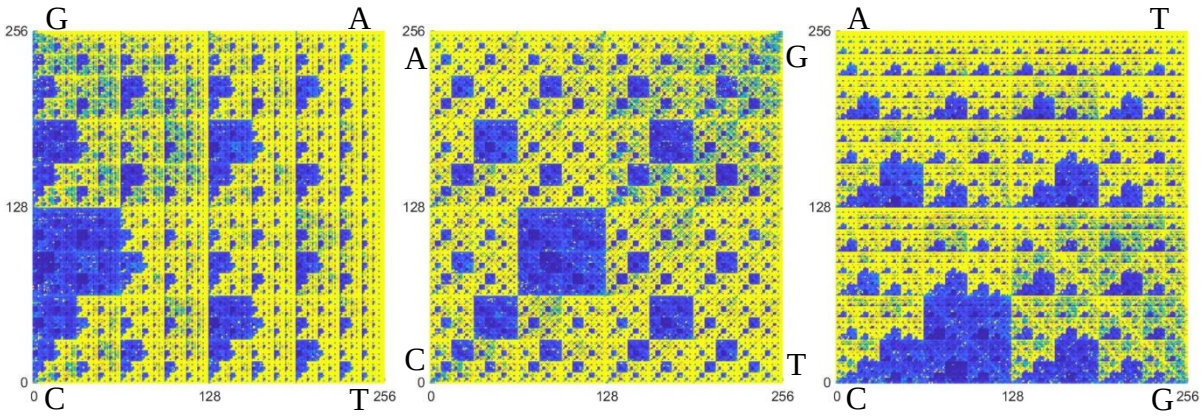
2.2.1. Matricų vizualizavimo metodas

Kiekvienam sekos variantui raidine seka turimi duomenys yra performuojami į skaičius pagal toliau aprašytas taisykles:

1. Išsaugomos trys sekos kopijos;
2. Pirma kopija perrašoma taip, kad nusakytų, ar nukleotidas yra pirimidinas ar purinas (jeigu nukleotidas yra pirimidinas, jam priskiriamas 0, jeigu purinas - 1);
3. Antra kopija perrašoma taip, kad nusakytų, kelis vandenilinius ryšius turi nukleotidas (jeigu nukleotidas turi silpną ryšį, jam priskiriamas 1, jeigu stiprų - 0);
4. Trečia kopija perrašoma taip, kad nusakytų, kokioms bazių grupėms nukleotidas priklauso (jeigu nukleotidas turi amino bazę, jam priskiriamas 0, jeigu keto - 1).

Pavyzdžiui, turint GCATATTGC seką, ji bus perrašoma taip:

1. 101010010 (pagal purinų – pirimidinų savybės taisyklę);
2. 001111100 (pagal vandenilių ryšių stiprumo taisyklę);
3. 100101110 (pagal keto – amino bazių grupių taisyklę).

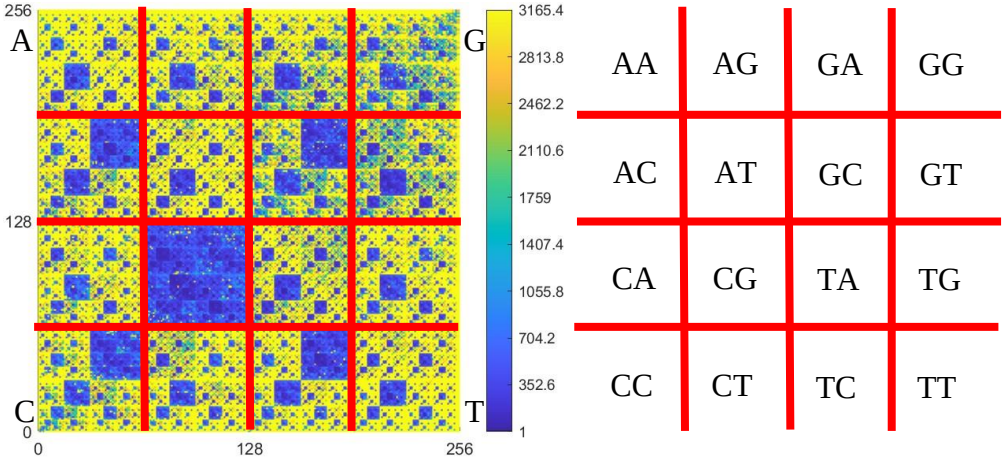


21 pav. Matricų vizualizavimo metodu pavaizduota trečios žmogaus chromosomos seka, kai $N = 8$. Pirmas vaizdas gautas naudojant ryšių stiprumo ir heterociklinių bazių koordinates, antras – naudojant bazių grupių ir heterociklinių bazių koordinates, trečias – naudojant bazių grupių ir ryšių stiprumo koordinates.

Turint tris skirtingas binarinių skaičių sekas, nusakančias genomo savybes, kiekvienai iš jų yra pritaikomas žemiau aprašytas algoritmas:

1. Pasirenkamas mažesnės sekos ilgis N.
2. Nuo pradinės binarinės K ilgio sekos pradžios paimama N ilgio seka, kuri yra paverčiama į dešimtainį skaičių (koordinatės dalį) ir yra išsaugoma.
3. Pirmas K ilgio sekos skaičius yra ištrinamas, ir nuo sekos pradžios paimama nauja N ilgio seka, kuri taip pat yra paverčiama į dšimtainį skaičių ir yra išsaugoma.
4. 2-3 žingsniai yra kartojasi tol, kol pasiekiamas K sekos galas.

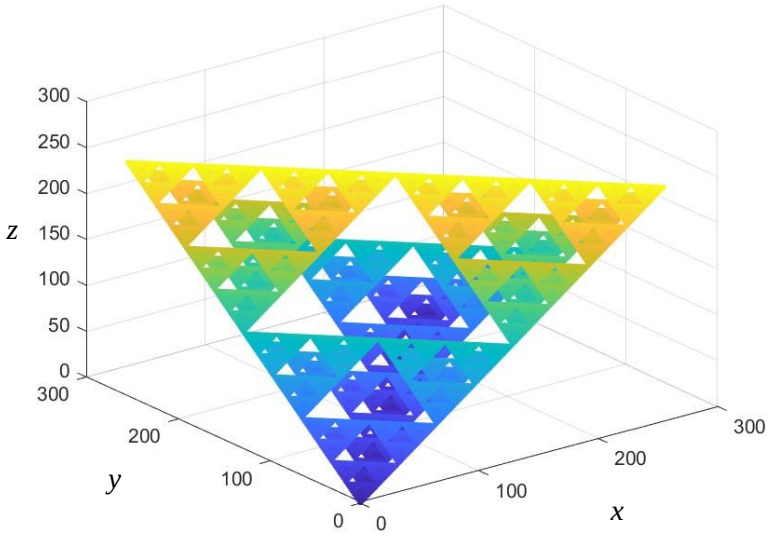
Algoritmas kartojamas pasirinkus kelis skirtingus mažesnės sekos ilgius N. Įvykdžius algoritmą turimos trys $K – N$ ilgio dešimtainių skaičių sekos (taškų koordinatės), kurias galima atidėti dvimatėje ir trimatėje plokštumose.



22 pav. Matricų vizualizavimo metodu pavaizduota trečios žmogaus chromosomos seka, kai $N = 8$. Vaizdo ašyse nurodytas matricos dydis, o spalvų skalė vaizduoja fragmentų dažnį. Raudonas tinklelis skirsto vaizdą į trumpesnius sekos fragmentus.

Dvimatėje plokštumoje seką galima pavaizduoti trimis skirtingais vaizdais. Pavaizduotos sekos pavyzdžiai pateikiami 21 paveiksle. Nors pavyzdyje pateiktos sekos vizualizacijos buvo gautos kai $N = 8$, gautus vaizdus galima nagrinėti ir trumpesnių sekų fragmentų atžvilgiu. Pavyzdžiui, 22 paveiksle pateiktas vaizdas gautas naudojant bazių grupių ir heterociklinių bazių koordinates. Iš pateikto pavyzdžio galima nuspręsti, jog sekoje yra mažiau CG nukleotidų kombinacijų. Pateiktą tinklelį galima pakeisti smulkesniu – tokiu, kuris nusakytų trijų fragmentų ar dar ilgesnes sekas. Trijų fragmentų tinklelio atveju būtų galima pastebėti mažesnį ACG, GCG, CCG ir TCG fragmentų skaičių – iš gaunamų dėsningumų galima spręsti, jog mažiausiai sekoje yra guanino nukleotidų.

Turimus koordinačių vektorius galima pavaizduoti ir trimatėje erdvėje kiekvieną koordinačių vektorių atidedant *x*, *y* ir *z* ašyse – gauto rezultato pavyzdys pateikiamas 23 paveiksle.



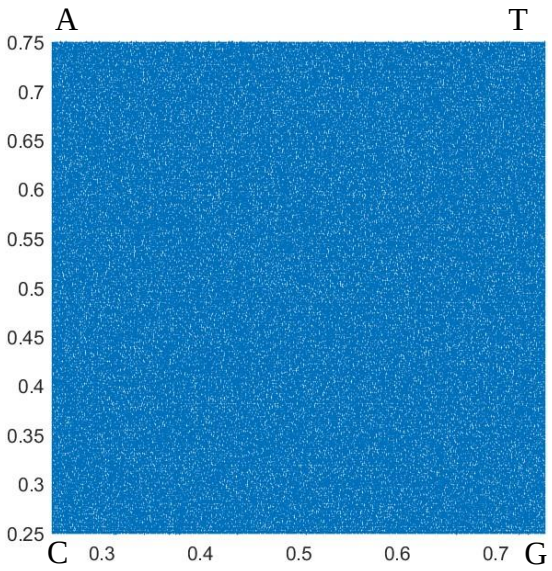
23 pav. Matricų vizualizavimo metodu pavaizduota trečios žmogaus chromosomos seka trimatėje erdvėje, kai $N = 8$. *X* ašyje atidėtos heterociklinių bazių koordinatės, *y* ašyje - ryšių stiprumo koordinatės, *z* ašyje – bazių grupių koordinatės.

2.2.2. Chaoso žaidimo vizualizavimo metodas

Chaoso žaidimo algoritmas sekos nedalina į mažesnius fragmentus. Kiekvienas nukleotidas vaizde pažymimas tašku nurodyta tvarka:

1. Pasirenkamas pradžios taškas (0;0).
2. Dvimatėje plokštumoje lygiais atstumais parenkami kvadratinės plokštumos kampai – koordinatės, atstovaujančios kiekvieną nukleotidą (A, T, G ir C). Pasirinktos kampų koordinatės yra tokios: A - (0,25; 0,75), C - (0,25; 0,25), G - (0,75; 0,25), T - (0,75; 0,75).
3. Iš pradinės *K* ilgio sekos pradžios pasirenkamas pirmas nukleotidas.
4. Skaičiuojamos nukleotido taško koordinatės plokštumoje – pradinio taško (0;0) atžvilgiu skaičiuojama tarpinė koordinatė, kuri atitinka pusę atstumo iki atitinkamo kampo plokštumoje. Gautas taškas laikomas pradžios tašku, ir toliau skaičiuojama kita tarpinė koordinatė antrojo sekos nukleotido atžvilgiu – ji taip pat atitinka pusę atstumo iki atitinkamo kampo plokštumoje. Taip suskaičiuojamos visos nukleotidus atitinkančių taškų koordinatės.

