

(11) Patento numeris: **6839** (51) Int. Cl. (2021.01): **G06N 3/00**

(21) Paraiškos numeris: **2019 524**

(22) Paraiškos padavimo data: **2019-09-27**

(41) Paraiškos paskelbimo data: **2021-05-10**

(45) Patento paskelbimo data: **2021-08-10**

(62) Paraiškos, iš kurios dokumentas išskirtas, numeris: —

(86) Tarptautinės paraiškos numeris: —

(86) Tarptautinės paraiškos padavimo data: —

(85) Nacionalinio PCT lygio procedūros pradžios data: —

(30) Prioritetas: —

(72) Išradėjas:

Laurynas KARPUS, LT
Vykintas JAUNIŠKIS, LT
Donatas REPEČKA, LT
Rolandas MEŠKYS, LT

(73) Patento savininkas:

Vilniaus Universitetas, Universiteto g. 3, LT-01513 Vilnius, LT
UAB „Biomatter Designs”, Žirmūnų g. 139 A, 09120 Vilnius, LT

(74) Patentinis patikėtinis/atstovas:

**Jurga PETNIŪNAITĖ, AAA Law, A. Goštauto g. 40B, Verslo centras „Dvyniai”,
LT-03163 Vilnius, LT**

(54) Pavadinimas:

**Būdas skirtas generuoti funkcionalias baltymų sekas naudojant generatyvinius
adversarinius tinklus**

(57) Referatas:

Išradimas yra susijęs su baltymų sekų ir funkcinių baltymų sekų generavimu. Konkrečiau, išradimas susijęs su funkcinių baltymų sekų, turinčių generatyvinius adversarinius tinklus, generavimo būdu. Aprašytas funkcinės sekos generavimo būdas apima daugybę žingsnių, kurių kiekvienas yra būtinas siekiant užtikrinti aukštą funkcinių sekų procentą galutiniame sekų rinkinyje: daugybės esamų baltymų sekų parinkimas, norint apibrėžti apytikslę sekos erdvę vėliau sugeneruotoms sintetinėms sekoms, perdurbant pasirinktas baltymų sekas, apytiksliai suderinant nežinomą tikrąjį iš anksto apdorotų sekų aminorūgščių pasiskirstymą, naudojant generatyvinių priešingų tinklų variantus, gaunant baltymų sekas iš apytikslio paskirstymo, gautų baltymų sekų perdurbimo. Aprašytas būdas suteikia išteklių prasme (pvz., laiko, sąnaudų) efektyvų būdą sintetinių baltymų sekų, kurios turi didelę tikimybę būti funkcionalios eksperimentiškai, gamybai.

IŠRADIMO SRITIS

Išradimas yra susijęs su baltymų sekų ir funkcionalių baltymų sekų generavimo sritimi. Detaliau, išradimas yra susijęs su funkcionalių baltymų generavimu naudojant generatyvinius adversarinius tinklus.

TECHNIKOS LYGIS

Baltymai yra molekulės susidedančios iš aminorūgščių grandinių galinčių susilankstyti trijų dimensijų erdvėje ir suformuoti molekulinės mašinas įvairių cheminių reakcijų katalizei. Rekombinantiniai baltymai yra ypatingai naudingi ir dažnai taikomi medicinoje, pvz. kaip antikūnai, vakcinos ir augimo faktoriai. Taip pat, baltymai turintys katalizines savybes (fermentai) yra aktyviai naudojami įvairiose industrijose, pvz. biokuro, maisto ir cheminės sintezės. Naudojant 20 dažniausiai aptinkamų proteogeninių aminorūgščių, baltymas susidedantis iš 100 aminorūgščių gali susidaryti iš 20^{100} skirtingų sekų variantų. Dėl šios priežasties sisteminė baltymų variantų paieška yra itin sudėtinga. Šioje astronominėje sekų erdvėje, tik 1 iš 10^{77} galimų baltymų sekų susilanksto į tinkamas trijų dimensijų struktūras galinčias atlikti biologines funkcijas (Keefe and Szostak 2001; Taverna and Goldstein 2002; Axe 2004). Standartinės atsitiktinės mutagenozės naudojimas naviguoti šią baltymų sekų erdvę (Romero and Arnold 2009a) yra dažnai neefektyvus, nes baltymų evoliucinis funkcionalumas mažėja eksponentiškai su kiekviena atsitiktine mutacija (Bloom et al. 2005; Guo, Choe, and Loeb 2004a). Todėl, rasti norimą funkcionalų baltymo variantą yra itin sudėtinga dėl didelės nefunkcionalių ar prastai susilankstančių baltymų erdvės. Eksperimentinės tikrinimo technikos yra taip pat ribotos, ir gali patikrinti tik 10^{6-9} baltymų variantų. Taip pat, iki 70% pavienių aminorūgščių pakeitimu lemia baltymų aktyvumo mažėjimą ir net 50% sukelia baltymų funkcijos išnykimą (Romero and Arnold 2009b; Bloom et al. 2006; Guo, Choe, and Loeb 2004b; Rennell et al. 1991; Axe, Foster, and Fersht 1998; Shafikhani et al. 1997; Rockah-Shmuel, Tóth-Petróczy, and Tawfik 2015; Sarkisyan et al. 2016). Tuo tarpu natūraliai egzistuojančių baltymų rekombinaciją sugeneruoja funkcionalius baltymus turinčius didelį kiekį mutacijų (Voigt et al. 2002; Hansson et al. 1999; Crameri et al. 1998). Pavyzdžiui, β -laktamazė turinti 75 mutacijas, gauta iš rekombinantinės bibliotekos turi 1016 kartų didesnį šansą būti funkcionali nei turinti 75 atsitiktines mutacijas (Drummond et al. 2005). Tačiau, ši strategija yra stipriai ribojama natūralių homologinių molekulių kiekio.

Naujausi giliojo mokymosi metodai pademonstravo aukštą struktūrinės, evoliucinės ir biofizikinės informacijos sugavimo potencialą natūraliose sekose potencialą, taip leidžiant tinklui išmokti baltymų savybių taisykles ir numatyti baltymo funkciją (Alley et al., n.d.).

Mašininio mokymosi modeliai aprašantys kompleksiškus epistatinis sekų sąryšius gali numatyti baltymų variantų funkcinis aktyvumus mokantis vien iš egzistuojančių sekų

(Riesselman, Ingraham, and Marks 2018). Tačiau, nors kompiuteriniai metodai galintys naviguoti baltymų sekų erdvę parodė didelį potencialą (Romero, Krause, and Arnold 2013; Yang, Wu, and Arnold 2019), iki šiol jie buvo naudojami funkcijos ir/ar funkcionalumo numatymui iš sekų. Giliojo mokymosi generatyviniai algoritmai gebantys produkuoti baltymų sekas buvo testuojami naudojant autoregresinius neuroninius tinklus (WO2019097014). Tačiau, šie metodai neužtikrina taisyklingo sugeneruotų baltymų susilankstymo arba cheminio aktyvumo, tad šis *in silico* metodas yra toks pat neefektyvus kaip dabar naudojami atsitiktiniai eksperimentiniai metodai.

Dėl šių priežasčių, yra didelis naujo metodo, skirto eksperimentiškai aktyvių sekų generavimui, poreikis.

IŠRADIMO ESME

Išradimas yra susijęs su baltymų sekomis ir funkcionalių baltymų sekų generavimu. Detaliau, išradimas yra susijęs su funkcionalių baltymų generavimu naudojant generatyvinius adversarinius tinklus

Aprašytas funkcinės sekos generavimo metodas apima daugybę žingsnių, kurių kiekvienas yra būtinas siekiant užtikrinti aukštą funkcinų sekų procentą galutiniame sukurtame sekų rinkinyje: daugybės sekų pasirinkimas siekiant apsibrėžti vėliau generuojamų sekų erdvę 6 pav. 601, pasirinktų sekų apdorojimas 6 pav. 602, nežinomo tikrojo aminorūgščių skirstinio aproksimacija pasirinktom baltymų sekom naudojant generatyvinius adversarinius tinklus 6 pav. 603, sekų gavimas iš aproksimuoto skirstinio 6 pav. 604, gautų sekų apdirbimas 6 pav. 605.

Apibūdintas metodas suteikia efektyvų būdą produkuoti sintetines baltymų sekas kurios turi aukštą šansą būti funkcionalios eksperimentiškai.

TRUMPAS BRĖŽINIŲ APRAŠYMAS

Neribojantys išradimo pavyzdžiai bus apibūdinti kartu su brėžiniais, kurie yra schematiniai ir nėra nupiešti atsižvelgiant į skalę. Brėžiniuose, kiekvienas identiškas ar beveik identiškas elementas yra pavaizduotas vienu numeriu. Dėl aiškumo, ne kiekvienas komponentas yra pažymėtas kiekviename brėžinyje, ir ne kiekvienas išradimo komponentas yra vaizduojamas, nes asmuo su tinkamu išsilavinimu gali suprasti aprašymą ir be jų.

Brėžiniuose:

1 pav. iliustruoja schemą apibūdinančią aukšto lygmens GAN architektūrą;

2 pav. iliustruoja schemą apibūdinančią generatoriaus tinklo architektūrą;

3 pav. iliustruoja schemą apibūdinančią Resnet bloko architektūrą generatoriaus

tinkle;

4 pav. iliustruoja schemą apibūdinančią diskriminatoriaus architektūrą tinkle;

5 pav. iliustruoja schemą apibūdinančią Resnet bloko architektūrą diskriminatoriaus tinkle;

6 pav. iliustruoja schemą apibūdinančią pagrindinius išradimo žingsnius;

7 pav. iliustruoja schemą apibūdinančią bendrą tinklo architektūrą naudojamą 1 pavyzdyje;

8 pav. iliustruoja generuotų sekų identitetą lyginant su artimiausia natūralia seka;

9 pav. iliustruoja generatoriaus ir diskriminatoriaus netekties funkciją mokymosi metu. Generatoriaus ir diskriminatoriaus nuostoliai po pradinės fazės tampa gana stabilūs ir galiausiai pasiekia plokštumą;

10 pav. iliustruoja Šenono entropijas generuotoms ir mokymo sekoms. Žemos Šenono entropijos vertės atspindi labai konservatyvias ir funkciškai svarbias pozicijas, tuo tarpu aukšta entropija rodo didelę aminorūgščių įvairovę tam tikroje padėtyje;

11 pav. iliustruoja kaip GAN išmoksta evoliuciškai svarbias aminorūgščių pozicijas;

12 pav. iliustruoja GAN galimybę atkurti pozicinius aminorūgščių skirstinius parodytus, kaip Pirsono sukurtų ir natūralių sekų koreliacijos koeficientus, įvertintus pagal kelių sekų derinimą. Pozicijos su mažesniais koreliacijos koeficientais atitinka pozicijas su didesniu sekos kintamumu. Pateikiamos tik pozicijos, kurių tarpų skaičius mažesnis nei 75%;

13 pav. iliustruoja aminorūgščių pozicijų asociacijų matricas natūralioms ir generuotoms baltymų sekoms. Teigiamos vertės rodo didesnę atstumą, lyginant su atsitiktinėmis sekomis su tuo pačiu aminorūgščių dažniu, t.y. sveikas skaičius rodo, kiek vidutiniškai aminorūgštys yra toliau viena nuo kitos nei atsitiktinėje sekoje;

14 pav. iliustruoja aminorūgščių porų koreliacijas tarp pagamintų sintetinių ir pasirinktų treniruočių sekų. Kiekvienas žemėlapių taškas parodo aminorūgščių porų tarp dviejų skirtingų duomenų rinkinių dažnių koreliaciją. Aukšta koreliacija reiškia, kad abiejuose duomenų rinkiniuose buvo rasta ta pati porinių aminorūgščių sąveika poromis;

15 pav. iliustruoja baltymų sekos erdvę, vizualizuotą transformuojant atstumo matricą, gautą naudojant k-tuple metodą baltymų sekų lygiavimui t-SNE įterpime. Taškų dydžiai parodo kiekvieno atstovo 70% tapatybės klasterio dydį;

16 pav. Iliustruoja CATH domenų diversiškumą gautą per visą „ProteinGAN“ evoliuciją. Kiekviename 1200 treniruočių etape buvo imamos 64 sekos ir ieškoma tipišku CATH domenų (E vertė $<1e-6$). Pradžia: „ProteinGAN“ sukurti nauji domenai, nerandami

natūraliose sekose, nes natūralių ir sugeneruotų sekų palyginimas su mutavusiomis atsitiktinėmis kontrolės sekomis parodė, kad sekų generavimas nebuvo atsitiktinis procesas (tiksliai Fišerio testo p vertė $< 8,2 \times 10^{-16}$);

17 pav. iliustruoja sekų įvairovės palyginimą tarp sugeneruotų ir natūralių MDH sekų. Sukurtos sekos yra sugrupuotos į įvairesnes grupes. Parodytas klasterių skaičiaus (Y ašies) santykis skirtingose sekos tapatumo ribose (X ašis);

18 pav. iliustruoja sintetinių MDH baltymų aktyvumo lygius bei natūralių MDH baltymų kontrolę;

19 pav. iliustruoja malato gamybos lygius sintetiniuose MDH baltymuose lyginant su natūraliais MDH baltymais.

DETALUS IŠRADIMO APRAŠYMAS

Toliau bus detalizuojamos išradimo sudėtinės dalys. Nors išradimas toliau bus apibūdinamas per pavyzdžius, turėtų būti suprasta jog šie pavyzdžiai neriboja išradimo apimtį. Kita vertus, šis išradimas apima ir alternatyvas, modifikacijas bei ekvivalentus, kurie gali būti naudojami išradimo apibrėžtyje.

Per visa išradimo atskleidimą, įvairūs išradimo aspektai gali būti pristatomi intervaliniu formatu. Turėtų būti suprasta, jog toks intervalinis apibūdinimas yra tik dėl aiškumo ir neturėtų būti suprastas kaip išradimo apimtį ribojantis faktorius.

Aprašytas funkcinės sekos generavimo metodas apima daugybę žingsnių, kurių kiekvienas yra būtinas siekiant užtikrinti aukštą funkcinių sekų procentą galutiniame sukurtame sekų rinkinyje: daugybės sekų pasirinkimas siekiant apibrėžti vėliau generuojamų sekų erdvę 6 pav. 601, pasirinktų sekų apdorojimas 6 pav. 602, nežinomo tikrojo aminorūgščių skirstinio aproksimacija pasirinktom baltymų sekomis naudojant generatyvinius adversarinius tinklus 6 pav. 603, sekų gavimas iš aproksimuoto skirstinio 6 pav. 604, gautų sekų apdirbimas 6 pav. 605.

Apibūdintas metodas suteikia efektyvų būdą produkuoti sintetines baltymų sekas kurios turi aukštą šansą būti funkcionalios eksperimentiškai.

Sąvokos:

Siekiant padėti suprasti išradimą, toliau pateikiami kai kurių sąvokų paaiškinimai.

Nebent nurodyta kitaip, visos techninės ir mokslinės sąvokos naudojamos toliau turi tokią pačią reikšmę kaip jas supranta žmonės turintys stiprų techninės srities supratimą. Nors bet kokie metodai gali būti naudojami apibrėžčių testavimui, aprašyme pateikti metodai yra pavyzdiniai.

Terminai „susideda iš“, „turi“ yra atviri terminai, nebent nurodyta kitaip.