Įvykdžius algoritmą turimas vienas *K* ilgio koordinačių dvimatėje plokštumoje rinkinys. Dvimatėje plokštumoje išdėliotus kampus galima išdėlioti ir trimatėje erdvėje – tuomet įvykdžius algoritmą bus gaunamas vienas *K* ilgio koordinačių trimatėje erdvėje rinkinys.



24 pav. Chaoso žaidimo vizualizavimo metodu pavaizduota trečios žmogaus chromosomos seka dvimatėje plokštumoje.

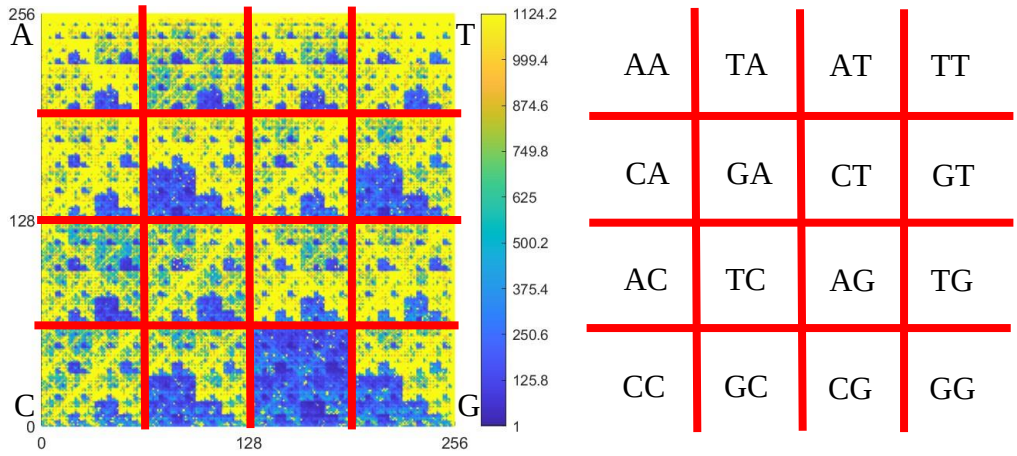
Kadangi gaunami koordinačių taškai yra unikalūs ir nepasikartoja, atliekamas paprastas taškų vaizdavimas – papildomai spalvomis dažniai neišskiriami. Chaoso žaidimo algoritmu gauto rezultato pavyzdys pateikiamas 24 paveiksle.

2.2.3. Chaoso žaidimo dažnių vizualizavimo metodas

Chaoso žaidimo dažnių vizualizavimo metodas naudojamas turint chaoso žaidimo metodu gautą rezultatą. Chaoso žaidimo dažnių vizualizavimo metodą galima nusakyti žemiau aprašytais žingsniais:

1. Pasirenkamas mažesnės sekos ilgis *N*.
2. Chaoso žaidimo vizualizavimo metodu gautas vaizdas suskirstomas į $2^N \times 2^N$ dydžio matricą.
3. Kiekviename gautos matricos langelyje surandamas chaoso žaidimo vizualizavimo metodu gautų taškų skaičius, patenkantis į atitinkamą langelį.
4. 2-3 žingsniai yra kartojasi tol, kol suskaičiuojami visi taškų dažniai, patenkantys į atitinkamus $2^N \times 2^N$ matricos langelius.

Priešingai nuo matricos vizualizavimo metodo, šiuo metodu yra gaunamas tik vienas vaizdas, tačiau jį taip pat galima nagrinėti trumpesnių sekos fragmentų atžvilgiu. Chaoso žaidimo dažnių vizualizavimo metodu gauto vaizdo pavyzdys ir vaizdo skirstymas į trumpesnius fragmentus pateikiamas 25 paveiksle. Nors matricų vaizdavimo ir chaoso žaidimo dažnių vaizdavimo metodais gautų rezultatų vaizdai skiriasi, abejais metodais gaunama informacija yra identiška: rečiausiai pastebima nukleotidų kombinacija sekoje yra CG. Pasirinkus mažesnį tinklelį – t.y. nusakantį trijų fragmentų ar dar ilgesnes sekas, fragmentai, kurių sekoje yra mažiau, sutaptų su matricos vaizdavimo metodu gautais fragmentais.



25 pav. Chaoso žaidimo dažnių vizualizavimo metodu pavaizduota trečios žmogaus chromosomos seka dvimatėje plokštumoje, kai N= 8. Vaizdo ašyse nurodytas matricos dydis, o spalvų skalė vaizduoja fragmentų dažnį. Raudonas tinklelis skirsto vaizdą į trumpesnius sekos fragmentus.

2.2.4. Vaizdų skirtumą palyginantys parametrai

Tyrime gaunamų vaizdų skirtumui įvertinti buvo pasirinkti trys parametrai – struktūrinio panašumo indeksas (SSIM), vaizdų atimtis ir Pirsono koreliacijos koeficientas. Kiekvieno parametro reikšmės apskaičiuojamos lyginant skirtingais skirtingų sekų vaizdus su vienodomis dimensijomis.

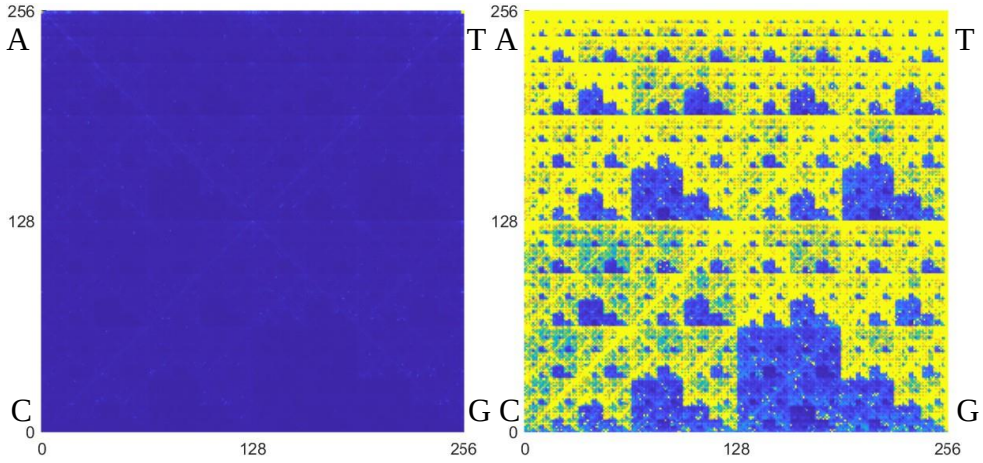
SSIM reikšmių intervalas yra [0; 1], kur 0 reiškia, jog vaizdai visiškai nepanašūs, o 1 rodo, jog vaizdai yra identiški. Kuo vaizdai yra panašesni, tuo struktūrinio panašumo indekso reikšmė yra artimesnė 1. Vaizdų atimties metodas naudojamas vizualiniam palyginimui. Metodo gaunamas rezultatas – matrica, kurios dimensijos atitinka atimtų vaizdų dimensijas, o reikšmės yra matricos elementų skirtumai. Pirsono koreliacijos koeficiento reikšmių intervalas yra [-1; +1] – kuo reikšmė artimesnė 1, tuo vaizdai yra panašesni. Pirsono koreliacijos koeficientą galima naudoti ir tuomet, kai vaizdų skirstiniai nėra normalieji.

Visi vaizdų skirtumą lyginantys parametrai yra naudojami pilkų atspalvių vaizdų skirtumams palyginti.

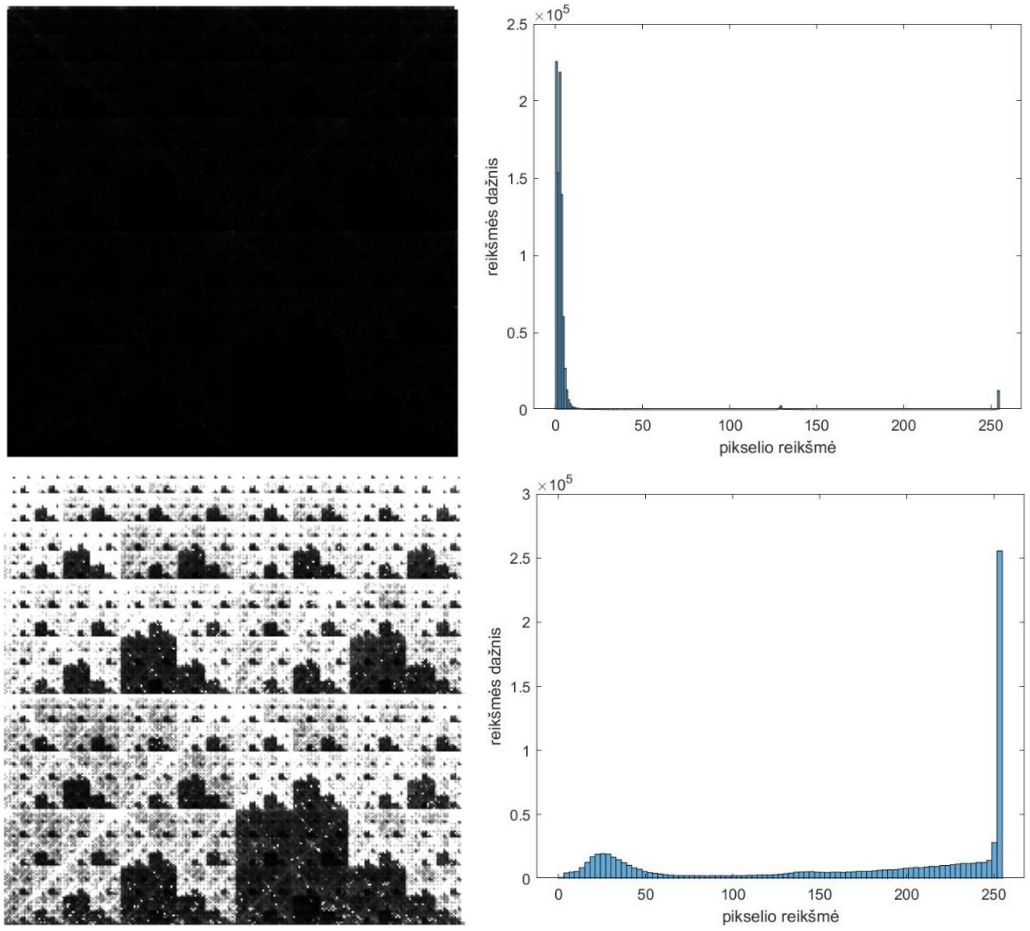
2.2.5. Vaizdų kontrasto didinimas

Pateiktuose pavyzdžiuose kiekvieno vaizdo kontrastas yra padidintas siekiant išryškinti vaizduose esantį raštą. Kai kuriuose kvadratuose (dažniausiai – vaizdo kampuose) taškų dažnis yra ypatingai didelis, tačiau tokių taškų yra labai mažai. Didžiausią vaizdo dalį sudaro mažą taškų dažnį rodantys kvadratai, todėl didžiausia dažnio reikšmė, kuriai yra priskirta geltona spalva, yra pakeičiama visų reikšmių vidurkiu. Vaizdų palyginimas pakoregavus vaizdo histogramos reikšmes ir jų nekoregavus pavaizduotas 26 paveiksle.

Vaizdų kontrasto koregavimas naudotas tik sekų vizualizacijoms pateikti. Skaičiuojant vaizdų skirtumą lyginančius parametrus gauti rezultatai vaizduojami ne RGB, o pilkumo skalėje, norint gauti teisingas pateiktų statistikų reikšmes. Skaičiuojant statistikas taip pat nebuvo koreguotos vaizdų histogramos. Rezultatai, pavaizduoti pilkumo skalėje, bei rezultatų reikšmių histogramos pateiktos 27 paveiksle.



26 pav. Trečios žmogaus chromosomos sekos vizualizacija, gauta chaoso žaidimo dažnių vaizdavimo metodu, prieš vaizdo kontrasto koregavimą (kairėje) ir po kontrasto koregavimo (dešinėje), kai $N = 8$.



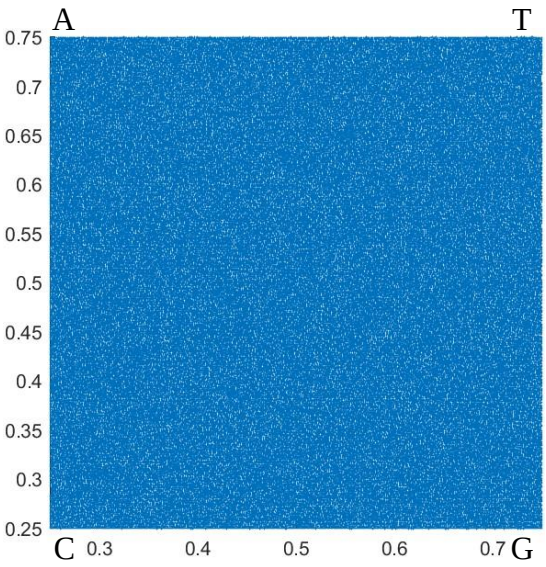
27 pav. Trečios žmogaus chromosomos sekos vizualizacijos, gautos chaoso žaidimo dažnių vaizdavimo metodu, kai $N = 8$. Kairėje viršuje matomas vaizdas prieš vaizdo kontrasto koregavimą, dešinėje – vaizdo pikselių reikšmių histograma. Kairėje apačioje matomas vaizdas po vaizdo kontrasto koregavimo, o dešinėje – vaizdo histograma.

3. Tyrimų rezultatai ir jų aptarimas

Skyriuje pateikiami matricų vizualizavimo ir chaoso žaidimo vizualizavimo metodu gauti rezultatai. Skyrius išskirtas į keturias dalis: chaoso žaidimo metodu gautų vaizdų analizė ir palyginimas, matricų vizualizavimo metodu gautų rezultatų analizė ir palyginimas, skirtingais metodais gautų vaizdų palyginimas bei optimalaus sekos fragmento ilgio matricų vizualizavimo metodui radimas remiantis vaizdų palyginimo parametrais.

3.1. Chaoso žaidimo vizualizavimo metodu gauti rezultatai

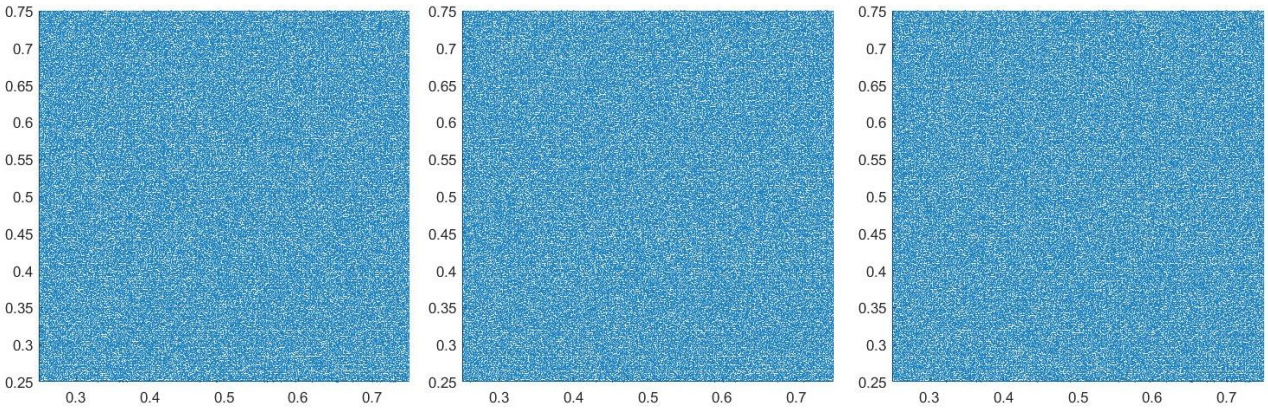
Chaoso žaidimo metodu vizualizuota 16 žmogaus chromosoma pavaizduota 28 paveiksle.



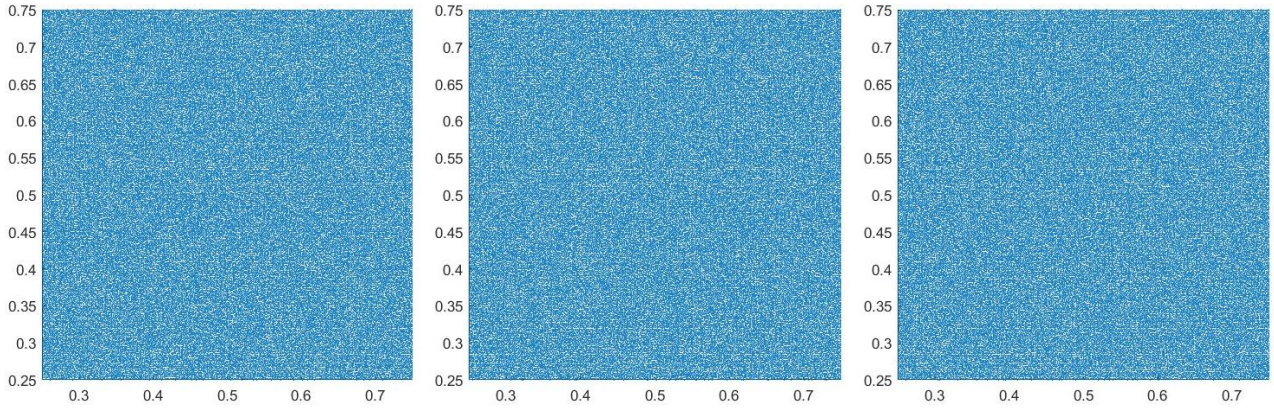
28 pav. Chaoso žaidimo vizualizavimo metodu pavaizduota 16-ta žmogaus chromosomos seka dvimatėje plokštumoje.

Pateiktame vaizde neįmanoma pastebėti jokių struktūros dėsningumų – taškai atrodo išsidėstę visiškai atsitiktinai.

Seka buvo išskirta į dvi dalis – koduojančią ir nekoduojančią genų dalį. Gauti rezultatai pavaizduoti 29 ir 30 paveiksluose.



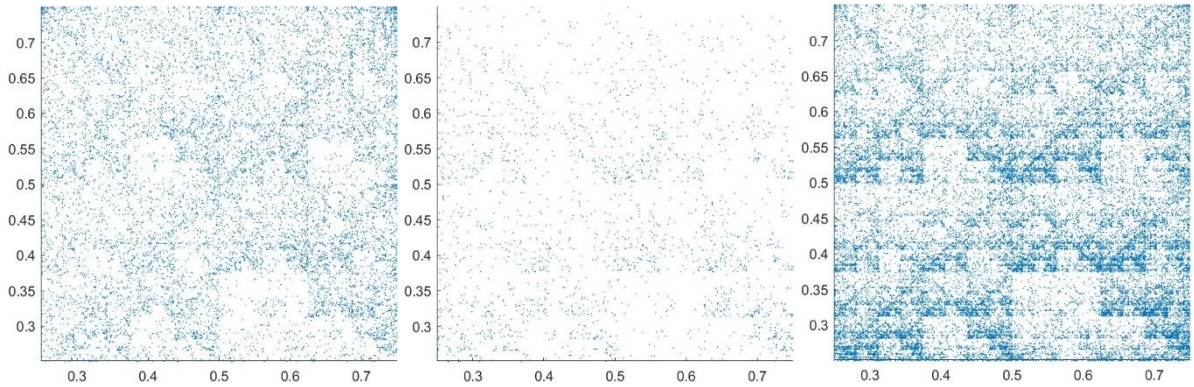
29 pav. Chaoso žaidimo vizualizavimo metodu pavaizduota 16-ta žmogaus chromosomos sekos geno nekoduojanti dalis dvimatėje plokštumoje. Trys vaizdai vaizduoja tris skirtingus sekos rėmelius.



30 pav. Chaoso žaidimo vizualizavimo metodu pavaizduota 16-ta žmogaus chromosomos sekos geną koduojanti dalis dvimatėje plokštumoje. Trys vaizdai vaizduoja tris skirtingus sekos rėmelius.

Galima pastebėti, jog ir išskirtose sekos dalyse nėra jokių taškų išsidėstymo dėsningumų. Todėl iš pateiktų sekos vizualizavimo pavyzdžių galima daryti išvadą, jog ilgoms sekoms pavaizduoti chaoso žaidimo metodas nėra tinkamas. Koreliacijos koeficientas tarp koduojančios ir nekoduojančios geno sekų dalių siekia 0,56 (visiems trims rėmeliams), o struktūrinio panašumo indeksas siekia vos 0,08 (taip pat visų trijų rėmelių atžvilgiu).

Tiesa, vaizduojant trumpesnę nukleotidų seką, chaoso žaidimo vizualizacijos metodu taip pat gaunamas fraktalines struktūras primenantis raštas. Trumpesnių sekų vizualizavimo pavyzdžiai pateikti 31 paveiksle.