Terminas „bio-molekulė“ arba „biomolekulė“ reiškia molekulę, kuri paprastai randama biologiniame organizme. Tokios biologinės molekulės gali būti biologinės makromolekulės, kurios paprastai yra polimerinės ir susideda iš kelių subvienetų (kitaip tariant „biopolimerai“). Tipinės biomolekulės gali būti RNR (susideda iš nukleotidų), DNR, polipeptidai (susideda iš aminorūgščių). Į tokias molekules, pavyzdžiui, įeina RNR, RNR analogai, DNR, DNR analogai, polipeptidai, polipeptidų analogai, RNR ir DNR kombinacijos. Biomolekulės gali taip pat gali, pavyzdžiui, būti lipidai, angliavandeniai, ar kitos organinės molekulės padarytos naudojant genetiškai užkoduojamas molekules (pvz. fermentai) ir panašiai.

Terminas „natūrali seka“ reiškia aminorūgščių sekas, kurios yra žinomos iš gamtos. Atitinkamai terminas „dirbtinė seka“ reiškia aminorūgščių sekas, kurios nėra žinomos gamtoje.

Terminas „sintetinė seka“ arba „sukurta seka“ reiškia baltymų sekas, sukurtas aprašytu išradimu.

Terminas „sekų erdvė“ reiškia erdvę, kurioje visus galimus baltymų kaimynus galima gauti iš eilės taškinių mutacijų.

Terminas „neuroninis tinklas“ arba „tinklas“ reiškia mašininio mokymosi modelį, kurį galima suderinti (pvz., apmokyti) remiantis įėjimais, kad būtų galima apytiksliai nustatyti nežinomus funkcijų sprendinius. Visų pirma, terminas neuroninis tinklas gali apimti tarpusavyje sujungtų neuronų, kurie bendrauja ir mokosi apytiksliai atlikti sudėtingas funkcijas bei generuoja išvestis, remiantis daugybe įvesties, pateiktos modeliui, modelį. Pavyzdžiui, terminas neuroninis tinklas apima vieną ar daugiau mašininio mokymosi algoritmų. Konkrečiai, terminas neuroninis tinklas gali apimti giliuosius konvoliucinius neuroninius tinklus, tokius kaip erdvinis transformatorių tinklas. Be to, neuroninis tinklas yra algoritmas (arba algoritmų rinkinys), įgyvendinantis gilias mokymosi metodus, kurie naudoja algoritmą aukšto lygio duomenų abstrakcijai modeliuoti.

Terminas „adversarinis mokymasis“ reiškia mašininio mokymosi algoritmą (pvz., generatyvinį adversarinį tinklą), kai kartu mokomasi priešingų mokymosi modelių. Visų pirma, terminas „konkuruojantis mokymasis“ apima daugybės mokymosi užduočių išsprendimą tame pačiame modelyje (pvz., iš eilės ar lygiagrečiai), atliekant užduotis atliekant vaidmenis ir apribojimus. Kai kuriuose įgyvendinimo variantuose konkurencinis mokymasis apima minimax funkcijos (pvz., minimax tikslo funkcijos) panaudojimą, kuris sumažina pirmojo tipo nuostolius ir padidina antrojo tipo nuostolius. Pvz., vaizdinėje kompozicinėje sistemoje naudojama adversariniu mokymusi, siekiant sumažinti nuostolius, susijusius su metmenų

parametrų generavimu naudojant geometrinį numatomąjį neuroninį tinklą, ir maksimaliai padidinti varžovų diskriminacijos neuroninio tinklo diskriminaciją su nerealiais vaizdais, kuriuos sukuria geometrinis prognozės neuroninis tinklas.

Sąvokos „polipeptidas“ ir „baltymas“ čia vartojamos pakaitomis, kad reikštų aminorūgščių liekanų polimerą (arba seką). Paprastai polimeras turi mažiausiai apie 30 aminorūgščių liekanų ir paprastai mažiausiai apie 50 aminorūgščių liekanų. Paprastai jie turi mažiausiai apie 100 aminorūgščių liekanų. Šie terminai taikomi aminorūgščių polimerams, kuriuose viena ar kelios aminorūgščių liekanos yra atitinkamų natūraliai atsirandančių aminorūgščių analogai, dariniai ar mimetikai, taip pat natūraliai atsirandantiems aminorūgščių polimerams. Pvz., polipeptidai gali būti modifikuoti arba daromi, pvz., pridėdant angliavandenių liekanų, kad susidarytų glikoproteinai. Sąvokos „polipeptidas“ ir „baltymas“ apima glikoproteinus, taip pat ne glikoproteinus.

„Aminorūgščių seka“ reiškia aminorūgščių, apimančių polipeptidą arba baltymą, eiliškumą ir tapatumą.

Terminas „atranka“ reiškia procesą, kurio metu nustatoma viena ar daugiau vienos ar kelių biologinių molekulių savybių. Pavyzdžiui, tipiški atrankos procesai apima tuos, kuriuose nustatomos vienos ar kelių vienos ar daugiau bibliotekų vieno ar kelių narių savybės.

Terminas „atranka“ taip pat reiškia procesą, kurio metu nustatoma, kad viena ar daugiau biologinių molekulių turi vieną ar daugiau dominančių savybių. Taigi, pavyzdžiui, biblioteką galima atrinkti taip, kad būtų nustatyta viena ar daugiau vienos ar kelių bibliotekos narių savybių. Jei nustatoma, kad vienas ar keli bibliotekos nariai turi dominančią savybę, ji pasirenkama. Atranka gali apimti bibliotekos nario izoliaciją, tačiau tai nėra būtina.

Terminas „fragmentas“ reiškia, bet kurią visos nukleorūgščių arba aminorūgščių sekos dalį.

Sąvokos „biblioteka“ arba „populiacija“ reiškia mažiausiai dviejų skirtingų molekulių ir (arba) simbolių stygų, tokių kaip nukleorūgščių sekos (pvz., genai, oligonukleotidai ir kt.), arba iš jų ekspresijos produktų (pvz., fermentų), rinkinį. Biblioteka ar populiacija paprastai apima daugybę skirtingų molekulių. Pavyzdžiui, bibliotekoje ar populiacijoje paprastai yra mažiausiai apie 10 skirtingų molekulių. Didelės bibliotekos paprastai apima mažiausiai apie 100 skirtingų molekulių, paprastai mažiausiai apie 1000 skirtingų molekulių. Kai kuriais atvejais bibliotekoje yra bent 10000 ar daugiau skirtingų molekulių.

Baltymų grandinių palyginimui naudojamas terminas „tapatumas“ (baltymų ir polipeptidų) aminorūgščių sekų atžvilgiu. „Sekos tapatumo“ apskaičiavimas tarp dviejų sekų atliekamas taip. Sekos yra suderintos optimaliausio palyginimo tikslais (pvz., siekiant

optimalaus suderinimo, gali būti įterpti tarpai tiek vienoje, tiek kitoje pirmoje ir antroje aminorūgščių sekose, o palyginimo tikslais į nehomologines sekas galima nepaisyti). Optimalus derinimas nustatomas kaip geriausias balas naudojant „search36“ programą FASTA36 programinės įrangos pakete (<http://faculty.virginia.edu/wrpearson/fasta/>) su „Blosum50“ balų matrica su atotrūkio nuobauda –10, o atotrūkio prailginimas – –2. Tada palyginami aminorūgščių likučiai atitinkamose aminorūgščių padėtyse. Kai pirmosios sekos poziciją užima ta pati aminorūgštis, liekana atitinkamoje antrosios sekos padėtyje, tada molekulės toje pačioje padėtyje yra identiškos. Procentinis tapatumas tarp dviejų sekų yra identiškų pozicijų, dalijamų sekomis, skaičiaus funkcija.

Terminas „funkcinis baltymas“ arba „funkcinė seka“ reiškia baltymą, kurio forma yra tokia, kokia jis apibūdinamas.

Terminas „žyma“, „žymos seka“ arba „baltymo žyma“ reiškia cheminį fragmentą - nukleotidą, oligonukleotidą, polinukleotidą arba aminorūgštį, peptidą ar baltymą ar kitą cheminę medžiagą, kuri, pridėjant prie kitos sekos, suteikia papildomo naudingumo. arba suteikia šiai sekai naudingų savybių, ypač aptikimo ar išskyrimo atveju. Taigi, pavyzdžiui, prie pradmens arba zondo sekos gali būti pridėjama homopolimerų nukleorūgščių seka arba nukleorūgščių seka, komplementuojanti gaudymo oligonukleotidą, kad palengvintų tolesnį pailginimo produkto ar hibridizuoto produkto išskyrimą. Baltymų žymių atveju histidino liekanos (pvz., nuo 4 iki 8 iš eilės esančių histidino liekanų) gali būti pridėamos prie baltymo amino- arba karboksi-galo, kad būtų lengviau atskirti baltymus chelatinės metalų chromatografijos būdu. Kaip alternatyva - aminorūgščių sekos, peptidai, baltymai ar sintezės partneriai, reprezentuojantys epitopus ar jungiančius determinantus, reaguojančius su specifinėmis antikūnų molekulėmis ar kitomis molekulėmis (pvz., vėliavos epitopas, c-myc epitopas, transmembraninis gripo A viruso epitopas, hemagglutinino baltymas, A baltymas, celiuliozė) jungiantis domenai, kalmoduliną surišantis baltymas, maltozę rišantis baltymas, chitiną surišantis domenai, glutationo S-transferazė ir panašiai) gali būti pridėjami prie baltymų, kad būtų lengviau atskirti baltymus tokiomis procedūromis, kaip giminingumo ar imunogiminingumo chromatografija. Cheminiai etikečių fragmentai apima tokias molekules kaip biotinas, kurios gali būti pridėamos prie nukleorūgščių arba baltymų ir palengvina išskyrimą ar aptikimą sąveikaujant su avidino reagentais ir panašiai. Išmoktas amatininkas žino ir gali įsivaizduoti daugybę kitų etikečių dalių, kurios, kaip manoma, patenka į šio apibrėžimo taikymo sritį.

Terminas „duomenų papildymas“ reiškia strategiją, kuri leidžia dirbtinai padidinti mokymui skirtų duomenų įvairovę, fiziškai nerenkant duomenų pavyzdžių. Vaizdų duomenų padidinimo metodų pavyzdžiai yra apkarpymas, užpildymas ir horizontalus apversimas.

Terminas „duomenų rinkinys“ reiškia elementų, kurie naudojami treniruoti ar įvertinti neuroninius tinklus, rinkinį.

Terminas „tikrasis paskirstymas“ reiškia paskirstymą, kuriame yra visi tikrieji elementai, įskaitant duomenų rinkinio elementus.

Terminas „bloškai“, kalbant apie neuroninius tinklus, reiškia architektūrinių neuroninių tinklų komponentų, sujungtų kartu ir pakartotinai panaudotų, grupę.

Terminas „diferencijuojamas diskretusis derinimas“ reiškia funkciją, kuri paverčia nuolatines reikšmes atskira erdve ir ši funkcija yra diferencijuojama.

Terminas „žodyno dydis“ reiškia keletą unikalių žetonų, naudojamų duomenų rinkinio elementams sudaryti. Šie žetonai yra atskiri (pvz., aminorūgštys).

Terminas „mokymo žingsnis“ reiškia neuroninio tinklo optimizavimo ciklą, kuriame apdorojamas elementų rinkinys, kai rinkinio dydis yra lygus grupės dydžiui.

Egzistuojančių sekų pasirinkimas ir paruošimas

Viename įgyvendinimo variantų egzistuojančios sekos gali būti specialiai parinktos generatyviam adversarinio tinklo mokymui. Pradinis sekų rinkinys(-iai) yra svarbi procedūra dėl kelių priežasčių: pasirinktos sekos apibrėžia sekų erdvę, kurioje pasirodys pagamintos funkcinės sintetinės sekos (i), pasirinktų sekų ypatybės apibrėžai nežinomą pasiskirstymą, kuris gali būti apytiksliai suderintame konkurencinio tinklo mokymosi etape (ii) ir savo ruožtu gali apibrėžti kai kurias pagamintų sintetinių sekų savybes. Eksperimentinis, duomenų pagrįstas pavyzdys parodytas 15 pav. Šiame paveiksle pateikiami natūralūs ir sintetiniai (aprašyto metodo išvestis), kur atstumai tarp skirtingų grupių yra panašūs į klasterių sekos panašumus ir kitas panašias charakteristikas. Kaip aprašyta anksčiau, sintetinės sekos atsiranda apytikslėse natūralių klasterių nustatytose ribose, todėl pirmasis metodo žingsnis - sekų parinkimas - yra labai svarbus.

Pavyzdžiui, norint ištirti sekų erdvę, kurioje yra funkciniai glicerolio-3-fosfato dehidrogenazės variantai, galima pasirinkti treniruočių sekas, patenkančias į tą sekos erdvės plotą. Tokios sekos gali būti glicerolio-3-fosfato dehidrogenazės homologai. Šios funkcinės sekos gali būti gaunamos iš viešųjų duomenų bazių, metagenomikos patikros, atsitiktinės mutagenezės patikros, racionalaus variantų patikrinimo ar kitų šaltinių. Tada surinktas duomenų rinkinys gali būti toliau modifikuojamas.

Tada pasirinktos sekos gali būti apdorotos bioinformatiniais algoritmais. Šis žingsnis yra labai svarbus, nes neperdirbtos sekos, naudojamos treniruojant generatyvų adversarinį tinklą, turi didelę tikimybę gauti nefunkcines ir (arba) netirpias galutines pagamintas sintetinių baltymų sekas.

Viename įgyvendinimų rinkinyje išankstinis pasirinktų baltymų sekų apdorojimas gali apimti sekų filtravimą pagal apibrėžtus kriterijus, tokius kaip sekos kilmė, panašumas, įvairovė, sekų grupių dydžiai, struktūros panašumas, domenų buvimas, funkcija ar funkcinės savybės, statistinės savybės (pvz., aminorūgščių dažnis arba nekanoninių aminorūgščių buvimas, darbo sąlygos), fizikinės ir cheminės savybės ar kiti panašūs būdai.

Duomenų didinimo būdai gali būti naudojami padidinant pasirinktų sekų skaičių ir (arba) įvairovę (pvz., įvykiuose, kai pasirinktos sekos skaičius yra per mažas, kad būtų galima naudoti aprašytą metodą), pavyzdžiui, invariantinių transformacijų įvedimas, interpoliacija, triukšmo įvedimas arba kitos technikos.

Dar kitame įgyvendinimo variantų rinkinyje pasirinktos sekos gali būti konvertuotos į skirtingas reprezentacijas, tokias kaip vieno karšto kodavimo, sekos įdėjimo (sekų konvertavimo į skaitines reikšmes) ar kita. Šiuos skirtingus vaizdus taip pat galima modifikuoti pridėdant arba pašalinant kiekybinę ar kokybinę informaciją tokiais būdais kaip sujungimas, įvesties daugyba ar kita.

Generatyvinių adversarinių tinklų architektūra baltymų sekų generavimui

Pasirinktos ir iš anksto apdorotos sekos gali būti naudojamos, kaip mokomosios (pavyzdžių) sekos generatyviniams adversariniams tinklams. Toliau aprašoma generatyvinių adversarinių tinklų, reikalingų funkcinėi baltymų sekai generuoti, architektūra.

Tolesniuose punktuose esantys nuorodų numeriai turėtų būti suprantami kaip pavyzdys, o kiti panašūs architektūros variantai taip pat gali būti perspektyvūs.

Generacinę adversarinio tinklo architektūrą sudaro du neuroniniai tinklai: generatoriaus tinklas 1 pav. 101 ir diskriminacinis tinklas 1 pav. 102. Generatoriaus tinklo 1 pav. 101 funkcija yra pateikti išėjimus 1 pav. 103, kurie, atrodo, yra paimti iš tikro duomenų rinkinio 1 pav. 104 paskirstymo, neturint prieigos prie paskirstymo mokymo metu. Diskriminacinis tinklas 102 gauna įvestis 1 pav. 104 iš duomenų rinkinio ir generatoriaus 1 pav. 101 ir jam pavesta atskirti sugeneruotus elementus nuo realių. Paprastai generalaus adversarinio tinklo mokymą sudaro: atsitiktinai pasirenkami taškai iš pasirinkto paskirstymo 1 pav. 105 ir pavyzdžiai sugeneruojami 1 pav. 103 naudojant generatorių 1 pav. 101 (i), atsitiktinai pasirenkant elementus iš duomenų rinkinio 1 pav. 104 (ii), naudojant diferenciatorių 1 pav. 102, kad būtų gauti balai 1 pav. 106 generuojamiems 103 ir duomenų rinkinio pavyzdžiams 1 pav. 104 (iii), naudojant diskriminavimo balus 1 pav. 106, siekiant optimizuoti diskriminavimo tinklą 1 pav. 102 ir generatoriaus tinklą 1 pav. 101 nepriklausomai (iv), pakartojant aprašytus „i-iv“ veiksmus, kol sugeneruoti mėginiai bus norimos kokybės, arba nesugeba atskirti sugeneruotų 1 pav. 103 pavyzdžių nuo realių 1 pav. 104. Diskriminatorių ir generatorių tinklai taip pat gali būti aprūpinti papildoma informacija 1 pav.

107, kad bendras generatyvusis adversarinis tinklas būtų sąlygotas pateiktos papildomos informacijos.

Viename įgyvendinimo variantų generatyvią adversarinio tinklo architektūrą sudaro du tinklai - generatorius 1 pav. 101 ir diskriminatorius 1 pav. 102 - kiekviename iš jų gali būti daugybė statybinių blokų, tokių kaip Resnet blokai 2 pav. 201, 4 pav. 401 (He ir kt., 2015). Kaip alternatyva „Resnet“ blokams gali būti naudojami konvoliuciniai sluoksniai, visiškai sujungti sluoksniai, daugiagalvis dėmesio mechanizmas (Vaswani ir kt., 2017) ar kiti architektūriniai statybiniai blokai.

Kitame įgyvendinimo variantų rinkinyje generatoriaus įvestis 2 pav. 105 gali būti vektorius, nubrėžtas iš bet kokio žinomo pasiskirstymo, tokio kaip vienodas arba normalus. Generatorių tinkle gali būti vienas ar daugiau visiškai sujungtų 2 pav. 201 konvoliucinių sluoksnių prieš „ResNet“ blokus 2 pav. 202 (pvz., 6 „Resnet“ blokai 202–1–6]), kad būtų galima transformuoti įvestį 2 pav. 105 į reikiamus matmenis. Generatorių tinklas gali turėti vieną ar kelis savitvarkos (Zhang ir kt., 2018) sluoksnius 2 pav. 203. Generatorių tinklą gali sudaryti vienas ar keli visiškai sujungti arba konvoliuciniai sluoksniai 2 pav. 204 su netiesine aktyvavimo funkcija, tokiais kaip nesandari ReLu 2 pav. 205, ReLu ir kiti. norimiems matmenims išvesti 2 pav. 103. Išvestis gali būti perduodama per netiesinę aktyvavimo funkciją (pavyzdžiui, Tahn, Softmax ir kt.), Taip pat diferencijuojamą diskretinį išvesties apytikslį, pvz., „Gumbel-Softmax 206“ arba „REINFORCE“ (Williams 1992). Be to, mokymų metu generatorių tinklas taip pat gali būti aprūpintas papildoma informacija 2 pav. 107, pavyzdžiui, klasės etikele, kuri gali būti užkoduota naudojant įterpimus, vienetiniu kodavimu arba transformuota kitais būdais ir tada sujungta su vienu ar keliais sluoksniais.

Kituose įgyvendinimo variantų rinkiniuose kiekvienas „Resnet“ blokas generatoriuje 2 pav. 201 gali būti sudarytas iš 1–10 perkeltų konvoliucijos sluoksnių 3 pav. 301 (pvz., 2 perkeltų konvoliucijos sluoksnių 301–1–2) ir 1–10 konvoliucijos sluoksnių 3 pav. 302 su filtro dydžiu (nuo 1 iki 100) x (nuo 1 iki 100). Konvoliucijos sluoksniuose gali būti išsiplėtimo greitis nuo 1 iki 10000. Blokuose gali būti daugybė normalizavimo sluoksnių, tokių kaip paketų normalizavimas (Ioffe ir Szegedy 2015), egzempliorių normalizavimas (Ulyanov, Vedaldi ir Lempitsky 2016) ir kiti. Be to, blokuose taip pat gali būti įvairių aktyvavimo funkcijų, tokių kaip nesandarus „ReLu 303“ („Maas 2013“) (pvz., 2 nesandarūs „ReLu“ aktyvinimai 303–1–2], „ReLu“ (Nair ir Hinton 2010) ir kitos. Blokuose taip pat gali būti 1-10 praleidžiamų jungčių 3 pav. 304, kurios gali būti sujungtos 3 pav. 305 su kitomis bloko dalimis. Norint padidinti sluoksnio išvesties matmenis, o ne perkeltą konvoliucinį sluoksnį 3 pav. 301, galima naudoti artimiausio kaimyno interpoliaciją, subpikselių judesį (Shi ir kt., 2016) ar kitus metodus.

Kituose įgyvendinimų rinkiniuose įvestis 4 pav. 104 į diskriminacinį tinklą gali būti

viena karšta koduotė, kurios žodyno dydis svyruoja nuo 10 iki 10 000 arba panašiai. Arba įvestis gali būti užkoduota naudojant aminorūgščių įterpimus arba fizikinius ir cheminius požymius. Diskriminuojančiame tinkle gali būti vienas ar keli 4 pav. 401 įterpimo, konvoliucijos arba visiškai sujungti sluoksniai prieš 4 pav. 402 „Resnet“ blokus (pvz., 6 „Resnet“ blokai 402–1–6]), kad būtų galima transformuoti įvestį 104. Be to, jame gali būti vienas ar daugiau savireguliacijos sluoksnių 4 pav. 403. Diskriminuojančiame tinkle gali būti sluoksnis, palaikantis didelę įvairovę tarp generuojamų sekų, tokių kaip minibatų standartinio nuokrypio sluoksnis 4 pav. 404, kaip aprašyta (Karras ir kt., 2017). Diskriminacinį tinklą gali sudaryti vienas ar keli 4 pav. 405 konvoliucijos, visiškai sujungti 4 pav. 406 sluoksniai arba bendras vidutinis telkimas su netiesinėmis aktyvavimo funkcijomis, tokiomis kaip nesandari ReLu 4 pav. 407, ReLu ir kitos, kad būtų gauti norimų matmenų išėjimai. Kai kurie sluoksniai gali būti išlyginti naudojant 4 pav. 408 sluoksnius. Išlyginamojo elemento rezultatas gali būti perduodamas per netiesinę aktyvavimo funkciją, tokią kaip „Softmax“, „Tanh“ ar kita.

Kituose įgyvendinimų rinkiniuose kiekvienas atkuriamasis blokas diskriminatoriuje gali turėti 1, 2, 3, 4, 5, 6, 7, 8, 9 arba 10 5 pav. 501 konvoliuciją (pvz., 3 konvoliucijos sluoksniai 501–1–3]) ir/arba visiškai sujungtus sluoksnius su filtru, kurio dydis yra nuo 1 iki 100) x (nuo 1 iki 100). Konvoliucijos sluoksniuose gali būti išsiplėtimo greitis (nuo 1 iki 10000). Blokuose gali būti daugybė normalizavimo sluoksnių, tokių kaip paketų normalizavimas 5 pav. 502 (pvz., 2 paketų normalizavimo sluoksniai 502 - [1-2]), egzempliorių normalizavimas ir kiti. Blokuose taip pat gali būti įvairių netiesinių aktyvavimo funkcijų, tokių kaip nesandari ReLu 5 pav. 503, ReLu ir kitos. Blokuose taip pat gali būti 1-10 praleidžiamų jungčių 5 pav. 504, kurios gali būti sujungtos 5 pav. 505 su kita bloko dalimi. Mokymų metu diskriminacinis tinklas taip pat gali būti aprūpintas papildoma informacija 1 pav. 107 kartu su iš anksto apdorotomis treniruočių sekomis, tokiomis kaip klasės etiketė, kuri gali būti užkoduota naudojant įterpimus, „karštas“ kodavimas arba transformuota kitais būdais, o tada sujungta su vienas ar keli diskriminuojančiojo tinklo sluoksniai.

Kituose įgyvendinimų rinkiniuose tinklo praradimui neprisotinamas (Goodfellow ir kt., 2014), neprisotinamas R1 sureguliuojamas (Mescheder, Geiger ir Nowozin 2018), vyras (Tran, Ranganath ir Blei 2017; Lim ir Ye 2017; Miyato ir kt., 2018), gali būti naudojami vyriai su reliatyvistiniu vidurkiu (Jolicoeur-Martineau 2018), Wasserstein (Arjovsky, Chintala ir Bottou 2017) ir Wasserstein su gradientu nuobauda (Gulrajani et al. 2017) ar kitos funkcijos. Norint užtikrinti Lipschitz apribojimų spektrinį normalizavimą (Miyato ir kt., 2018 m.), Gali būti naudojamos nuolydžio nuobaudos (Gulrajani ir kt., 2017 m.).

Kituose įgyvendinimų rinkiniuose generuotų išėjimų matmenys priklauso nuo maksimalios sekos, kurią reikia generuoti, ilgio ir naudojamo diskriminuojančiojo tinklo kodavimo tipo. Pvz., jei didžiausias sekos ilgis yra 400 aminorūgščių ir vienas karštas

kodavimas, kurio žodyno dydis yra 21, generuojamos išvesties matmenys būtų 400x21.

Priklausomai nuo pasirinktų sugeneruotų išėjimo matmenų, sekos, pasirinktos mokymui, gali būti toliau filtruojamos, kad būtų pašalintos sekos, turinčios daugiau aminorūgščių, nei leidžia išėjimo matmenys. Pvz., jei sugeneruotų išėjimų matmenys yra 400x21, sekų duomenų rinkinį galima filtruoti, kad būtų pašalintos sekos, kuriose yra daugiau nei 400 aminorūgščių. Duomenų rinkinys taip pat gali būti suskirstytas į grupes su konkrečiomis tapatybėmis. Pavyzdžiui, to galima pasiekti naudojant grupavimo įrankius, tokius kaip mmseq2 ar kitus. Klasterizavimas leidžia subalansuoti generacinį adversarinio tinklo mokymo procesą, kuris yra svarbus norint pasiekti sintetinę funkcinę sekos variaciją. Sekas pagal jų klasterio dydį galima sugrupuoti į įvairaus dydžio kaušus (1,2,3,5,10,20,30 ir kt.). Tada padidėjęs mėginių ėmimo koeficientas nustatomas padalijus maksimalų kaušo dydį iš klasterio kaušo dydžio visiems kaušams. Šis koeficientas yra naudojamas treniruočių metu atrenkant reprezentatyvias grupes. Dalį duomenų rinkinio galima pasirinkti atsitiktinai arba racionaliai ir išimti iš mokymo duomenų rinkinio. Tokios sekos gali veikti kaip patvirtinimo sekos, kurių tinklas nematys mokymo metu, bet vėliau gali būti panaudotos tinklo našumo analizės tikslams.

Kituose įgyvendinimų rinkiniuose, siekiant optimizuoti neuroninio tinklo svorius, generatoriaus ir diskriminacinių tinklų kūrimui gali būti naudojami ADAM optimizatorius (Kingma ir Ba 2014), Stochastinis gradiento nusileidimas (Kiefer ir Wolfowitz 1952), RMSProp (Graves 2013) ir kiti optimizatoriai. Mokymosi lygis gali būti palaipsniui mažinamas tiek generatoriaus, tiek diskriminuojančiojo atžvilgiu, siekiant padidinti mokymo stabilumą ir palengvinti konvergenciją. Pavyzdžiui, laipsniškas mokymosi greičio mažėjimas gali būti nuo $1e-3$ iki $5e-5$. Santykis tarp generatoriaus ir diskriminuojančiojo mokymo žingsnių gali būti 1:1, 1:2, 1:5 ar kitas.

Dar viename išradimų rinkinyje, norint normalizuoti duomenų klasterio dydžius vykstant konkurencinio tinklo mokymui, gali būti dinamiškai imami nepakankamai pavaizduoti sekų rinkiniai. Tai gali būti pasiekta imant nepakankamai atstovaujamas grupes (dubliuojančias sekas klasterio viduje), imant ankstesnių etapų skaičiavimo koeficientą. Šis procesas gali būti kartojamas visos kartos adversarinio tinklo mokymo metu, siekiant išsaugoti sekos kitimą. Sekos gali būti dinamiškai užklijuotos, kad būtų pažymėta, kad nėra aminorūgšties. Tai gali būti naudojama trumpesnėms sekoms kompiliuoti, jei sukonstruotame tinkle yra sluoksnių, kuriems reikia fiksuoto dydžio įvesties, pavyzdžiui, visiškai sujungtų. Sekos gali būti užklijuotos iš kairės, dešinės arba iš abiejų pusių. Plėvelė pašalinama iš sugeneruotų sekų, kai gaminama galutinė išvestis (pavyzdžiui, kai viena karšta koduotė sekos virsta vienos raidės aminorūgščių sekomis).

Dar kitoje realizacijų aibėje, norint žinoti tinklo galią, sugeneruoti duomenys turėtų būti įvertinami mokymo proceso metu. Pavyzdžiui, kas 1200 iteracijų sugeneruotos sekos galėtų būti automatiškai palyginamos su mokymosi ir validavimo duomenų aibių sekomis naudojant BLAST ar panašius algoritmus. Taip pat, mokymosi metu sugeneruotos sekos periodiškai galėtų būti naudojamos skaičiuojant biosum⁴⁵, e-value ir tapatumo įverčius.

Dar kitoje realizacijų aibėje, po generatyvinio adversarinio tinklo mokymosi, norint gauti baltymų sekas iš išmokto skirstinio, pasirenkami atsitiktiniai taškai, kurie buvo naudoti mokantis. Siekiant pagerinti sugeneruotų pavyzdžių kokybę, naudojamo skirstinio standartinis nuokrypis gali būti sumažintas imties įvairovės kaina. Tuomet šie taškai yra tiesiogiai praskleidžiami per ištreniruotą generatorių ir gaunamos sugeneruotos reprezentacijos iš nustatyto tikrojo skirstinio kuris buvo išmoktas mokymosi metu. Gauta reprezentacija (užkoduota vienas-nulis arba įterpiniau) tuomet paverčiama į aminorūgščių seką ir visi sekos pradžioje ir pabaigoje esantys tarpai yra pašalinami.

Gautų sintetinių baltymų sekų apdorojimas

Sintetinių baltymų sekos gautos iš generatyvinio adversarinio tinklo nustatyto skirstinio gali būti toliau apdorojamos naudojant bioinformatikos metodus. Šis žingsnis labai svarbus, nes gali ženkliai padidinti eksperimentiškai veikiančių baltymų sekų radimo tikimybę.

Vienoje realizacijų aibėje apdorojimas gali apimti skaitmeninį gautų sintetinių sekų filtravimą. Tokios filtravimo procedūros gali būti naudojamos ranguoti sintetines sekas pagal apsibrėžtus kriterijus, tokius kaip diskriminatoriaus įvertis, sugeneruotas kokybinis ar kiekybinis aprašas, įverčiai ar priklausymas tam tikru modeliu (pvz. mašininio mokymosi modeliais, kiekybiniais struktūros-savybės priklausomybės modeliais, struktūrinės ar molekulinės dinamikos modeliais) prognozuojamai klasei ar kt.