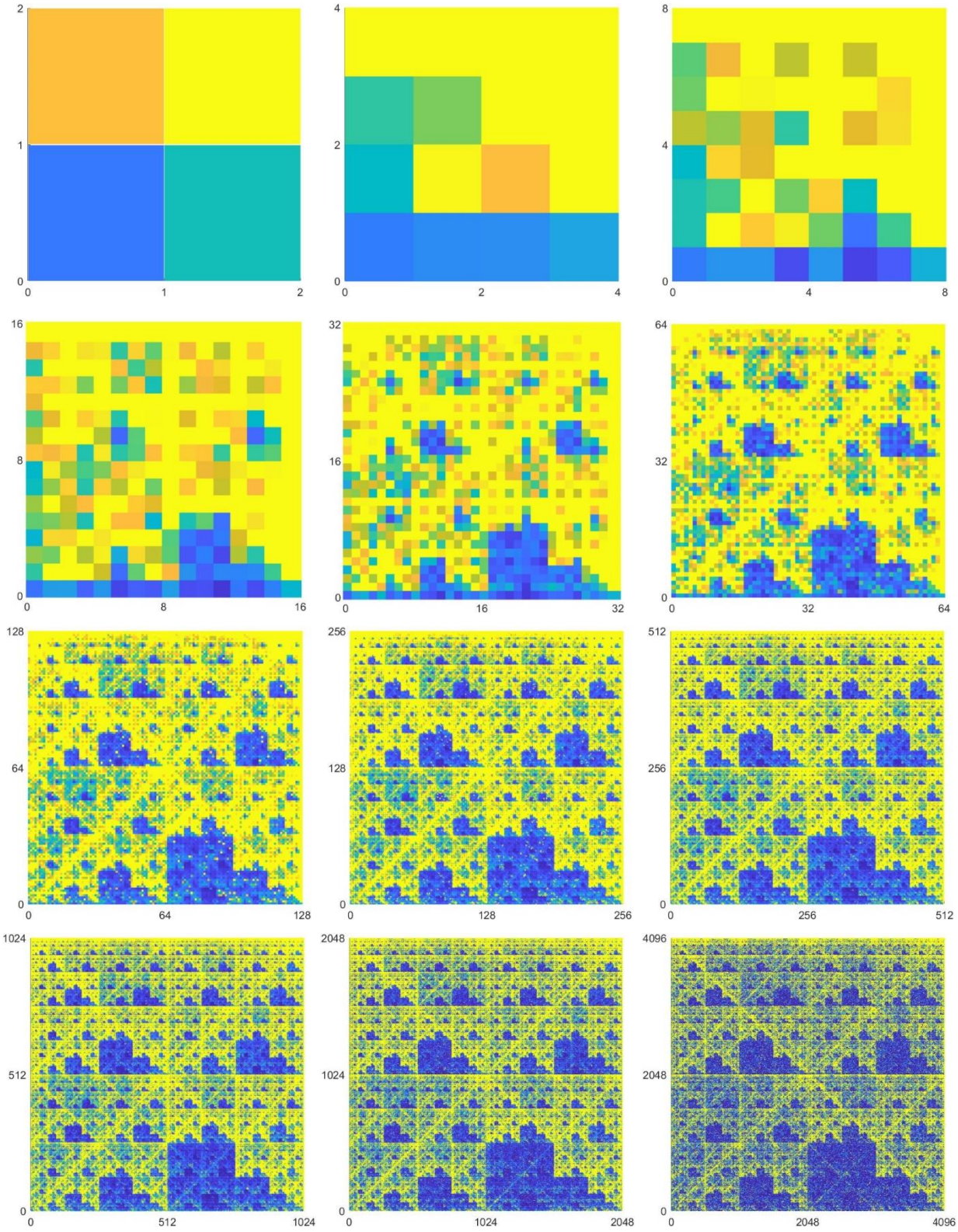


31 pav. Chaoso žaidimo vizualizavimo metodu pavaizduoti trys 16 – tos žmogaus chromosomos sekos genai dvimatėje plokštumoje. Kairėje vaizduojamas GINS2, viduryje – CTRL, dešinėje – BCAR1 genai.

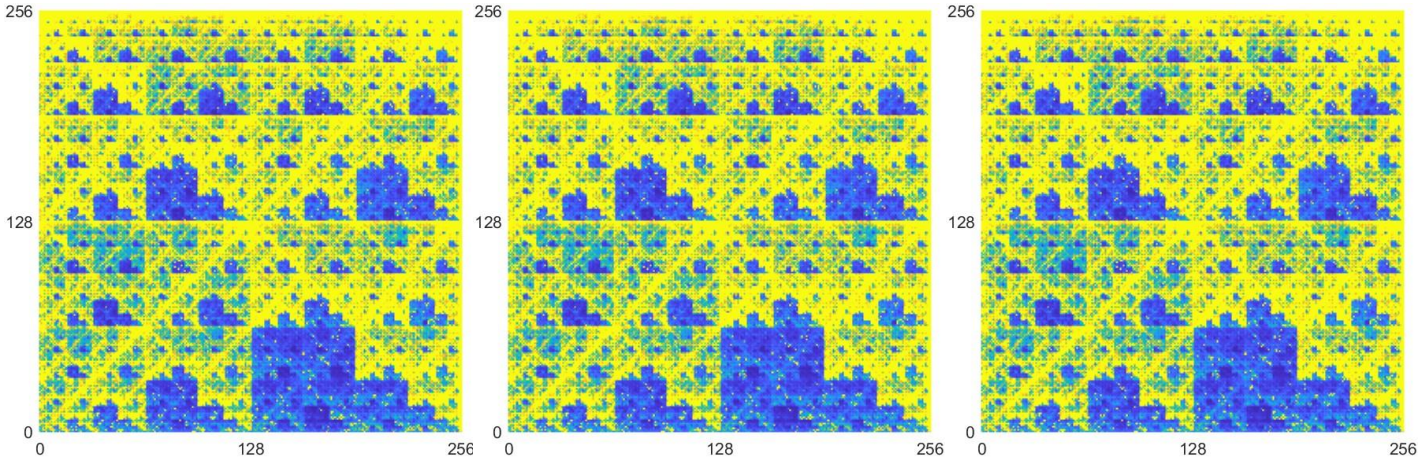
3.2. Chaoso žaidimo dažnių vizualizavimo metodu gauti rezultatai

Chaoso žaidimo dažnių vizualizavimo metodu gaunamų vaizdų rašto formavimasis kintant sekos fragmento ilgio parametrui pavaizduotas 32 paveiksle. Priešingai nei chaoso žaidimo vizualizavimo metodu gautuose vaizduose, 32 paveiksle pateiktuose vaizduose galima stebėti fraktalinių struktūrų formavimąsi. Tiesa, vis didėjant fragmento ilgiui matomas raštas po truputį nyksta, o fragmento ilgį pasirinkus didesnį, nei 12, raštą bus galima išskirti ne reikšmių dažnių kitimu, o atsirandančiu fragmentų trūkumu.

Svarbu paminėti, jog 32 paveiksle pateikiamų vaizdų histogramos yra išlygintos. Hostogramų neišlyginant vaizdai nebus kontrastingi, o raštas atsiras vizualizavimui pasirinkus ilgesnius fragmentų ilgius *N*.

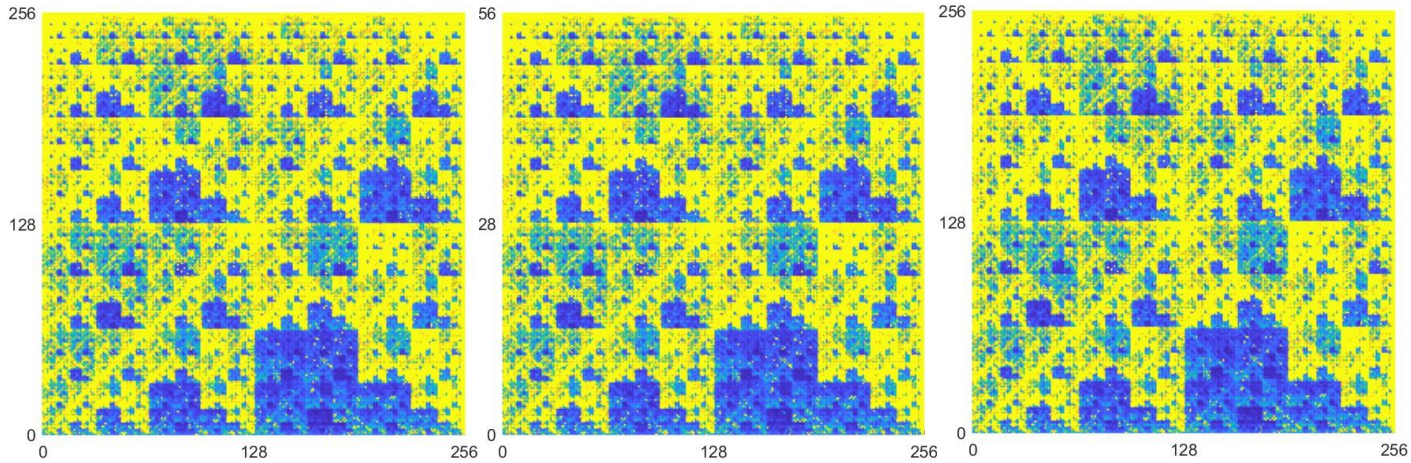


32 pav. Chaoso žaidimo dažnių vizualizavimo metodu pavaizduota 16-ta žmogaus chromosomos seka dvimatėje plokštumoje kintant sekos fragmento ilgiui. Vaizduose sekos fragmento ilgis *N* kinta nuo 1 iki 12.



33 pav. Chaoso žaidimo dažnių vizualizavimo metodu pavaizduota 16-ta žmogaus chromosomos sekos geno nekoduojanti dalis dvimatėje plokštumoje, kai $N = 8$. Trys vaizdai vaizduoja tris skirtingus sekos rėmelius.

Kaip ir chaoso žaidimo vizualizacijos metode, taip ir dažnių vizualizavimo metode seka buvo išskirta į dvi dalis: dalį, kuri koduoja geną, ir dalį, kuri geno nekoduoja. Vaizdų pavyzdžiai pateikiami 33 ir 34 paveiksluose. Galima pastebėti, jog visi išskirti sekų variantai vizualiai primena visos sekos vizualizacijose gaunamą raštą – tačiau kai kuriose vaizdų srityse matomi ryškesni neatitikimai ir su visos chromosomos vizualizacija, ir tarp koduojančių ir nekoduojančių sekos dalių.