Kitoje realizacijų aibėje, sintetinių sekų apdorojimas gali būti sekų modifikavimas pridėdant stabilizuojančias mutacijas, jungiamąsias sekas, baltymų žymeklius, sujungiant sekas su kitų baltymų sekomis ar kt.

Sukurtos veikiančių baltymų bibliotekos naudojimas

Apibūdinto metodo išvestis – gerai veikiančių baltymų sekų biblioteka – gali būti taikoma įvairiai, pavyzdžiui, baltymų sąveikos eksperimentuose, duomenų papildyme ir kt. Veikiančių sekų biblioteka gali būti fiziškai sukonstruota naudojant genų ar baltymų sintetinio metodo. Tada fizinė biblioteka gali būti eksperimentiškai tiriama naudojant standartinius metodus, tokius kaip *in vitro/in vivo* baltymų ekspresija ir charakteristikų nustatymas, mikroskopsčių sistemos ir kt. Baltymai gali būti rūšiuojami pagal įvairias charakteristikas, tokias kaip baltymo varianto atliekamos cheminės reakcijos, aktyvumo lygis,

termostabilumas, tirpumas ir kt. Pavyzdinis veikiančių baltymų bibliotekos generavimas ir eksperimentinis tyrimas apibūdinamas 1 pavyzdyje. Veikiančių sekų biblioteka sukurta apibrėžtu metodu taip pat gali būti naudojama duomenų papildymo tikslais. Tokiais atvejais metodas naudojamas praplėsti sekų aibę naudojamą kito mašininio mokymosi algoritmo sekomis sukurtomis apibrėžto išradimo. Tokių algoritmų pavyzdžiai galėtų būt optimalios fermento katalitinės temperatūros prognozavimas, antrinės baltymo struktūros prognozavimas ir kt.

PAVYZDŽIAI

Čia išradimas apibūdinamas detaliau su išnašomis į pavyzdžius, nors šio išradimo techninė apimtis nėra limituojama šiais pavyzdžiais.

1 pavyzdys. Funkcinių malato dehidrogenazės sekų kūrimas

Šis pavyzdys yra funkcinės malato dehidrogenazės (E.C. 1.1.1.37) sintetinių baltymų sekų kūrimas naudojant apibūdintą išradimą. Šio pavyzdžio tikslas yra pavaizduoti kaip kiekvienas metodo žingsnis galėtų būti įvykdytas.

Šiame pavyzdyje generatyvinio adversarinio tinklo architektūra sudaryta iš dviejų neuroninių tinklų – diskriminatoriaus ir generatoriaus – kurių kiekvienas naudoja ResNet blokus. Šio generatyvinio adversarinio tinklo bendra architektūra pavaizduota schemeje 7 pav. Kiekvienas diskriminatoriaus blokas susideda iš 3 konvoliucinių tinklų, kurių dydis 3x3, 2 grupinės normalizacijos (batch normalization) sluoksnių ir kiaurų (leaky) ReLU aktyvacijų. Generatoriaus likutiniai blokai sudaryti iš dviejų transponuotų konvoliucinių sluoksnių, vieno konvoliucinio sluoksnio su tuo pačiu filtro dydžiu (3x3) ir kiaurų ReLU aktyvacijų. Kiekvienas tinklas turi po vieną save stebintį (self-attention) sluoksnį. Transponuotos konvoliucijos technika buvo pasirinkta praturtinimui (up-sampling), nes pademonstravo geriausių eksperimentinius rezultatus. R1 regularizacija naudota kaip nesotinanti nuostolių funkcija. Mokymo stabilumui užtikrinti spektrinė normalizacija buvo naudojama visuose sluoksniuose.

Diskriminatoriaus įvestis užkoduota vienas-nulis (one-hot) formatu naudojant 21 simbolio dydžio žodyną (20 kanoninių aminorūgščių ir ženklas skirtas užkoduoti likusią vietą sekos pradžioje ir pabaigoje). Vektorius su 128 reikšmėmis spėtomis iš atsitiktinio skirstinio su vidurkio verte 0 ir vidutinio nuokrypio verte 0,5 naudotas generatoriaus įvesčiai, išskyrus vertes, kurių dydis nutolęs daugiau nei 2 vidutiniai nuokrypiai nuo vidurkio – jos pakartotinaai spėjamos. Sugeneruotos išvesties dimensijos yra 512x21, kurioje kai kurios pozicijos žymi likusią vietą.

Šiame pavyzdyje bakterinės malato dehidrogenazės (MDH) sekos buvo surinktos iš viešai prieinamos baltymų sekų duomenų bazės Uniprot. Sekos ilgesnės nei 512 aminorūgščių arba turinčios nekanoninių aminorūgščių buvo išmetamos. Galutinis duomenų

rinkinys susidėjo iš 16898 sekų, kurios siekiant subalansuoti duomenų rinkinį mokymo metu buvo suklasterizuotos į 70% sekos identiškumo klasterius naudojant MMseqs2 įrankį (Steinegger and Söding 2017). 20% klasterių su mažiau nei 3 sekomis buvo atsitiktiniu būdu atskirti validacijai (192 sekos) ir visa likusi duomenų rinkinio dalis naudota mokymui (16706 sekos). Aštuonios reprezentatyvios, natūralios MDH sekos iš mokymo rinkinio prisegtos (SEQ ID Nr.1 – SEQ ID Nr. 8).

Generatoriaus ir diskriminatoriaus mokymo žingsnių santykis buvo parinktas 1:1. ADAM algoritmas buvo naudotas optimizuoti abu neuroninius tinklus. Mokymo greitis palaipsniui mažintas nuo $1e-3$ iki $5e-5$ generatoriui ir diskriminatoriui kartu. Siekiant išvengti tendencingumo (bias) sekoms kurios turi didelį homologiškų sekų skaičių, maži sekų klasteriai buvo dinamiškai praturtinami mokymo metu. Mokymo našumo sekimui kartu su GAN nuostoliais, sugeneruoti duomenys buvo nuolatos vertinami. Nestabdant mokymo proceso, kas 1200 mokymo žingsnių sugeneruotos sekos buvo automatiškai prilygiuojamos mokymo ir validacijos duomenų rinkiniais naudojant BLAST (8 pav.). Mokymas užtruko 210 valandų (~9 dienas) naudojant NVIDIA Tesla P100 (16 GB).

Po 2,5 M mokymo žingsnių, nutraukus mokymą, vidutinis sekų identiškumas tarp sugeneruotų ir natūralių sekų rinkinių pasiekė plynaukštę (plateau) (sekų identiškumo mediana panašiausioms natūralioms sekoms buvo 61.3%, (9 pav.). Po pradinio kokybės įvertinimo, 20 000 sekų buvo sugeneruotos tolimesnei išmokyto neuroninio tinklo analizei.

Neuroninio tinklo gebėjimas išmokti kurios pozicijos sekoje yra konservatyvios ir kurios yra variabilios įvertintas skaičiuojant Šenono entropiją kiekvienai pozicijai neuroninio tinklo sugeneruotoms ir natūralioms sekoms (1 pav.).

Variabilumas sugeneruotų sekų atskirose pozicijose buvo ypač panašus į natūralių sekų, kur pikai (aukšta entropija) ir įdubos (žema entropija) matoma panašiose pozicijose sekų palyginyje. Iš tiesų matoma beveik ideali koreliacija tarp entropijos verčių sugeneruotose ir natūraliose sekose (Pirsono $r = 0,89$, P-vertė $< 1e-16$). Sugeneruotos sekos išlaikė substratą jungiančias ir katalizinio centro aminorūgštis išmokdamas konservatyvių aminorūgščių pozicijas, kurios kritiškai svarbios katalizei (11 pav.).

Tolimesnė sugeneruotų ir natūralių sekų palyginamoji analizė atskleidė, kad net ir ypatingai variabiliuose sekų regionuose individualių aminorūgščių dažniai idealiai koreliavo (Pirsono $r = 0,96$, P-vertė $< 1e-16$, 12 pav.).

Rezultate mūsų specifinė neuroninio tinklo architektūra išmoko specifinius fizikocheminius bruožus variabiliuose sekų regionuose, kurie yra unikalūs kiekvienai homologiškai sekai, tačiau sudaro tokį patį suminį fizikocheminį parašą. Pavyzdžiui, nepaisant didelio sekų variabilumo, hidrofobinių, aromatinių, įkrautų ir cisteiną turinčių

aminorūgščių dalys buvo tokios pačios sugeneruotose sekose (Vikloksono rangų sumos (Wilcoxon rank-sum) testo P-vertė $> 0,05$) kaip ir natūraliose. Nepaisant skirtumų hidrofobinių ir polinių krūvio neturinčiose aminorūgščių (P-vertė = $7e-5$; $1e-28$, atitinkamai) dalyse, neuroninis tinklas išmoko bendrus aminorūgščių bruožus atspindinčius evoliucinį ir fizikocheminį kontekstą (1 lentelė).

1 lentelė. Fizikocheminės aminorūgščių savybės.

Aminorūgštis	Statistika	p-vertė	Savybės
W	-44.5535	0	-
T	-32.6757	3.45E-234	-
N	-31.0134	3.55E-211	-
P	5.414836	6.13E-08	-
F	36.10193	2.12E-285	-
A	-7.00421	2.48E-12	-
G	2.373175	0.017636	-
I	10.88373	1.38E-27	-
L	24.0913	3.08E-128	-
H	0.883687	0.376865	-
R	16.52561	2.40E-61	-
M	2.927137	0.003421	-
V	-37.0289	3.93E-300	-
E	-17.0094	7.00E-65	-
Y	-2.82555	0.00472	-
V, I, L, F, W, Y, M	-0.64435	0.519345	Hidrofobinė
S, T, H, N, Q, E, D, K, R	-3.965	7.34E-05	Hidrofobinė
F, W, Y, H	-1.00434	0.315217	Aromatinė
P, G, A, S	7.263205	3.78E-13	Maža
K, R, H	5.612656	1.99E-08	Teigiama

D, E	-22.965	1.04E-116	Neigiama
V, I, L, M	-2.04672	0.040685	Alifatinė
S, C, T, M	-0.12037	0.904194	Hidroksi grupę turinti
S, T, C, M, N, Q	-11.1188	1.02E-28	Polinė neturinti krūvio
H, K, R, E, D	1.296764	0.194713	Krūvį turinti

Baltymuose augelis aminorūgščių porų nutolusios pirminėje struktūroje yra erdviškai artimos ir sąveikauja tretinėje struktūroje, stabilizuodamos tinkamą baltymo struktūrą ir užtikrindamos funkciją. Mes ištyrėme ar neuroninis tinklas sugeba išmokyti tokias lokalias ir globalias aminorūgščių sąsajas nagrinėdami ilgo atstumo aminorūgščių porų sąsajas visame MDH sekos ilgyje. Visoms sugeneruotoms MDH sekoms mes suskaičiavome aminorūgšties asociacijos įverčius naudodami minimalaus proksimalumo funkciją Z_m (Santoni et al. 2016). Funkcija $Z_m(A,B)$ skaičiuoja artimiausią vidutinį atstumą nuo kiekvienos aminorūgšties „A“ iki kitos bet kurios aminorūgšties „B“ sekoje kuri yra išreikšta kaip visų porų tikimybių matrica (13 pav.).

Natūralių (mokymo rinkinio) ir sugeneruotų sekų matricos buvo 88% identiškios ir dauguma atvejų skyrėsi triptofano aminorūgšties statistika, nes 22% natūralių sekų neturėjo šios aminorūgšties. Toliau tirdami aminorūgščių porų sąsajas mes apskaičiavome visų įmanomų aminorūgščių porų kombinacijų koreliaciją imdami pozicijas iš daugybinio sekų palyginio sudaryto iš natūralių ir sugeneruotų sekų (vidurkintas Pirsono $r = 0,95$, Pav. 14), kuri rodo, kad aminorūgščių porų sąsajos yra ypatingai panašios abiejuose sekų rinkiniuose.

Tęsiant mintį, mes tikrinome ar sugeneruotos MDH sekos turėjo du pagrindinius Pfam (Finn et al. 2014) domenų, kurie yra identifikuojami ($E\text{-vertė} < 1e-10$) natūraliose MDH sekose (Ldh_1_N and Ldh_1_C). Iš tiesų – mes atskleidėme, kad 98% sugeneruotų sekų turėjo abiejus domenų, o likusi dalis bet kurį vieną iš jų. Šie rezultatai parodo, kad sekos sugeneruotos mūsų išrastu metodu yra aukštos kokybės ir atitinka MDH baltymus tiek aminorūgščių skirstiniais individualiose pozicijose, tiek ilgo atstumo sąsajose tarp aminorūgščių porų, kurie pastebimi pirminėje sekoje visoje MDH baltymų šeimoje.

Toliau mes siekėme patikrinti ar mūsų išmokytas neuroninis tinklas taip pat išmoko generalizuoti MDH baltymų šeimą ir sugeneruoti įvairias natūralias sekas. Pirmiausia mes vizualizavome sugeneruotų ir natūralių sekų įvairovę t-paskirstytu stochastiniu kaimynų įdėjimo (t-SNE) metodu dimensijų redukavimui (Maaten and Hinton 2008). Kadangi daugelis natūralių MDH sekų buvo labai panašios (poros sekų identiškumo mediana 92%), jos buvo

sugrupuotos į klasterius ir sugeneruotos sekos pateko į natūralių sekų klasterių rėmus parodant, kad MDH sekų erdvės rėmai buvo išmokti (15 pav.).

Siekdami įvertinti ar įvairios sugeneruotos sekos turi naujų ir funkciškai reikšmingų biologinių savybių, mes atlikome visų CATH duomenų bazės (Dawson et al. 2017) sekų modelių, atitinkančių visus žinomus 3D struktūrinius baltymų domenų, paiešką. Pirmiausia mes įvertinome ar neuroninis tinklas kito mokymo metu iš sugeneruotų struktūrinių domenų įvairovės (16 pav.).

Nors surastų struktūrinių domenų skaičius nustojo augti ankstyvoje mokymo stadijoje (po 0.2M mokymo žingsnių), pasiekdamas 79% visų domenų, struktūriniai CATH domenai buvo randami viso mokymo metu. Iš viso 119 naujų struktūrinių sekų motyvų ($E\text{-vertė} < 1e-16$) buvo identifikuoti (16 pav.) sugeneruotose sekose, kurie neegzistavo natūralių bakterinių malato dehidrogenazių fermentų šeimos. Paskiau mes įvertinome ar sugeneruota struktūrinių domenų įvairovė yra ne atsitiktinis įvykis. To patikrinimui mes atsitiktine tvarka įvedėme aminorūgščių pakeitimus į natūralias sekas kaip kontrolę, išlaikydami natūralius aminorūgščių dažnių skirstinius ir mutacijų dažnius atkartodami natūralių sekų variabilumą (16 pav.). Struktūrinė domenų įvairovė sumažėjo iki 38,9% modifikuotose natūraliose sekose ir 97,4% modifikuotų motyvų buvo randami natūraliose sekose, atspindint, kad atsitiktinės modifikacijos neatkūrė biologiškai reikšmingos sekų įvairovės (16 pav. įsiuvas, Fišerio tikrojo (exact) testo $p\text{-vertė} < 812e-16$). Bendrai, virš 95% sugeneruotų sekų nebuvo daugiau nei 10% panašios viena kitai (17 pav.), lyginant tik su 17% natūralių sekų toje pačioje sekų identiškumo kartelėje, išplėčiant natūralių žinomų MDH sekų įvairovę 4 kartus (17 pav. įsiuvas).