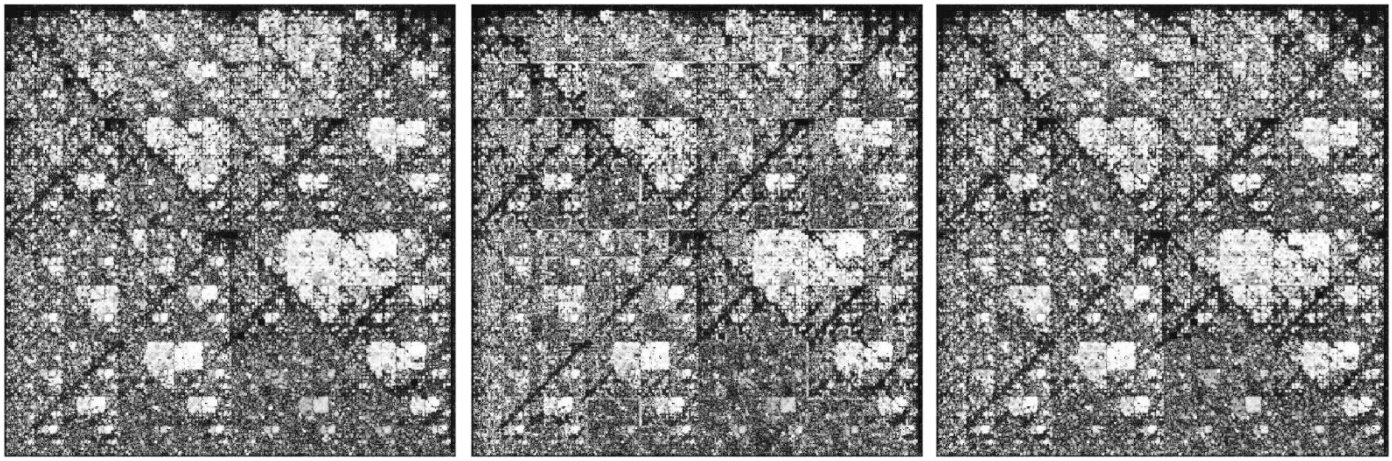


34 pav. Chaoso žaidimo dažnių vizualizavimo metodu pavaizduota 16-ta žmogaus chromosomos sekos geną koduojanti dalis dvimatėje plokštumoje, kai $N = 8$. Trys vaizdai vaizduoja tris skirtingus sekos rėmelius.

Skirtumas tarp koduojančios bei nekoduojančios dalių pateikiamas 35 paveiksle. Didžiausias skirtumas tarp vaizdų pažymėtas balta spalva. Galima pastebėti, jog visuose gautuose skirtumų raštuose taip pat matomas pasikartojantis raštas, kuris skiriasi nuo ankstesnių vaizdų raštų. Koreliacijos koeficientas tarp koduojančios ir nekoduojančios geno sekų dalių siekia 0,92 (pirmo rėmelio atveju), 0,9 (antro ir trečio rėmelių atveju), o struktūrinio panašumo indeksas siekia 0,76 (pirmo rėmelio atveju), 0,68 (antro rėmelio atveju) ir 0,72 (trečio rėmelio atveju). Ir statistiškai, ir vizualiai chaoso žaidimų dažnių vizualizavimo metodu gauti rezultatai yra tarpusavyje panašesni nei chaoso žaidimo metodu gauti vaizdai.

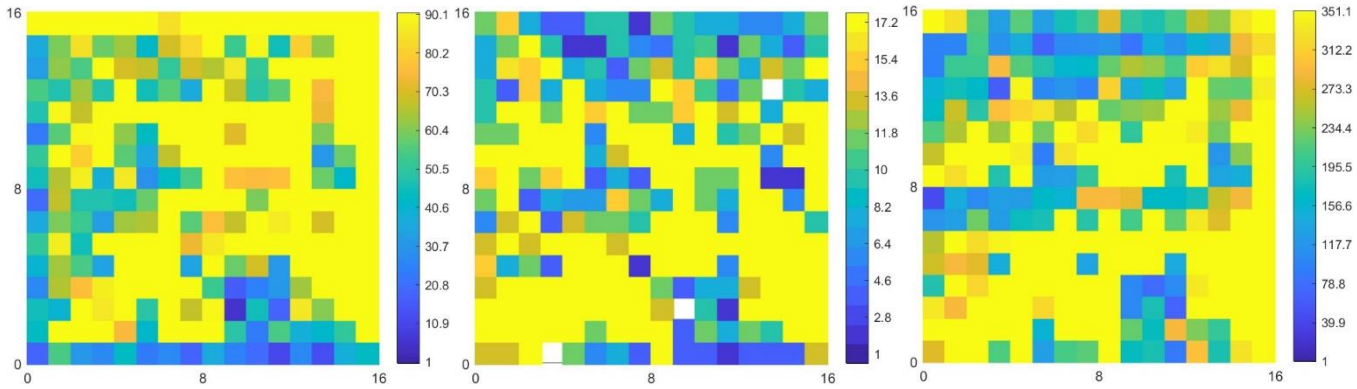
Chaoso žaidimo dažnių vaizdavimo metodu pavaizduoti ir genai, kurių sekų vizualizacijos buvo pateiktos naudojant chaoso žaidimo vaizdavimo metodą (31 paveiksle). Dėl trumpo sekos ilgio

vizualizavimui pasirinkus ilgesnį fragmentą (pvz. N = 8 ir daugiau) gauti rezultatai bus beveik identiški gautiems chaoso žaidimo vizualizacijos metodu – dėl šios priežasties sekos perskaičiuotos su mažesniu N fragmento ilgiu (N = 4). Gauti rezultatai pateikiami 36 paveiksle.



35 pav. Chaoso žaidimo dažnių vizualizavimo metodu pavaizduoti 16-tos žmogaus chromosomos sekos geną koduojančios ir nekoduojančios dalių skirtumai dvimatėje plokštumoje, kai N = 8. Trys vaizdai vaizduoja trijų skirtingų sekos rėmelių skirtumus.

Nors 36 paveiksle pavaizduotų rezultatų spalvų pasiskirstymas yra panašus, vaizdų pikselių reikšmių dažniai, pavaizduoti šalia kiekvieno vaizdo, smarkiai skiriasi. Trumpiausio CTRL geno vizualizacijoje matomi keli balti kvadratai – tai trūkstami sekos fragmentai GGCC, GCAG ir GCGT.

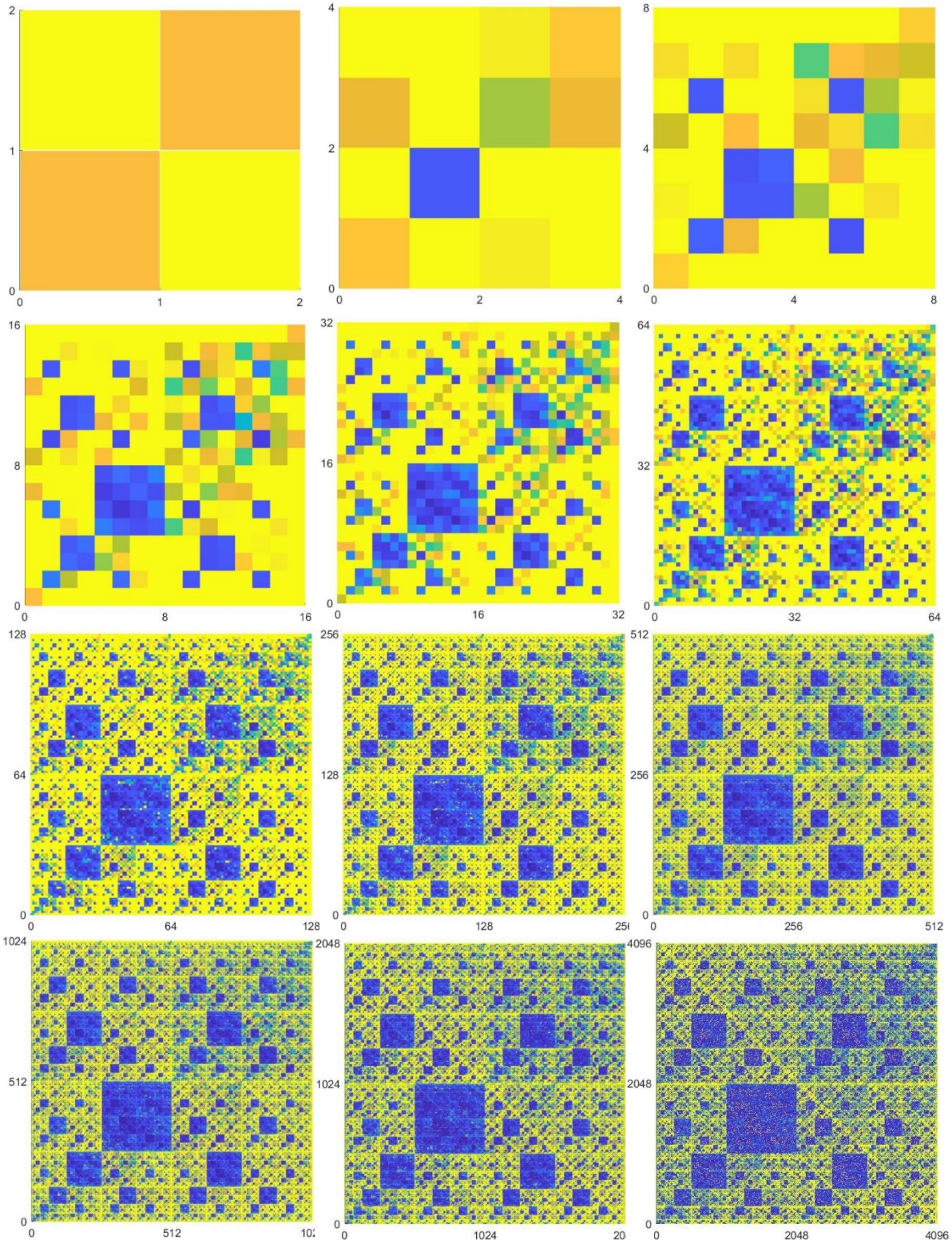


36 pav. Chaoso žaidimo dažnių vizualizavimo metodu pavaizduoti trys 16 – tos žmogaus chromosomos sekos genai dvimatėje plokštumoje, kai N = 4. Kairėje vaizduojamas GINS2, viduryje – CTRL, dešinėje – BCAR1 genai.

3.3. Matricų vizualizavimo metodu gauti rezultatai

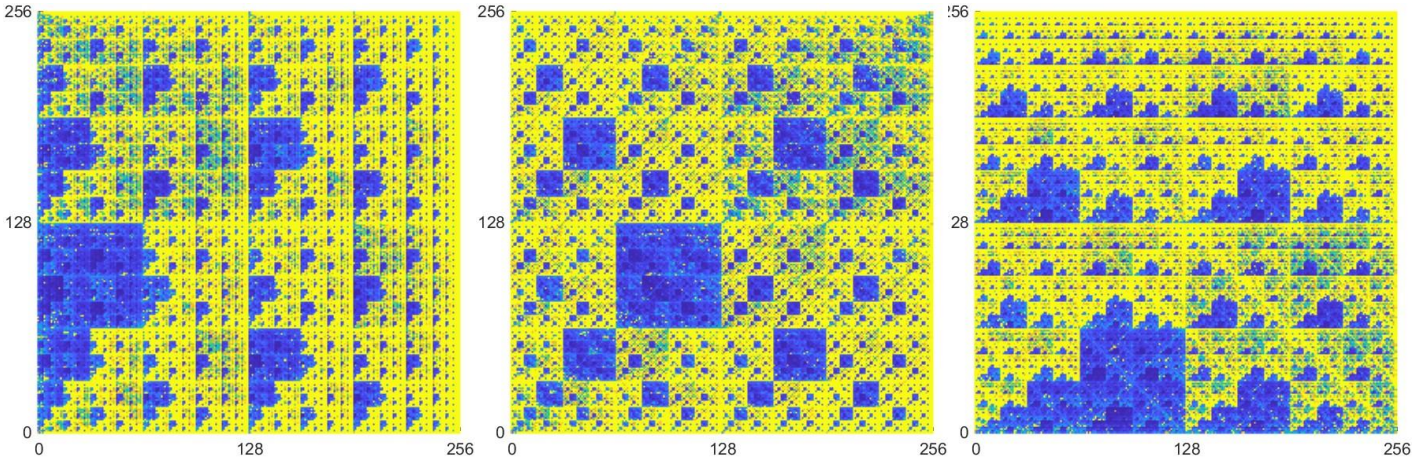
Matricų vizualizavimo metodu gaunamų vaizdų rašto formavimasis kintant sekos fragmento ilgio parametrui pavaizduotas 37 paveiksle. Vaizdai gauti vaizdavimui pasirinkus bazių grupes ir heterociklines bazes nusakančias koordinates. Visų galimų skirtingų raštų pavyzdžiai pateiktas 38 paveiksle. Visų raštų evoliucijos pateikiamos 1 ir 2 priede.

Galima pastebėti, jog visi pateikiami vaizdai taip pat turi pastebimas fraktalines struktūras – o 37 paveiksle pateiktas raštas skiriasi nuo iki šiol pateiktų fraktalinių struktūrų pavyzdžių. Ilgėjant pasirinkto sekos fragmento ilgiui, raštas susiformuoja, o vėliau pradeda tamsėti – pasikartojantys fragmentų dažniai retėja, o po kurio laiko visai išnyksta.

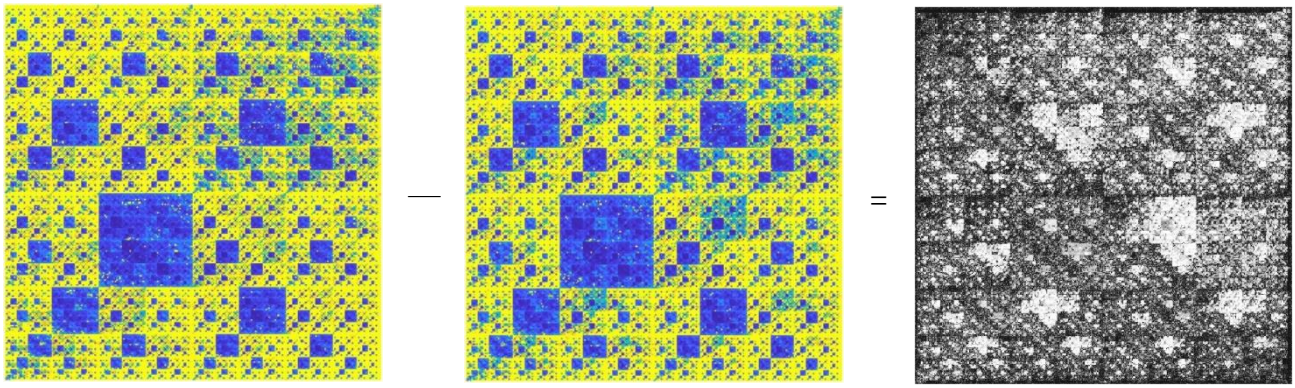


37 pav. Matricų vizualizavimo metodu pavaizduota pirma žmogaus chromosomos seka dvimatėje plokštumoje kintant sekos fragmento ilgiui. Vaizduose sekos fragmento ilgis kinta nuo 1 iki 12.

Sekos taip pat vaizduotos ir koduojančioms bei nekoduojančioms genų sekos dalims – tačiau rezultatai beveik nesiskyrė nuo viso sekos vaizdavimo rezultato ir ankstesniais metodais gautų vaizdų. Pirmo rėmelio koduojančios ir nekoduojančios geno sekos fragmentai ir jų skirtumas pateikti 39 paveiksle. Koreliacijos koeficientas tarp koduojančios ir nekoduojančios geno sekų dalių siekia 0,94 (pirmo rėmelio atveju) o struktūrinio panašumo indeksas siekia 0,86 (taip pat pirmo rėmelio atveju).

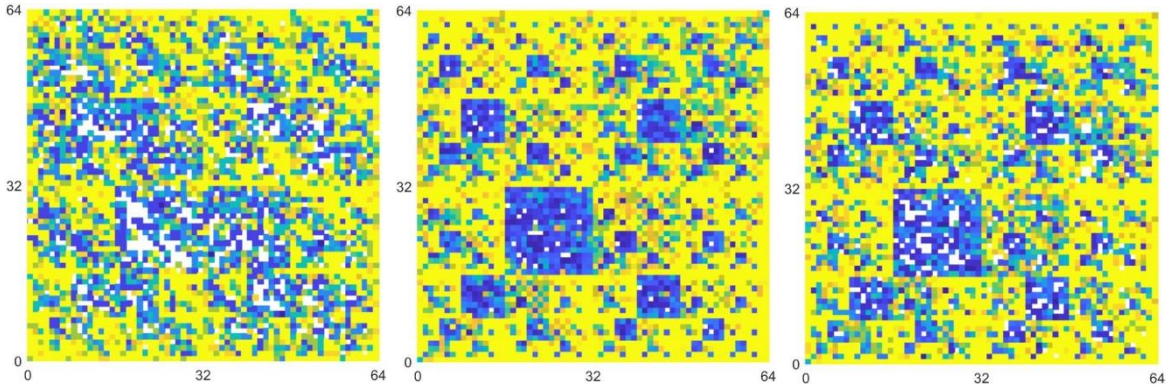


38 pav. Matricų vizualizavimo metodu pavaizduota pirma žmogaus chromosomos seka dvimatėje plokštumoje, kai N = 8. Pirmas vaizdas gautas naudojant ryšių stiprumo ir heterociklinių bazių koordinates, antras – naudojant bazių grupių ir heterociklinių bazių koordinates, trečias – naudojant bazių grupių ir ryšių stiprumo koordinates.



39 pav. Matricų vizualizavimo metodu pavaizduoti 1-os žmogaus chromosomos sekos geną koduojančios ir nekoduojančios dalių skirtumai dvimatėje plokštumoje, kai N = 8. Nespalvotas vaizdas vizualizuoja skirtumus tarp vaizdų (baltai žymimas didžiausias skirtumas).

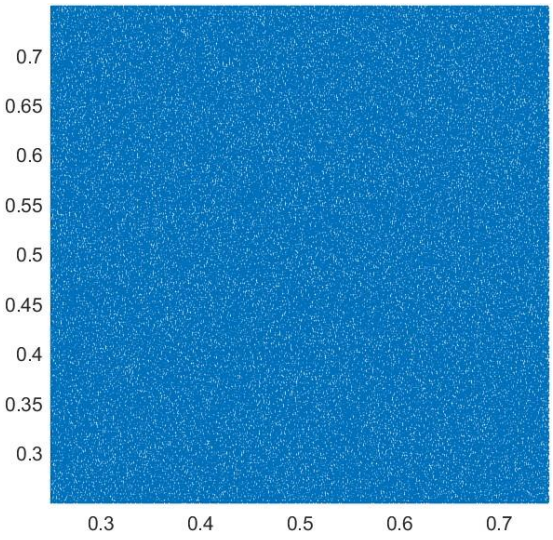
Matricų vaizdavimo metodu vizualizuoti ir trys pirmos žmogaus chromosomos genų sekos, kurių vaizdai pateikiami 40 paveiksle. Dėl trumpų sekos ilgių vizualizavimui pasirinktas trumpesnis fragmentas (N = 6), tačiau rezultatuose vistiek pastebimi trūkstami sekos fragmentai (žymimi balta spalva).



40 pav. Matricų vizualizavimo metodu pavaizduoti trys 1 – os žmogaus chromosomos sekos genai dvimatėje plokštumoje, kai N = 6. Kairėje vaizduojamas MTHFR, viduryje – MTOR, dešinėje – AGT genai.

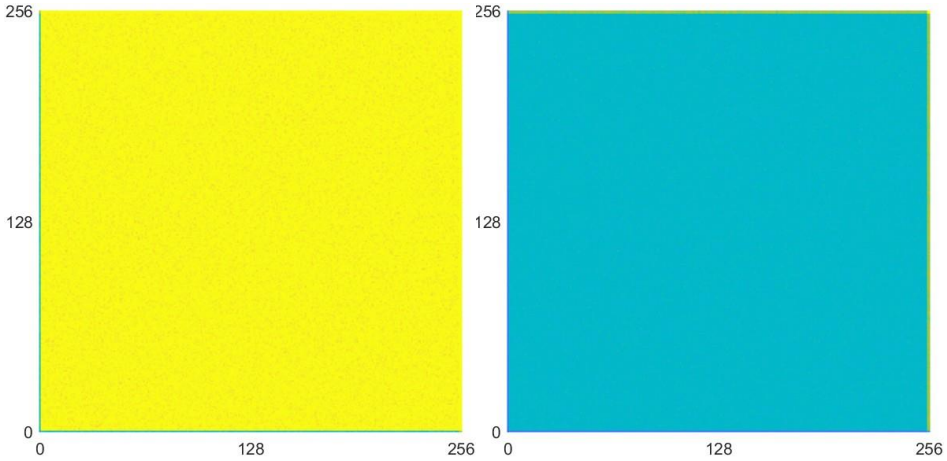
3.4. Atsitiktinės sekos generavimas

Atsitiktinai sugeneruotos sekos vaizdavimas chaoso žaidimo vizualizacijos metodu pateikiamas 41 paveiksle. Galima pastebėti, jog chaoso žaidimo metodu vizualizuota seka vizualiai nesiskiria nuo vizualizuotos DNR sekos.