Įprastai iki 70% visų atsitiktinių mutacijų gali būti inaktyvuojančios baltymų funkcijas (Romero and Arnold 2009a; Bloom et al. 2006; Guo, Choe, and Loeb 2004a; Rennell et al. 1991; Axe, Foster, and Fersht 1998; Shafikhani et al. 1997; Rockah-Shmuel, Tóth-Petróczy, and Tawfik 2015; Sarkisyan et al. 2016), mes norėjome eksperimentiškai patvirtinti, kad sugeneruota natūraliai atrodančių naujų homologiškų fermentų įvairovė pasižymi malato dehidrogenazės aktyvumu.

Baltymų sekos buvo patikrintos įvairiais bioinformatikiais metodais prieš eksperimentinį testavimą norint padidinti funkcionalių baltymų sekų dalį sugeneruotame sekų rinkinyje. Sugeneruotos sekos buvo atrenkamos remiantis šiais kriterijais: priskiriant diskriminatoriaus įvertį kiekvienai sekai, tik sekos iš pirmojo diskriminatoriaus įverčio kvartilio buvo pasirinktos (i), sintetinės sekos buvo sulygiuotos su sekomis naudotomis neuroninio tinklo mokymui ir tos sintetinės sekos su žemesniu sekų identiškumu nei 60% pačiai panašiausiai natūraliai sekai buvo išmetamos (ii), likusios sintetinės sekos buvo struktūriškai

palygintos su sekomis naudotomis neuroninio tinklo mokymui (iii).

Sintetinių ir natūralių sekų struktūrinis palyginimas ir įvertinimas yra daugiažingsnė procedūra. Pačios panašiausios natūralios sekos kurių tretinė struktūra žinoma buvo parinktos ir priskirtos kiekvienai sugeneruotai sekai. Kiekvienai aminorūgščiai priskirtos kitos aminorūgštys toje sekoje, kurių šoninės grandinės struktūroje yra arti erdvėje. Tuomet kiekviena sugeneruota seka buvo sulygiuota su iš pradžių priskirta natūralia seka. Jeigu aminorūgštys dviejose skirtingose sekose nebuvo vienodos toje pačioje palyginio pozicijoje – kontaktų skaičius priskiriamas tai pozicijai. Galų gale sintetinės sekos su žemiausiu suminiu įverčiu buvo parenkamos (sekos, kurių aminorūgščių šoninių grandinių kontaktai mažiausiai keitėsi).

Iš visų likusių sintetinių sekų mes atsitiktine tvarka pasirinkome 40 sekų, kurios turėjo nuo 64% iki 98% sekos identišumą lyginant su panašiausia natūralia MDH seka, arba nuo 6 iki 45 mutacijų lyginant su panašiausia natūralia MDH seka. Pasirinktų baltymų genų sekos buvo susintezuotos ir rekombinantiniu būdu buvo atlikta baltymų raiška *Escherichia coli* bakterijose. Baltymai išgryninti chromatografiniu metodu ir *in vitro* patikrintas jų MDH katalitinis aktyvumas.

Detali eksperimentinė eiga apibūdinama sekančioje pastraipoje. Baltymų genų sekos išrastu metodu buvo susintezuotos, įklonuotos į pET21a raiškos vektorių ir patvirtintos kompanijoje Twist Bioscience. Prie kiekvienos fermento sekos C-galo pridėtas peptidinis jungtukas ir keturios histidino aminorūgštys (AAALEHHH). Kartu su šešiomis histidino aminorūgštimis iš raiškos vektoriaus kiekvienas baltymas galų gale turėjo deka-His žymę, kuri buvo naudojama baltymų gryninime. Raiškos plazmidės su tiksliniais genais buvo transformuotos į BL21(DE3) *E. coli* ekspresijos kamieną. Iš gauto transformantų mišinio 15 µl inokuluota į 500 µl LB augimo terpės su 100 µg/ml karbenicilino. Ląstelės buvo auginamos per naktį 32°C 96 gilių šulinėliu lėkštelėje purtant orbitinėje purtyklėje 700 sūkių per minutę. Baltymų raiška atlikta praskiedus naktines kultūras santykiu 1:30 su 1 ml autoindukcinę TB terpę su mikroelementais (Formedium, JK) ir 100 µg/ml karbenicilino ir auginant 4 val. 37°C ir per naktį 18°C purtant orbitinėje purtyklėje 700 sūkių per minutę. Ląstelės surinktos centrifugavimu ir užšaldytos -80°C per naktį. Rekombinantinių baltymų gryninimui ląstelės buvo atšildytos ir resuspenduotos 200 µl lizės buferyje (50 mM HEPES pH 7.4, 5% glicerolio, 300 mM NaCl, 0,5 mM TCEP, 0,5 mg/ml lizocimo, 10 U/ml DNazės I, 2 mM MgCl₂), ir inkubuotos 30 min kambario temperatūroje. Siekiant paskatinti lizę triton-X-100 buvo pridėtas iki galutinės 0,125% koncentracijos (v/v), ir lizatas buvo užšaldytas -80°C 30 min. Atšildžius lizatą kambario temperatūros vandens vonioje, lizatai buvo centrifuguojami 10 min 3000 x g ir supernatantas perkeltas į naują 96 šulinėlių lėkštę su 50 µl Talon dervos kiekviename šulinėlyje (Takara Bio, Japonija). Nespecifinis baltymų prisijungimas prie dervos buvo

mažinamas pridedant imidazolo iki galutinės 10 mM koncentracijos į kiekvieną šulinėlį. Lėkštelė buvo inkubuojama kambario temperatūroje 30 min purtant 400 sūkių per minutę, po kurio lizatai su derva buvo perkelti į 96 šulinėlių filtro lėkštelę (Thermo Scientific, JAV, Nunc 96-well filter plates), kuri padėta ant 96 šulinėlių surikimo lėkštelės ir centrifuguojama 1 min 500 x g specializuotoje centrifugoje. Derva tris kartus plaunama su 200 µl plovimo buferiu (50 mM HEPES pH 7,4, 5% glicerolio, 300 mM NaCl, 0,5 mM TCEP, 40 mM imidazolo) ir baltymai eliuojami nuo dervos dviejose frakcijose naudojant eliacijos buferį (50 mM HEPES pH 7,4, 5% glicerolio, 300 mM NaCl, 0,5 mM TCEP, 250 mM imidazolo). Abi frakcijos buvo sumaišytos ir perkeltos į 96 šulinėlių nudruskinimo lėkštę (Thermo Scientific, JAV, Zeba Spin Desalting Plate, 7K MWCO) ekvilibruotą su buferiu (50 mM HEPES pH 7,4, 5% glicerolio, 300 mM NaCl, 0,5 mM TCEP). Lėkštelė centrifuguojama 1000 x g 1 min ir surinkta baltymų frakcija analizuojama SDS-PAGE gelyje dažytame Kumasi (Coomassie) dažu. Tirpūs baltymai buvo naudojami tolimesniam charakterizavimui. Malato dehidrogenazės aktyvumo patikrinimui, dalis išgrynintų baltymų buvo sumaišoma su reakcijos mišiniu turinčiu 0,15 mM NADH, 0,2 mM oksaloacto rūgšties, 20 mM HEPES buferio (pH 7,4). Galutinis reakcijos tūris buvo 100 µl, reakcija atlikta kambario temperatūroje UV peršviečiamoje 96 šulinėlių pusės ploto lėkštelėje (UV-Star Microplate, Greiner, Austrija). Aktyvumas buvo stebimas po tris kartus sekant NADH oksidaciją į NAD⁺ matuojant 340 nm sugertį kas 30 s. per 15 min. laiko tarpą naudojant BMG Labtech SPECTROstar Nano spektrofotometrą. Nespecifinė oksidacija buvo stebima kontrolėje be substrato ir šios sugerties vertės atimamos iš tikslinio mėginio. LC-MS/MS kvantifikavimas atliktas pasirinktiems aktyviems fermentams. Aktyvumo tyrimas atliktas būdu apibrėžtu aukščiau naudojant skirtingų koncentracijų tikslius fermentus tarp 10 ir 250 nM. Reakcijos nutrauktos po 45 min praskiedžiant reakcijos mėginius vandenyje iki 1 µg/ml pradinės oksalo acetato koncentracijos.

Chromatografiniam atskyrimui buvo naudojami „Zorbax Eclipse Plus C18“ 50 mm × 2,1 mm × 1,8 µm (Agilent) su Nexera serijos HPLC (Shimadzu). Judriąją A fazę sudarė H₂O (MiliQ HPLC rūšis) su 0,1% skruzdžių rūgštimi (Sigma); mobiliąją B fazę buvo metanolis (Sigma) su 0,1% skruzdžių rūgštimi (Sigma). Orkaitės temperatūra buvo 40°C. Chromatografinis gradientas buvo nuosekliai didinamas nuo 0% iki 100%, sulaikomas, sumažinamas nuo 100% iki 0% ir išlaikomas atitinkamai per 60 s., 30 s., 30 s. ir 30 s. Automatinio mėginio ėmimo temperatūra buvo 15°C, o įpurškimo tūris buvo 0,5 µl, įpurškiant visą kilpą. MS kiekybiniam įvertinimui buvo naudojama QTRAP® 6500 sistema (Sciex), veikianti neigiamu režimu, naudojant daugiareakcijos stebėjimo (MRM) parametrus, optimizuotus obuolių rūgščiai, remiantis paskelbtais parametrais (McCloskey ir Ubhi 2014). Elektros purškimo jonizacijos parametrai buvo optimizuoti esant 0,8 ml / min. Srautui ir buvo tokie: elektropurkštuvo įtampa –4500 V, temperatūra 500°C, užuolaidų dujos 40, CAD dujos

nustatytos vidutinei, o 1 ir 2 dujos - 50 ir 50, atitinkamai, psi. Priemonė buvo masiškai kalibruojama naudojant polipropilenglikolio (PPG) standartų mišinį. Rezultatų analizei ir kiekybiniam įvertinimui naudota programinė įranga Analyst 1.7 (Sciex) ir MultiQuant 3 (Sciex).

Dešimt iš šių 40 baltymų variantų (25%) buvo ekspresuojami aukštu lygiu ir jų buvo tirpioje frakcijoje po ląstelių lizės, tai rodo baltymo sulankstytą konformaciją. Tai iš tiesų didelis pasisekimo procentas, turint omenyje tai, kad sisteminiuose tyrimuose net ekspresuojant natūralius fermentus *E. coli*, tirpių fermentų frakcija gali būti tik 20% (Huang ir kt., 2015; Bastard ir kt., 2017). 10 tirpių baltymų buvo išgryninti afininės chromatografijos būdu ir įvertintas malato dehidrogenazės aktyvumas, fluorescencingai stebint NADH vartojimą. 8 iš 10 (80%) tirpių fermentų, įskaitant variantą su 45 aminorūgščių pakaitomis, parodė tvirtą katalitinį aktyvumą (SEQ ID Nr. 9 - SEQ ID Nr. 16, 18 pav.) Su panašia kinetika, kaip laukinio tipo sekos (SEQ ID Nr. 17 ir SEQ ID Nr. 18, 18 pav.). Norėdami patvirtinti reakcijos specifiškumą, mes stebėjome produkto susidarymą naudodami LC-MS/MS, veikiantį pasirinktu reakcijos stebėjimo režimu. Mes patvirtinome oksalacetato virtimą malatu (SEQ ID Nr. 9 - SEQ ID Nr. 16, 19 pav.) su panašiais reakcijos rezultatais, kaip laukinio tipo MDH analogai (SEQ ID Nr. 17 - SEQ ID Nr. 18, 19 pav.).

Apibendrinant, mūsų pateiktas eksperimentinis pavyzdys rodo, kad mūsų daugiapakopis funkcinų baltymų sekų generavimo būdas užtikrintai atspindi daugybę natūralių baltymų savybių, tokių kaip sekų motyvai, pozicijai būdinga aminorūgščių sudėtis ir ilgo diapazono aminorūgščių sąveika, kartu užtikrindamas kataliziškai aktyvių, funkcinų ir skirtingų sekų generavimą. Eksperimentiškai patvirtinome tvirtą fermentinį aktyvumą 80% tirpių generuotų fermentų. Taigi išrastas būdas įgalina didelius šuolius į neištyrinėtus sekų skyrius ir leidžia tirti labai įvairius naujus funkcinus baltymus, atsižvelgiant į išmokus biologinius fermentų šeimų apribojimus, ekonomiškai ir efektyviai naudojant išteklius. Panaudojant šį virtualų būdą galima greitai ir efektyviai rasti naujus veiksmingus fermentus, o tai sumažina resursų ir laiko panaudojimą būsimuose gyvuose eksperimentuose.

LITERATŪRA

1. Alley, Ethan C., Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. n.d. "Unified Rational Protein Engineering with Sequence-Only Deep Representation Learning." <https://doi.org/10.1101/589333>.
2. Arjovsky, Martin, Soumith Chintala, and Léon Bottou. 2017. "Wasserstein GAN." <http://arxiv.org/abs/1701.07875>.
3. Axe, Douglas D. 2004. "Estimating the Prevalence of Protein Sequences Adopting Functional Enzyme Folds." *Journal of Molecular Biology* 341 (5): 1295–1315.
4. Axe, Douglas D., Nicholas W. Foster, and Alan R. Fersht. 1998. "A Search for Single Substitutions That Eliminate Enzymatic Function in a Bacterial Ribonuclease†." *Biochemistry*. <https://doi.org/10.1021/bi9804028>.
5. Bastard, Karine, Alain Perret, Aline Mariage, Thomas Bessonnet, Agnès Pinet-Turpault, Jean-Louis Petit, Ekaterina Darii, et al. 2017. "Parallel Evolution of Non-Homologous Isofunctional Enzymes in Methionine Biosynthesis." *Nature Chemical Biology* 13 (8): 858–66.
6. Bloom, Jesse D., Sy T. Labthavikul, Christopher R. Otey, and Frances H. Arnold. 2006. "Protein Stability Promotes Evolvability." *Proceedings of the National Academy of Sciences of the United States of America* 103 (15): 5869–74.
7. Bloom, Jesse D., Jonathan J. Silberg, Claus O. Wilke, D. Allan Drummond, Christoph Adami, and Frances H. Arnold. 2005. "Thermodynamic Prediction of Protein Neutrality." *Proceedings of the National Academy of Sciences of the United States of America* 102 (3): 606–11.
8. Cramer, A., S. A. Raillard, E. Bermudez, and W. P. Stemmer. 1998. "DNA Shuffling of a Family of Genes from Diverse Species Accelerates Directed Evolution." *Nature* 391 (6664): 288–91.
9. Dawson, Natalie L., Tony E. Lewis, Sayoni Das, Jonathan G. Lees, David Lee, Paul Ashford, Christine A. Orengo, and Ian Sillitoe. 2017. "CATH: An Expanded Resource to Predict Protein Function through Structure and Sequence." *Nucleic Acids Research* 45 (D1): D289–95.
10. Drummond, D. Allan, Jonathan J. Silberg, Michelle M. Meyer, Claus O. Wilke, and Frances H. Arnold. 2005. "On the Conservative Nature of Intragenic Recombination." *Proceedings of the National Academy of Sciences of the United States of America* 102 (15): 5380–85.

11. Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Networks." <http://arxiv.org/abs/1406.2661>.
12. Graves, Alex. 2013. "Generating Sequences With Recurrent Neural Networks." <http://arxiv.org/abs/1308.0850>.
13. Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. "Improved Training of Wasserstein GANs." <http://arxiv.org/abs/1704.00028>.
14. Guo, H. H., J. Choe, and L. A. Loeb. 2004a. "Protein Tolerance to Random Amino Acid Change." *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.0403255101>.
15. ———. 2004b. "Protein Tolerance to Random Amino Acid Change." *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.0403255101>.
16. Hansson, Lars O., Robyn Bolton-Grob, Tahereh Massoud, and Bengt Mannervik. 1999. "Evolution of Differential Substrate Specificities in Mu Class Glutathione Transferases Probed by DNA Shuffling 1 Edited by R. Huber." *Journal of Molecular Biology*. <https://doi.org/10.1006/jmbi.1999.2607>.
17. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Deep Residual Learning for Image Recognition." <http://arxiv.org/abs/1512.03385>.
18. Huang, Hua, Chetanya Pandya, Chunliang Liu, Nawar F. Al-Obaidi, Min Wang, Li Zheng, Sarah Toews Keating, et al. 2015. "Panoramic View of a Superfamily of Phosphatases through Substrate Profiling." *Proceedings of the National Academy of Sciences of the United States of America* 112 (16): E1974–83.
19. Ioffe, Sergey, and Christian Szegedy. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." <http://arxiv.org/abs/1502.03167>.
20. Jang, Eric, Shixiang Gu, and Ben Poole. 2016. "Categorical Reparameterization with Gumbel-Softmax." <http://arxiv.org/abs/1611.01144>.
21. Jolicoeur-Martineau, Alexia. 2018. "GANs beyond Divergence Minimization." <http://arxiv.org/abs/1809.02145>.
22. Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. "Progressive Growing of GANs for Improved Quality, Stability, and Variation." <http://arxiv.org/abs/1710.10196>.

23. Keefe, A. D., and J. W. Szostak. 2001. "Functional Proteins from a Random-Sequence Library." *Nature* 410 (6829): 715–18.
24. Kiefer, J., and J. Wolfowitz. 1952. "Stochastic Estimation of the Maximum of a Regression Function." *Annals of Mathematical Statistics* 23 (3): 462–66.
25. Kingma, Diederik P., and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." <http://arxiv.org/abs/1412.6980>.
26. Lim, Jae Hyun, and Jong Chul Ye. 2017. "Geometric GAN." <http://arxiv.org/abs/1705.02894>.
27. Maas, Andrew L. 2013. "Rectifier Nonlinearities Improve Neural Network Acoustic Models." <https://pdfs.semanticscholar.org/367f/2c63a6f6a10b3b64b8729d601e69337ee3cc.pdf>.
28. Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research: JMLR* 9 (Nov): 2579–2605.
29. McCloskey, Douglas, and Baljit K. Ubhi. 2014. "Quantitative and Qualitative Metabolomics for the Investigation of Intracellular Metabolism." *SCIEX Tech Note*, 1–11.
30. Mescheder, Lars, Andreas Geiger, and Sebastian Nowozin. 2018. "Which Training Methods for GANs Do Actually Converge?" <http://arxiv.org/abs/1801.04406>.
31. Miyato, Takeru, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. "Spectral Normalization for Generative Adversarial Networks." <http://arxiv.org/abs/1802.05957>.
32. Nair, Vinod, and Geoffrey E. Hinton. 2010. "Rectified Linear Units Improve Restricted Boltzmann Machines." In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 807–14. Omnipress.
33. Rennell, D., S. E. Bouvier, L. W. Hardy, and A. R. Poteete. 1991. "Systematic Mutation of Bacteriophage T4 Lysozyme." *Journal of Molecular Biology* 222 (1): 67–88.
34. Riesselman, Adam J., John B. Ingraham, and Debora S. Marks. 2018. "Deep Generative Models of Genetic Variation Capture the Effects of Mutations." *Nature Methods* 15 (10): 816–22.
35. Rockah-Shmuel, Liat, Ágnes Tóth-Petróczy, and Dan S. Tawfik. 2015. "Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations." *PLoS Computational Biology* 11 (8): e1004421.
36. Romero, Philip A., and Frances H. Arnold. 2009a. "Exploring Protein Fitness Landscapes by Directed Evolution." *Nature Reviews. Molecular Cell Biology* 10 (12): 866–76.

37. ———. 2009b. "Exploring Protein Fitness Landscapes by Directed Evolution." *Nature Reviews. Molecular Cell Biology* 10 (12): 866–76.
38. Romero, Philip A., Andreas Krause, and Frances H. Arnold. 2013. "Navigating the Protein Fitness Landscape with Gaussian Processes." *Proceedings of the National Academy of Sciences of the United States of America* 110 (3): E193–201.
39. Sarkisyan, Karen S., Dmitry A. Bolotin, Margarita V. Meer, Dinara R. Usmanova, Alexander S. Mishin, George V. Sharonov, Dmitry N. Ivankov, et al. 2016. "Local Fitness Landscape of the Green Fluorescent Protein." *Nature* 533 (7603): 397–401.
40. Shafikhani, S., R. A. Siegel, E. Ferrari, and V. Schellenberger. 1997. "Generation of Large Libraries of Random Mutants in *Bacillus Subtilis* by PCR-Based Plasmid Multimerization." *BioTechniques* 23 (2): 304–10.
41. Shi, Wenzhe, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network." <http://arxiv.org/abs/1609.05158>.
42. Steinegger, Martin, and Johannes Söding. 2017. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology* 35 (11): 1026–28.
43. Taverna, Darin M., and Richard A. Goldstein. 2002. "Why Are Proteins Marginally Stable?" *Proteins* 46 (1): 105–9.
44. Tran, Dustin, Rajesh Ranganath, and David M. Blei. 2017. "Hierarchical Implicit Models and Likelihood-Free Variational Inference." <http://arxiv.org/abs/1702.08896>.
45. Ulyanov, Dmitry, Andrea Vedaldi, and Victor Lempitsky. 2016. "Instance Normalization: The Missing Ingredient for Fast Stylization." <http://arxiv.org/abs/1607.08022>.
46. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." <http://arxiv.org/abs/1706.03762>.
47. Voigt, Christopher A., Carlos Martinez, Zhen-Gang Wang, Stephen L. Mayo, and Frances H. Arnold. 2002. "Protein Building Blocks Preserved by Recombination." *Nature Structural Biology* 9 (7): 553–58.
48. Williams, Ronald J. 1992. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning." *Machine Learning* 8 (3-4): 229–56.
49. Yang, Kevin K., Zachary Wu, and Frances H. Arnold. 2019. "Machine-Learning-

Guided Directed Evolution for Protein Engineering." *Nature Methods* 16 (8): 687–94.

50. Zhang, Han, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018. "Self-Attention Generative Adversarial Networks." <http://arxiv.org/abs/1805.08318>.

51. WO2019097014

APIBRĖŽTIS

1. Funkcinių sintetinių baltymų sekų gamybos būdas besiskiriantis tuo, kad apima šiuos žingsnius:

a) apytikslų sintetinių sekų, kurios bus gaminamos, erdvės ribų apibrėžimas, pasirenkant daugybę esamų baltymų sekų,

b) pasirinktų baltymų sekų apdorojimas,

c) nežinomo, tikrojo, iš anksto paruoštų sekų aminorūgščių pasiskirstymo aproksimavimas, naudojant generatyvinius adversarinius tinklus,

d) sintetinių baltymų sekų gavimas pagal aproksimuotą pasiskirstymą,

e) gautų baltymų sekų apdorojimas.

2. Būdas pagal 1 punktą besiskiriantis tuo, kad pagamintos funkcinės sintetinių baltymų sekos yra fermentai.

3. Būdas pagal 1 punktą besiskiriantis tuo, kad išankstinis pasirinktų baltymų sekų apdorojimas apima sekų filtravimą pagal jų biologines savybes.

4. Būdas pagal 1 punktą besiskiriantis tuo, kad save stebintys sluoksniai yra įtraukti į generatyvinio adversarinio tinklo architektūrą.

5. Būdas pagal 1 punktą besiskiriantis tuo, kad išsiplėtęs konvoliucinis sluoksnis yra įtrauktas į generatyvinio adversarinio tinklo architektūrą.

6. Būdas pagal 1 punktą besiskiriantis tuo, kad generatyviniai adversarinio tinklo sluoksniai yra normalizuojami naudojant spektrinį normalizavimą.

7. Būdas pagal 1 punktą besiskiriantis tuo, kad generatyvinio adversarinio tinklo mokymo metu dinamiškai paimami nepakankamai atstovaujamų treniruočių sekų klasteriai.

8. Būdas pagal 1 punktą besiskiriantis tuo, kad papildoma informacija yra teikiama į diskriminatoriaus ir generatoriaus tinklus.

9. Būdas pagal 1 punktą besiskiriantis tuo, kad aminorūgštys yra koduojamos naudojant „one-hot“ kodavimą.

10. Būdas pagal 9 punktą besiskiriantis tuo, kad generatoriaus tinklas sukuria „one-hot“ koduotą išėjimą, naudojant diferencijuotą diskretinę aproksimaciją.

11. Būdas pagal 1 punktą besiskiriantis tuo, kad aminorūgštys yra užkoduotos naudojant įterpimus.

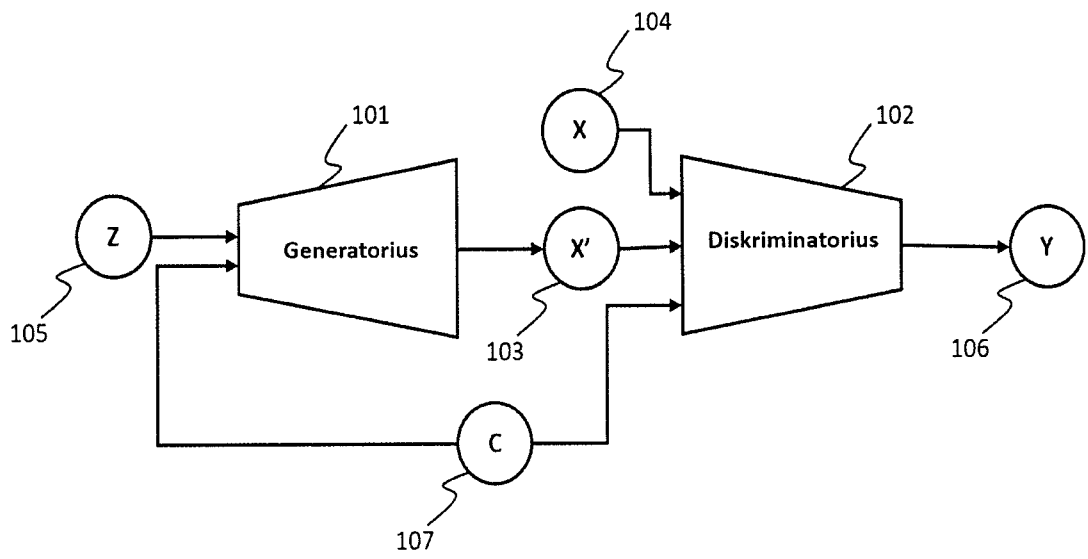
12. Būdas pagal 1 punktą besiskiriantis tuo, kad gautų sintetinių baltymų sekų

apdorojimas apima sekų filtravimą pagal balą, kurį nustato diskriminatoriaus tinklas.

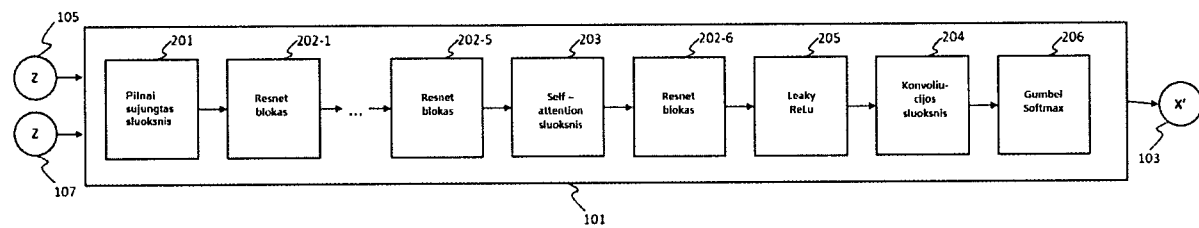
13. Būdas pagal 1 punktą besiskiriantis tuo, kad gautų sintetinių baltymų sekų apdorojimas apima sekų filtravimą, pritaikant jas mašininio mokymosi modeliams.

14. Funkcinių baltymų sekų, gautų būdu pagal 1 punktą, panaudojimas eksperimentiniam baltymų patikrinimui.

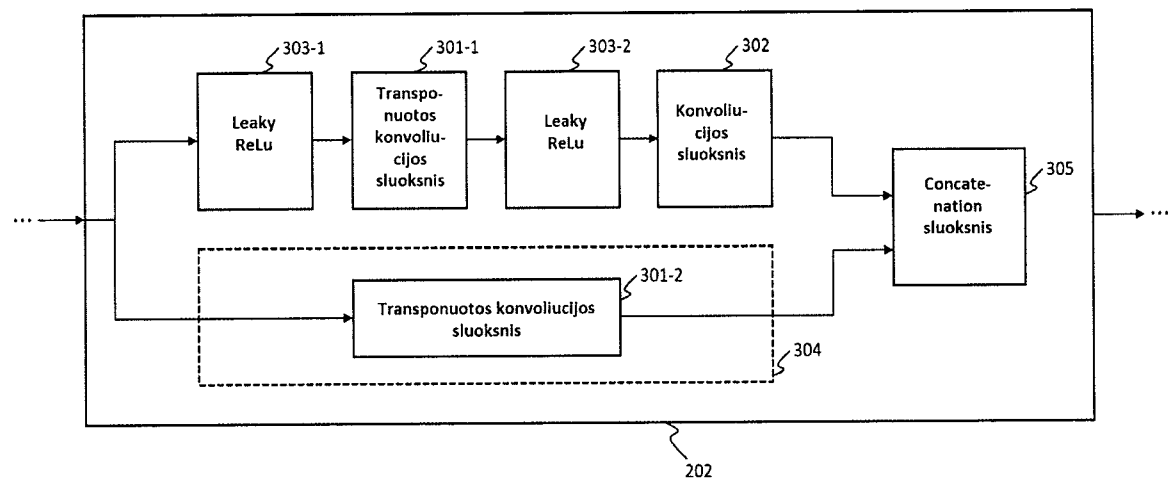
15. Funkcinių baltymų sekų, gautų būdu pagal 1 punktą, panaudojimas duomenų padidinimui.



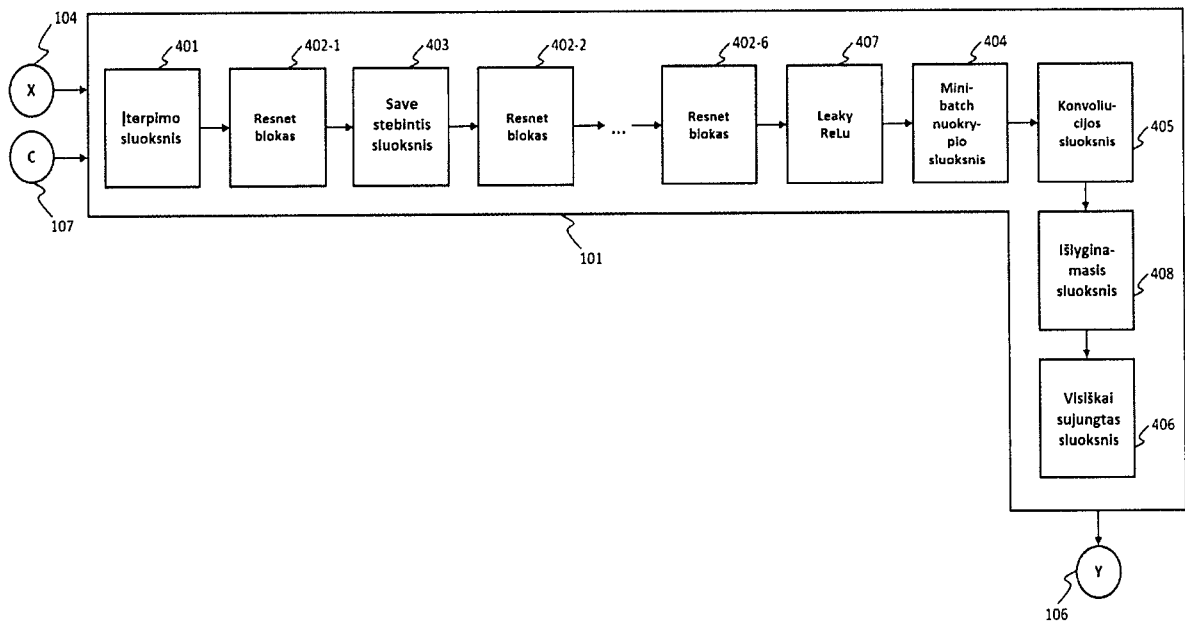
1 pav.