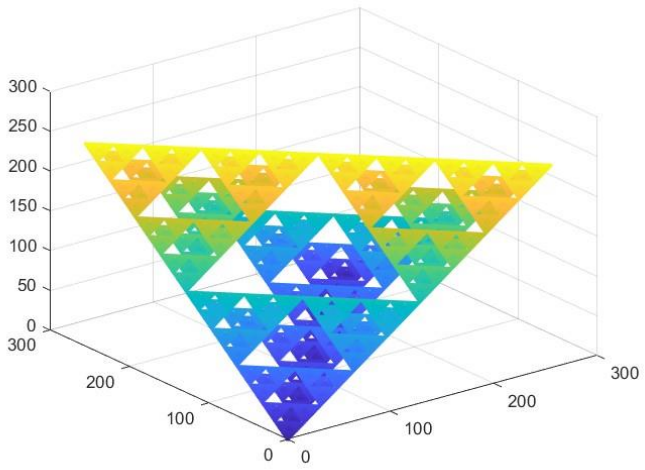
41 pav. Chaoso žaidimo vizualizavimo metodu pavaizduota atsitiktinė seka dvimatėje plokštumoje.

Atsitiktinę seką pavaizdavus chaoso žaidimo dažnių vizualizavimo metodu galima pastebėti akivaizdų skirtumą tarp atsitiktinai sugeneruotos sekos ir DNR sekos – pavaizdavus atsitiktinai sugeneruotą seką fraktalinės struktūros nėra pastebimos. Pavyzdys pateiktas 42 paveiksle.



42 pav. Chaoso žaidimo dažnių vizualizavimo metodu pavaizduota atsitiktinė seka dvimatėje plokštumoje. Kairėje pateikiamas vaizdas gautas išlyginus histogramą, dešinėje – neišlyginus.

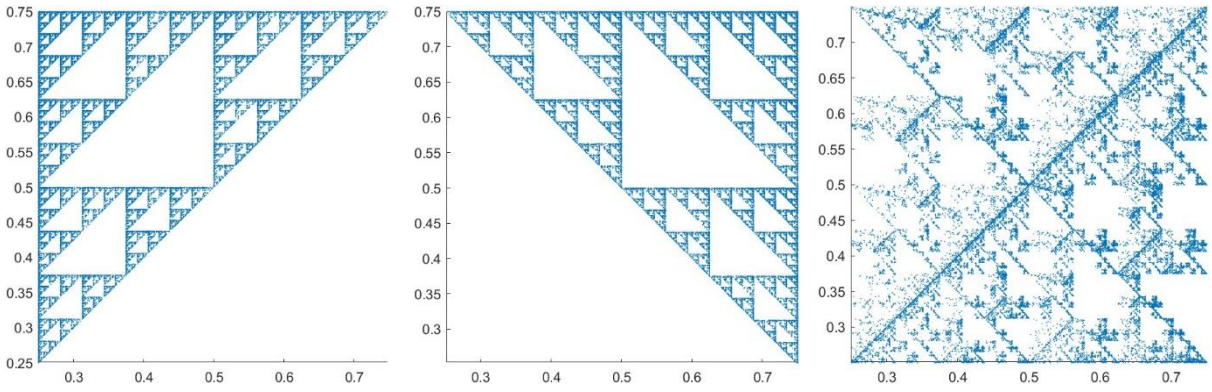
Matricų vaizdavimo metodu gaunamas vaizdas yra panašus, tačiau pavaizdavus atsitiktinę seką trimatėje erdvėje, gaunama fraktalinė struktūra – ji pavaizduota 43 paveiksle. Tai parodo, jog trimatėje erdvėje gauta fraktalinė struktūra susidaro dėl metodo specifikos, ne dėl sekos dėsningumo.



43 pav. Matricų vaizdavimo metodu gaunama atsitiktinės sekos vizualizacija

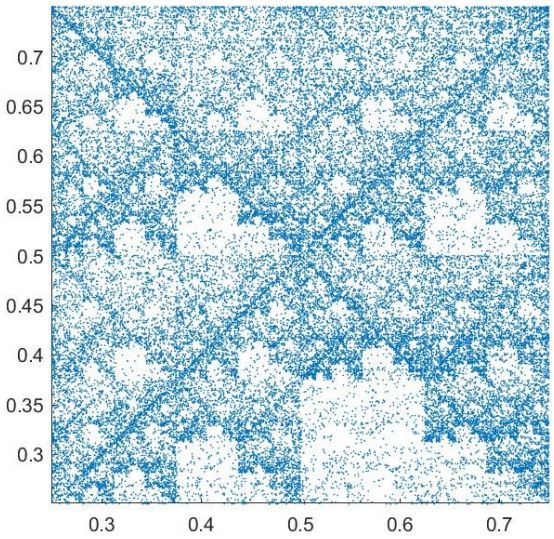
3.5. Logistine regresija sugeneruota seka

Logistinei regresijai pasirinkta nagrinėti vieną iš trumpesnių – VDR geną, esantį žmogaus 12 – oje chromosomoje. Gauti trys skirtingi sekų variantai - nagrinėjant centrinio nukleotido priklausomybę nuo tik iš kairės, tik iš dešinės, ir iš abiejų pusių esančių kaimynų. Gautų sekų vizualizacijos pateikiamos 44 paveiksle.



44 pav. Chaoso žaidimo vizualizacijos metodu pavaizduotos logistine regresija sugeneruotos sekos. Kairėje pavaizduota nukleodito priklausomybė iš kairės, viduryje – iš dešinės, o dešinėje – iš abiejų pusių.

Originali geno seka, naudota logistinės regresijos sekai sugeneruoti, pavaizduota 45 paveiksle. Galima pastebėti, jog visi trys sugeneruotų sekų variantai turi fraktalus primenančius raštus, tačiau nei vienas vaizdas neatitinka originalios sekos vizualizacijos. Nors visais trimis atvejais logistinės regresijos modeliais galima prognozuoti nukleotido bazę, žinant jo kaimyninius nukleotidus (Chi kvadrato reikšmės visais atvejais lygios $0 < 0{,}05$), sugeneruotų sekų vaizdai originalios sekos vizualizacijos neprimena. Tai parodo DNR sekos unikalumą.



45 pav. Chaoso žaidimo vizualizacijos metodu pavaizduotas VDR genas.

Išvados

1. Apžvelgus mokslinę literatūrą genetinės sekos vizualizavimui buvo pasirinkti matricų vizualizavimo, chaoso žaidimo vizualizavimo ir chaoso žaidimo dažnių vizualizavimo metodai. Vaizdų palyginimui buvo pasirinkti keli kriterijai: vaizdų atėmimo, struktūrinio panašumo indekso ir Pirsono koreliacijos koeficiento parametrai.
2. Naudojantis pasirinktais metodais vizualizuotos kelios skirtingos genetinės sekos, įskaitant tris žmogaus chromosomas, koduojančias ir nekoduojančias geno chromosomų dalis ir kelis skirtingus žmogaus genus. Sekos vizualizuotos dvimatėje ir trimatėje plokštumose. Pastebėta, jog matricų vizualizavimo metodu gaunamas raštas skiriasi nuo chaoso žaidimo ir chaoso žaidimo dažnių vizualizavimo metodais gaunamų raštų, o chaoso žaidimo vizualizavimo metodas yra tinkamas tik trumpoms sekoms pavaizduoti. Raštas, primenantis fraktalines struktūras, vaizduose pastebimas tuomet, kai vizualizuojamos sekos fragmento ilgis yra didesnis už 3.
3. Remiantis biologinėmis žiniomis žmogaus chromosomos atskirtos į koduojančias ir nekoduojančias genų sekas, o atskiri genai pasirinkti iš nukleotidų duomenų bazės. Pastebėta, jog trumpų sekų vizualizavimui tinkamas ir chaoso žaidimo vizualizavimo metodas – juo pavaizdavus seką pastebimas fraktalinis raštas. Visais atvejais pavaizduotos sekos nesiskyrė visos chromosomos vizualizacijos – raštas visuomet išliko panašus, tačiau keliuose vaizdo fragmentuose buvo galima pastebėti pokyčių (fragmentų sumažėjimo).
4. Palygintos žmogaus chromosomos, atsitiktinai sugeneruotos ir logistine regresija gautų sekų vizualizacijos vaizdų atėmimo, struktūrinio panašumo indekso ir Pirsono koreliacijos koeficiento parametrais. Gauti vaizdai pastebimai skyrėsi tarpusavyje, o pasirinkti parametrai tai tik patvirtino. Atsitiktinai sugeneruota seka dvimatėje plokštumoje fraktalinių struktūrų nesudarė, tačiau trimatėje erdvėje gautas rezultatas primena Sierpinskio trikampius – todėl padaryta išvada, jog trimatėje erdvėje fraktalas gaunamas dėl pasirinkto vaizdavimo metodo specifikos.
5. Nustatyta, kad dauguma atvejų įmanoma prognozuoti nukleotido bazę žinant jam iš šonų esančius kaimyninius nukleotidus, tačiau logistine regresija gautos sekos neatitinka realių DNR sekų vizualizacijų, nors sugeneruotoje sekoje taip pat susidarė unikalus raštas. Tai parodo, jog DNR sekų struktūra yra unikali.

Literatūros sąrašas

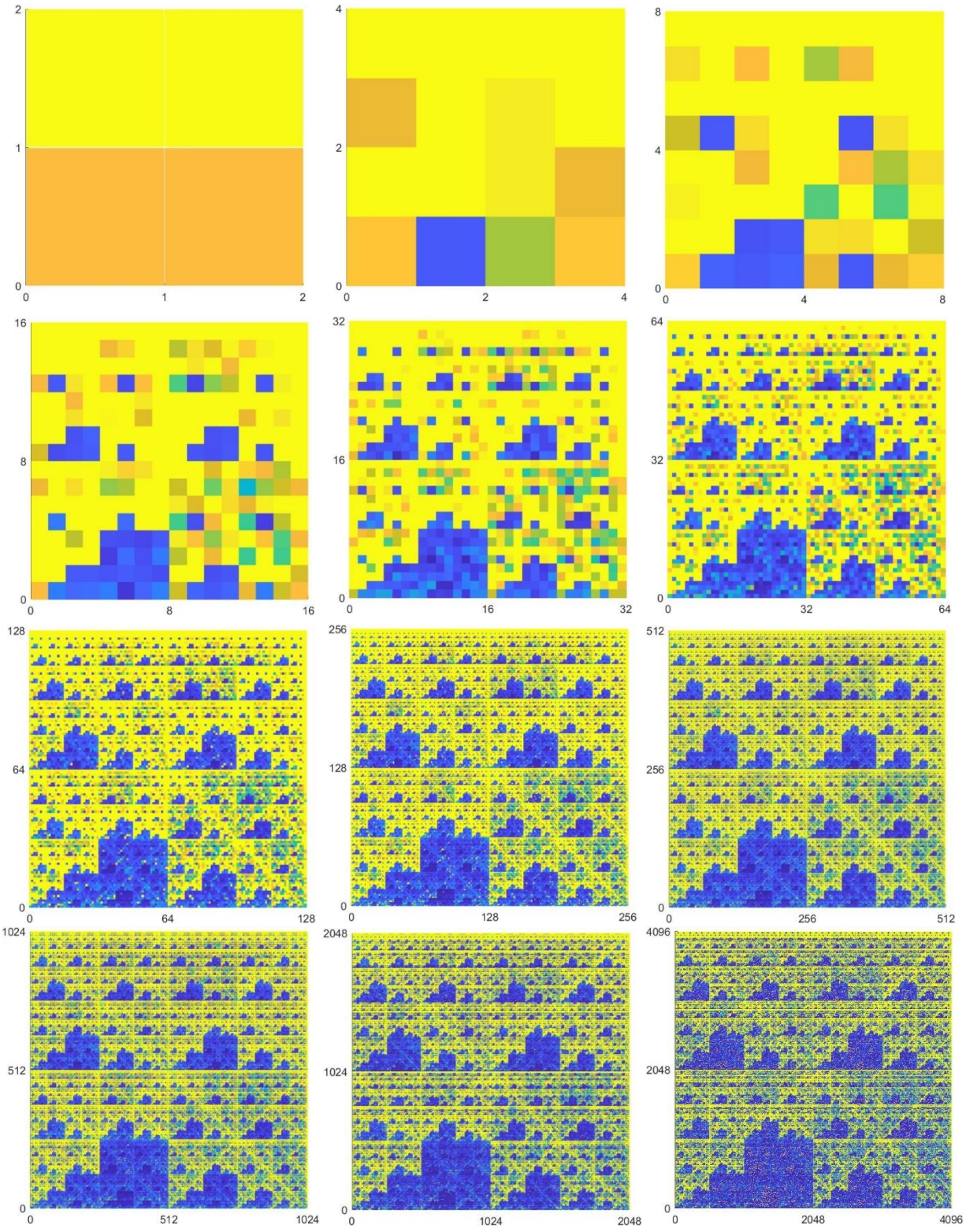
1. Molekulinė biologija. Vadovėlis. – Vilnius: Baltijos kopija, 2014. 412 p. Bibliogr. 397 p. http://www.esparama.lt/documents/10157/490675/Molekulin%C4%97s+biologijos+vadov%C4%97lis.pdf/118fdb87-cae7-49ac-ba59-c090d1dbb958
2. Povolotskaya, I.S., Kondrashov, F.A., Ledda, A. et al. Stop codons in bacteria are not selectively equivalent. Biol Direct 7, 30 (2012). https://doi.org/10.1186/1745-6150-7-30
3. Hershberg R. Mutation--The Engine of Evolution: Studying Mutation and Its Role in the Evolution of Bacteria. Cold Spring Harb Perspect Biol. 2015 Sep 1;7(9):a018077. doi: 10.1101/cshperspect.a018077. PMID: 26330518; PMCID: PMC4563715.
4. J. Židanavičiūte (2008) Logit analysis of genetic data, Mathematical Modelling and Analysis, 13:1, 135-144, DOI: 10.3846/1392-6292.2008.13.135-144
5. Avery, Peter J., and Daniel A. Henderson. "Fitting Markov Chain Models to Discrete State Series Such as DNA Sequences." Journal of the Royal Statistical Society. Series C (Applied Statistics) 48, no. 1 (1999): 53–61. http://www.jstor.org/stable/2680818.
6. Wu J, Liu Y and Zhao Y (2021) Systematic Review on Local Ancestor Inference From a Mathematical and Algorithmic Perspective. Front. Genet. 12:639877. doi: 10.3389/fgene.2021.639877
7. Khodaei A, Feizi-Derakhshi MR, Mozaffari-Tazehkand B. A Markov chain-based feature extraction method for classification and identification of cancerous DNA sequences. Bioimpacts. 2021;11(2):87-99. doi: 10.34172/bi.2021.16. Epub 2020 Mar 24. PMID: 33842279; PMCID: PMC8022238
8. Ching, W. K., Fung, E. S., & Ng, M. K. (2003). Higher-Order Hidden Markov Models with Applications to DNA Sequences. Lecture Notes in Computer Science, 535–539. doi:10.1007/978-3-540-45080-1_73
9. Cappelletti, L.; Fontana, T.; Di Donato, G.W.; Di Tucci, L.; Casiraghi, E.; Valentini, G. Complex Data Imputation by Auto-Encoders and Convolutional Neural Networks—A Case Study on Genome Gap-Filling. Computers 2020, 9, 37. https://doi.org/10.3390/computers9020037
10. E. Chen, J. Chu, J. Zhang, R. L. Warren and I. Birol, "GapPredict – A Language Model for Resolving Gaps in Draft Genome Assemblies," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 6, pp. 2802-2808, 1 Nov.-Dec. 2021, doi: 10.1109/TCBB.2021.3109557.
11. Zhang Y, Jia C, Fullwood MJ, Kwoh CK. DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. Brief Bioinform. 2021 Mar 22;22(2):2073-2084. doi: 10.1093/bib/bbaa039. PMID: 32227075.
12. Zheng, X., Xu, S., Zhang, Y. et al. Nucleotide-level Convolutional Neural Networks for Pre-miRNA Classification. Sci Rep 9, 628 (2019). https://doi.org/10.1038/s41598-018-36946-4
13. Dilliott, A.A., Abdelhady, A., Sunderland, K.M. et al. Contribution of rare variant associations to neurodegenerative disease presentation. npj Genom. Med. 6, 80 (2021). https://doi.org/10.1038/s41525-021-00243-3
14. Bertram L, Tanzi RE. The genetic epidemiology of neurodegenerative disease. J Clin Invest. 2005 Jun;115(6):1449-57. doi: 10.1172/JCI24761. PMID: 15931380; PMCID: PMC1137006.

15. D. M. Montserrat, C. Bustamante and A. Ioannidis, "Lai-Net: Local-Ancestry Inference with Neural Networks," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 1314-1318, doi: 10.1109/ICASSP40776.2020.9053662.
16. A. Samir et al., "A new approach for detection Alzheimer's Disease with machine learning using Whole Genomic and Single Nucleotide Polymorphism-Chip data," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), NV, USA, 2021, pp. 1535-1540, doi: 10.1109/CCWC51732.2021.9376121.
17. Takwa Mohamed, Sabah Sayed, Akram Salah, Essam H. Houssein, "Long Short-Term Memory Neural Networks for RNA Viruses Mutations Prediction", Mathematical Problems in Engineering, vol. 2021, Article ID 9980347, 9 pages, 2021. https://doi.org/10.1155/2021/9980347
18. Salama, M.A., Hassanien, A.E. & Mostafa, A. The prediction of virus mutation using neural networks and rough set techniques. J Bioinform Sys Biology 2016, 10 (2016). https://doi.org/10.1186/s13637-016-0042-0
19. Sirakoulis, G. C., Karafyllidis, I., Mizas, C., Mardiris, V., Thanailakis, A., & Tsalides, P. (2003). A cellular automaton model for the study of DNA sequence evolution. Computers in Biology and Medicine, 33(5), 439–453. doi:10.1016/s0010-4825(03)00017-9
20. Huang, Q., Wang, X., Li, H., He, F., & Wu, X. (2012). Visualization of DNA Sequence Features Based on Cellular Automata. Lecture Notes in Electrical Engineering, 77–82. doi:10.1007/978-3-642-25778-0_12
21. Athens, Josie & José, Marco. (2005). Mathematical properties of DNA sequences from coding and noncoding regions. Revista Mexicana de Física. 51.
22. Hannah Franziska Löchel, Dominik Heider, Chaos game representation and its applications in bioinformatics, Computational and Structural Biotechnology Journal, Volume 19, 2021, Pages 6263-6271, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2021.11.008.
23. Pasechnik, Alexey & Mylläri, A. & Salakoski, Tapio. (2012). Dynamical Visualization of the DNA Sequence and Its Nucleotide Content.
24. Mylläri, A., Salakoski, T., & Pasechnik, A. (2005). On the visualization of the DNA sequence and its nucleotide content. ACM SIGSAM Bulletin, 39(4), 131. doi:10.1145/1140378.1140385
25. Berger, J. A., Mitra, S. K., Carli, M., & Neri, A. (2004). Visualization and analysis of DNA sequences using DNA walks. Journal of the Franklin Institute, 341(1-2), 37–53. doi:10.1016/j.jfranklin.2003.12.002
26. Anitas EM. Fractal Analysis of DNA Sequences Using Frequency Chaos Game Representation and Small-Angle Scattering. International Journal of Molecular Sciences. 2022; 23(3):1847. https://doi.org/10.3390/ijms23031847
27. Hannah Franziska Löchel, Dominik Heider, Chaos game representation and its applications in bioinformatics, Computational and Structural Biotechnology Journal, Volume 19, 2021, Pages 6263-6271, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2021.11.008.
28. Stepanyan, Ivan V., and Sergey V. Petoukhov. 2017. The Matrix Method of Representation, Analysis and Classification of Long Genetic Sequences, Information 8, no. 1: 12. https://doi.org/10.3390/info8010012
29. Stepanyan, I. V., & Khussein, A. M. (2019). Scaling and Visualization of Nucleotide Sequences. EPJ Web of Conferences, 224, 03007. doi:10.1051/epjconf/201922403007

30. Stepanyan, Ivan, and Michail Lednev. 2022. "Spectral Decomposition of Mappings of Molecular Genetic Information in the System Basis of Single Nucleotide Functions" Symmetry 14, no. 5: 844. https://doi.org/10.3390/sym14050844
31. Stepanyan, Ivan & Lednev, Michail. (2022). Parametric Multispectral Mappings and Comparative Genomics. Symmetry. 14. 2517. 10.3390/sym14122517.
32. D. Brunet, E. R. Vrscay and Z. Wang, "On the Mathematical Properties of the Structural Similarity Index," in IEEE Transactions on Image Processing, vol. 21, no. 4, pp. 1488-1499, April 2012, doi: 10.1109/TIP.2011.2173206.
33. Raman B. Paranjape, 1 - Fundamental Enhancement Techniques, In Biomedical Engineering, Handbook of Medical Imaging, Academic Press, 2000, Pages 3-18, https://doi.org/10.1016/B978-012077790-7/50004-7.
34. A. M. Neto, A. C. Victorino, I. Fantoni, D. E. Zampieri, J. V. Ferreira and D. A. Lima, "Image processing using Pearson's correlation coefficient: Applications on autonomous robotics," 2013 13th International Conference on Autonomous Robot Systems, Lisbon, Portugal, 2013, pp. 1-6, doi: 10.1109/Robotica.2013.6623521.
35. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb). 2013;23(2):143-9. doi: 10.11613/bm.2013.018. PMID: 23894860; PMCID: PMC3900058.

Priedai

1 priedas. Matricų vizualizavimo metodu pavaizduota 1 – a žmogaus chromosoma. Vaizdams gauti naudotos ryšių stiprumo ir heterociklinių bazių koordinatės.



2 priedas. Matricų vizualizavimo metodu pavaizduota 1 – a žmogaus chromosoma. Vaizdams gauti naudotos bazių grupių ir ryšių stiprumo koordinatės.