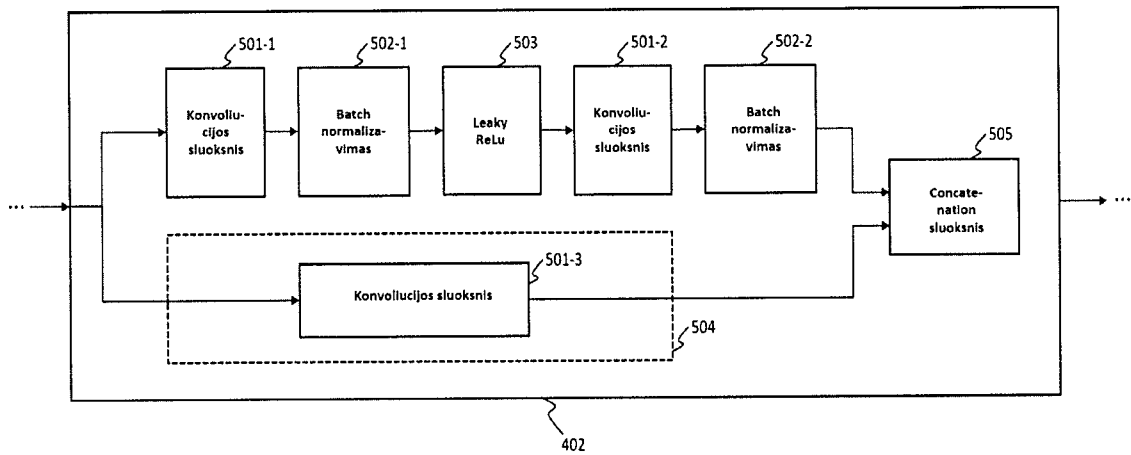
2 pav.



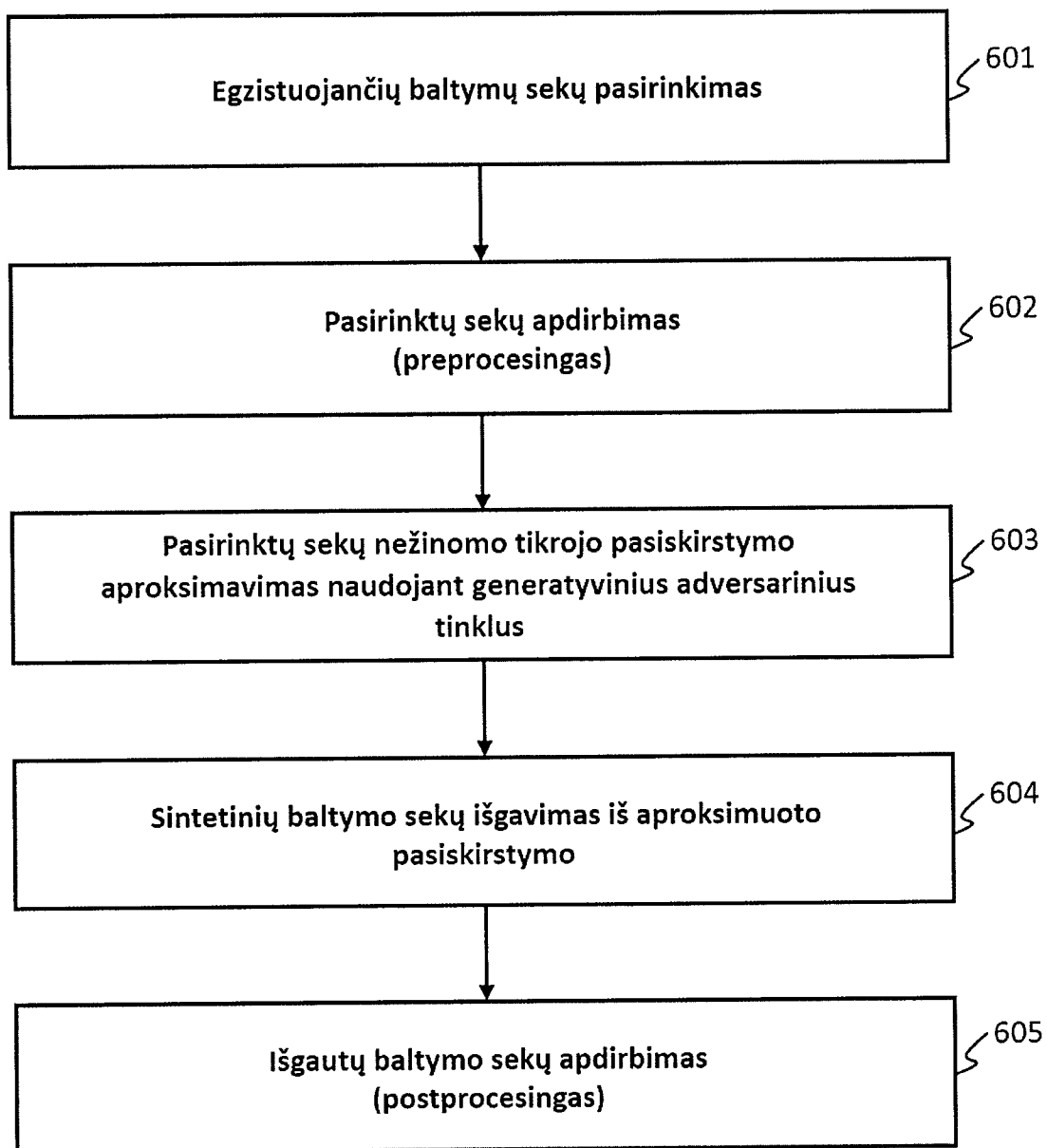
3 pav.



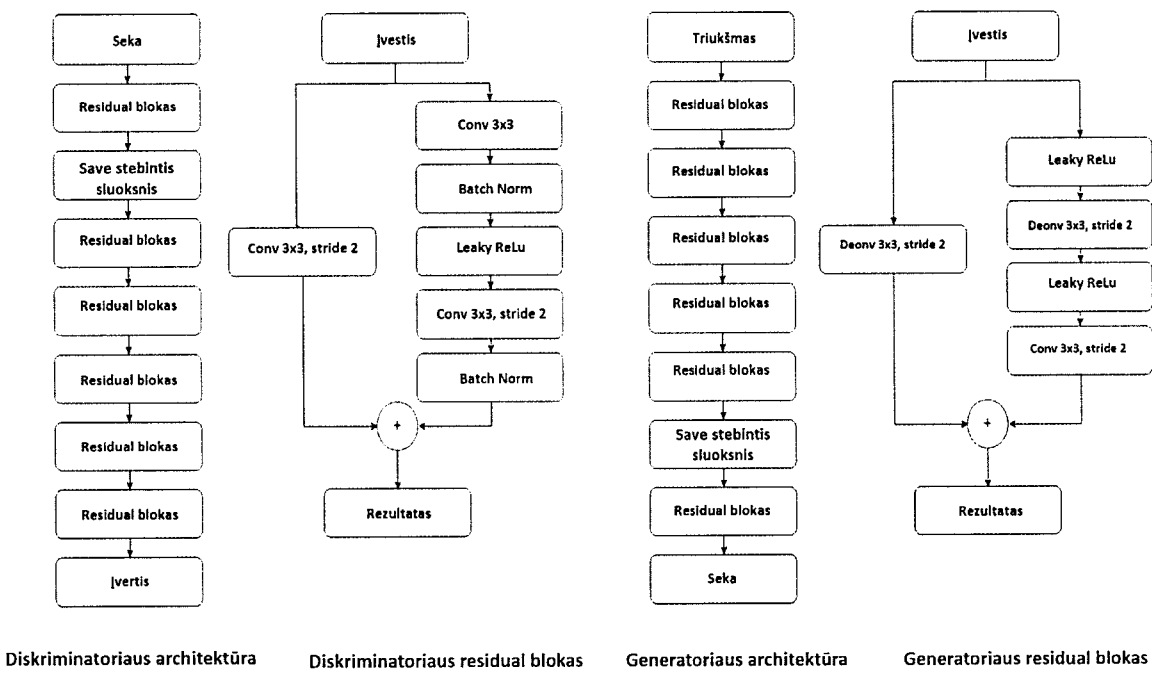
4 pav.



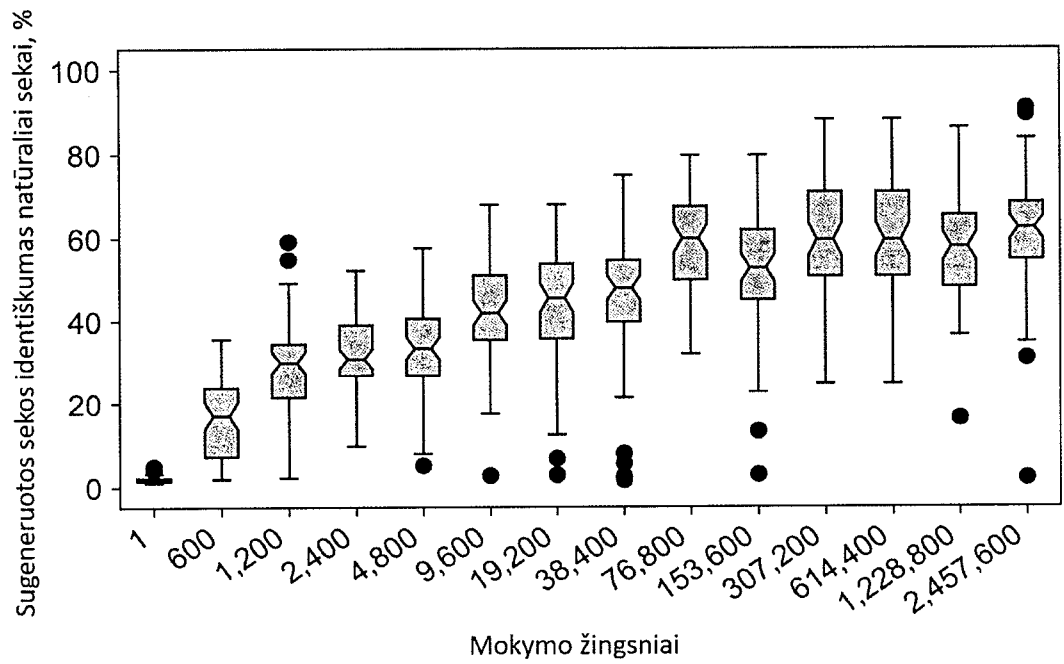
5 pav.



6 pav.

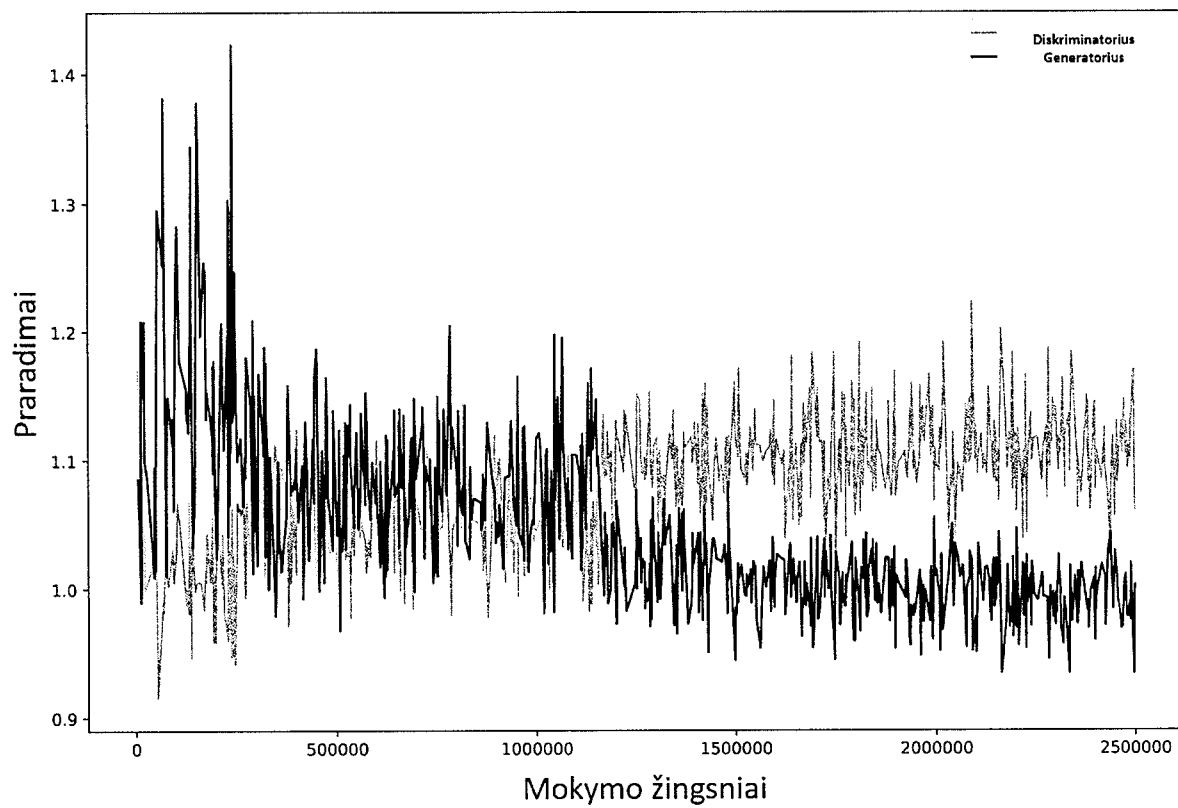


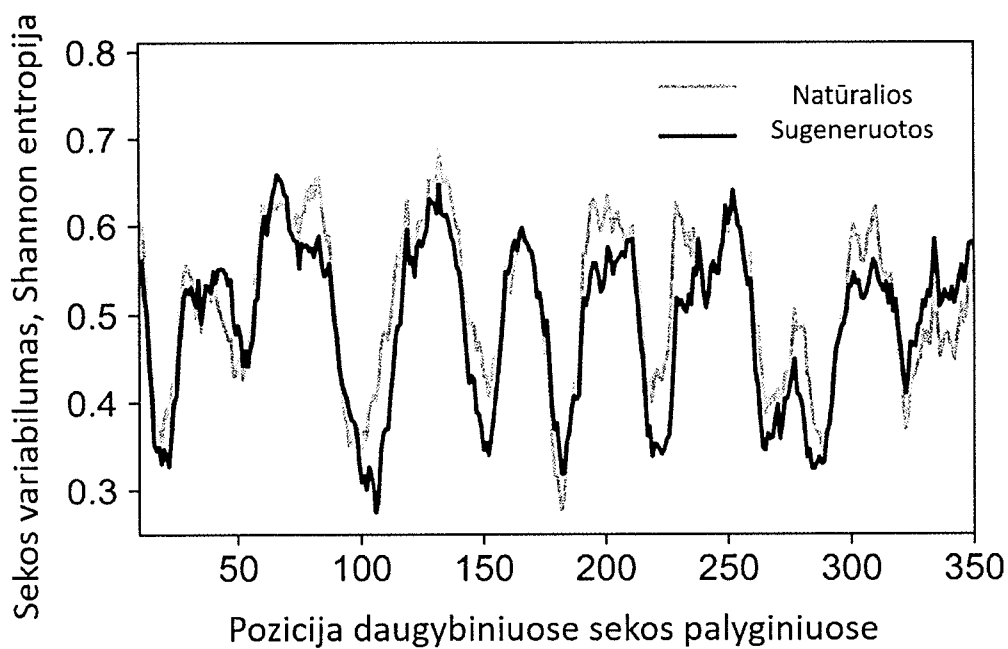
7 pav.



8 pav.

Praradimai naudojant diskriminatorių arba generatorių

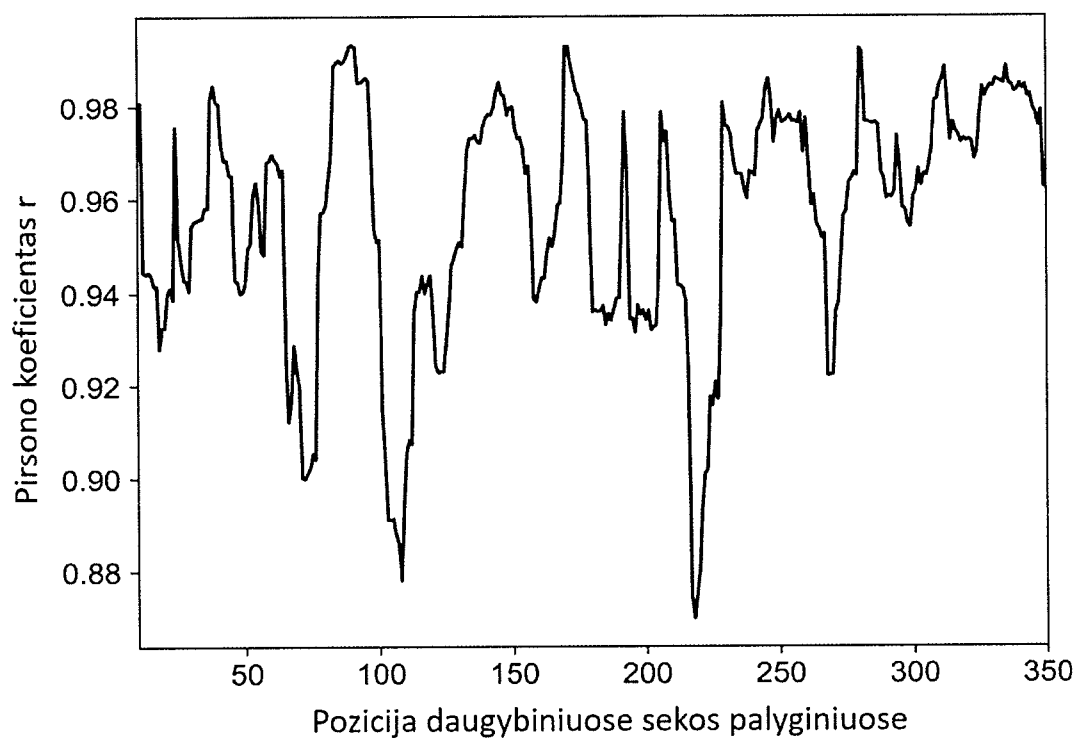




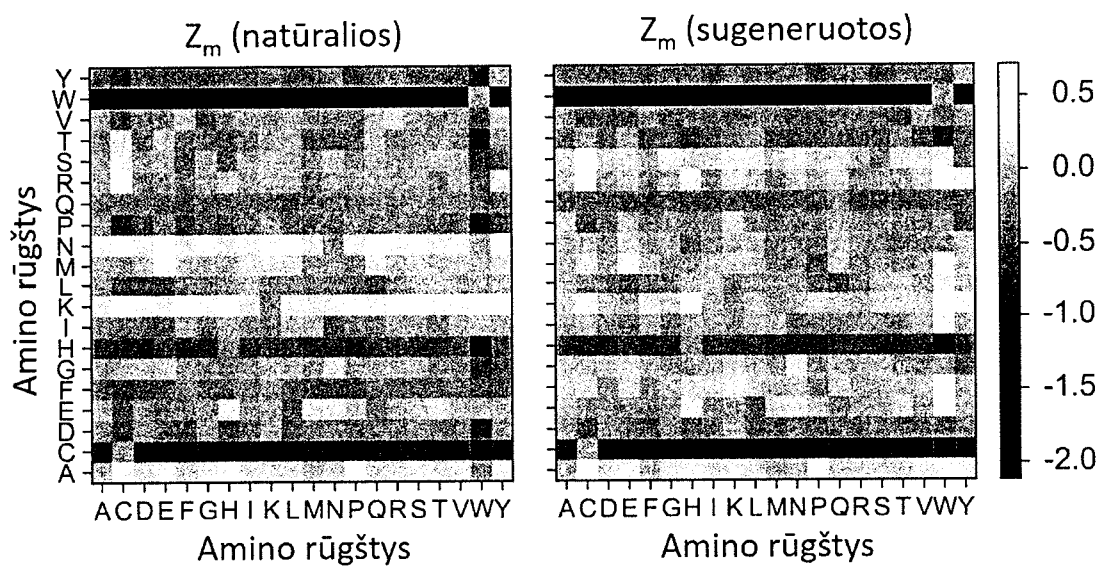
10 pav.



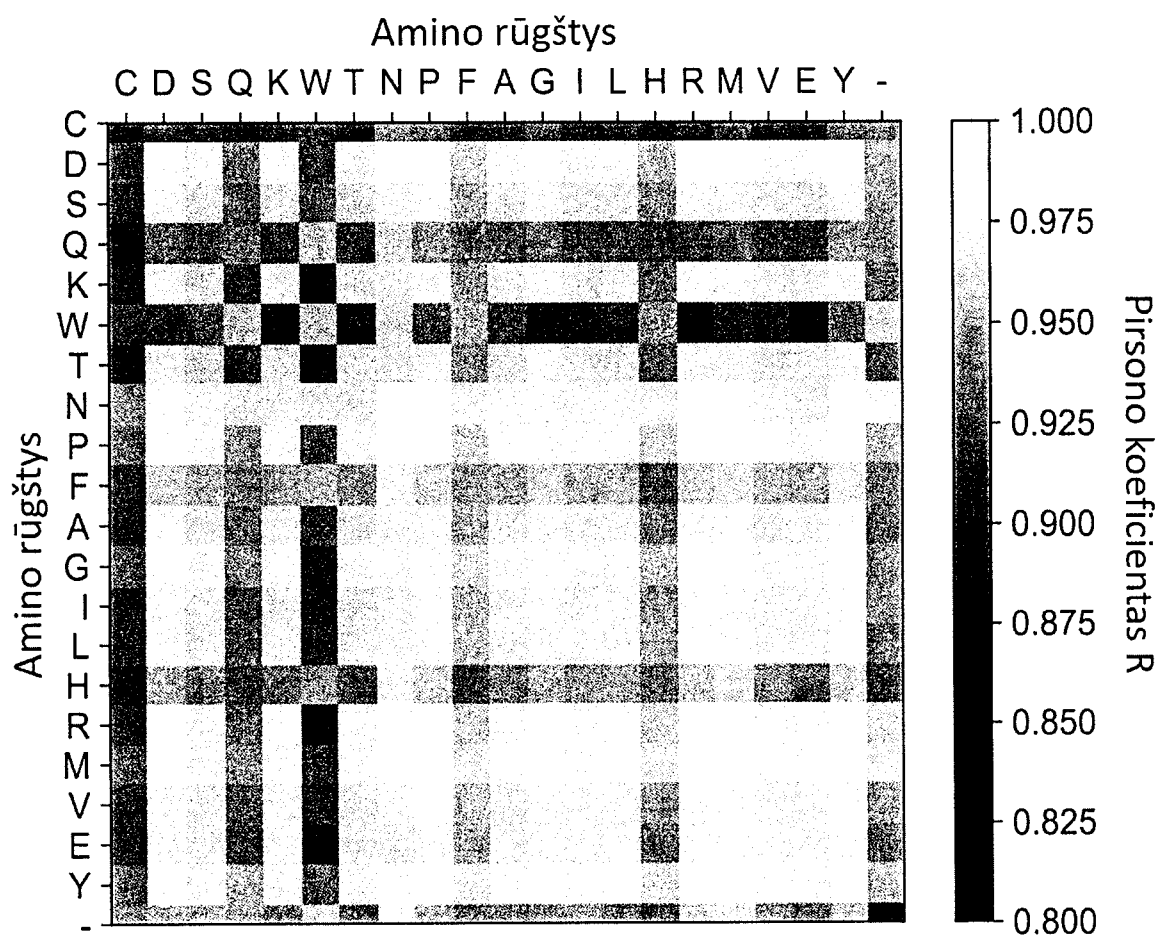
11 pav.

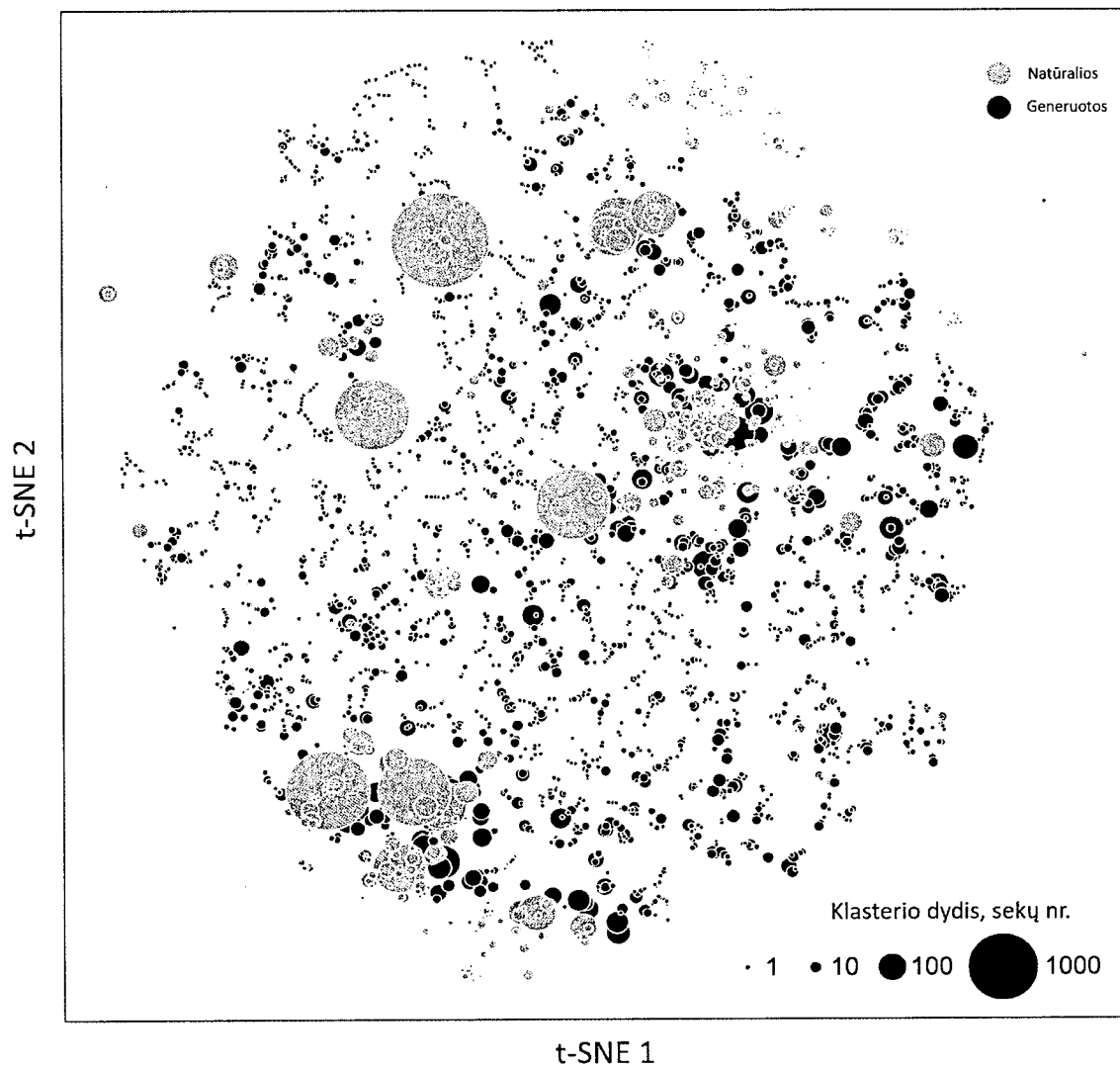


12 pav.

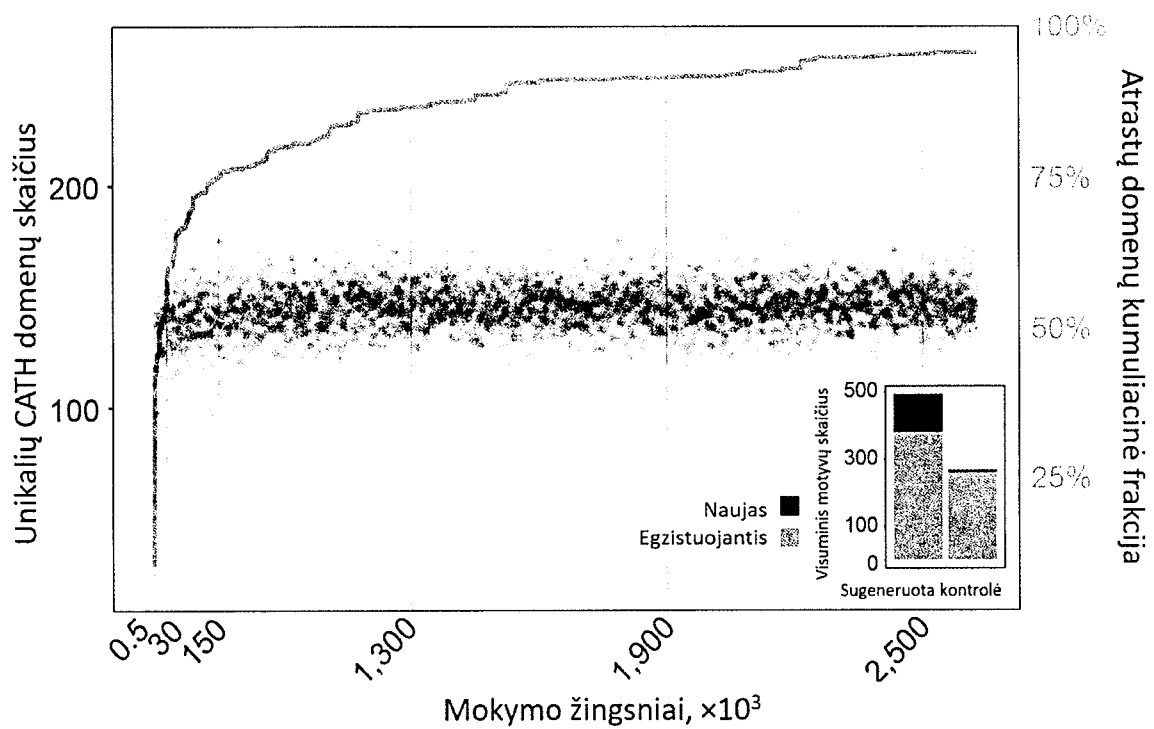


13 pav.

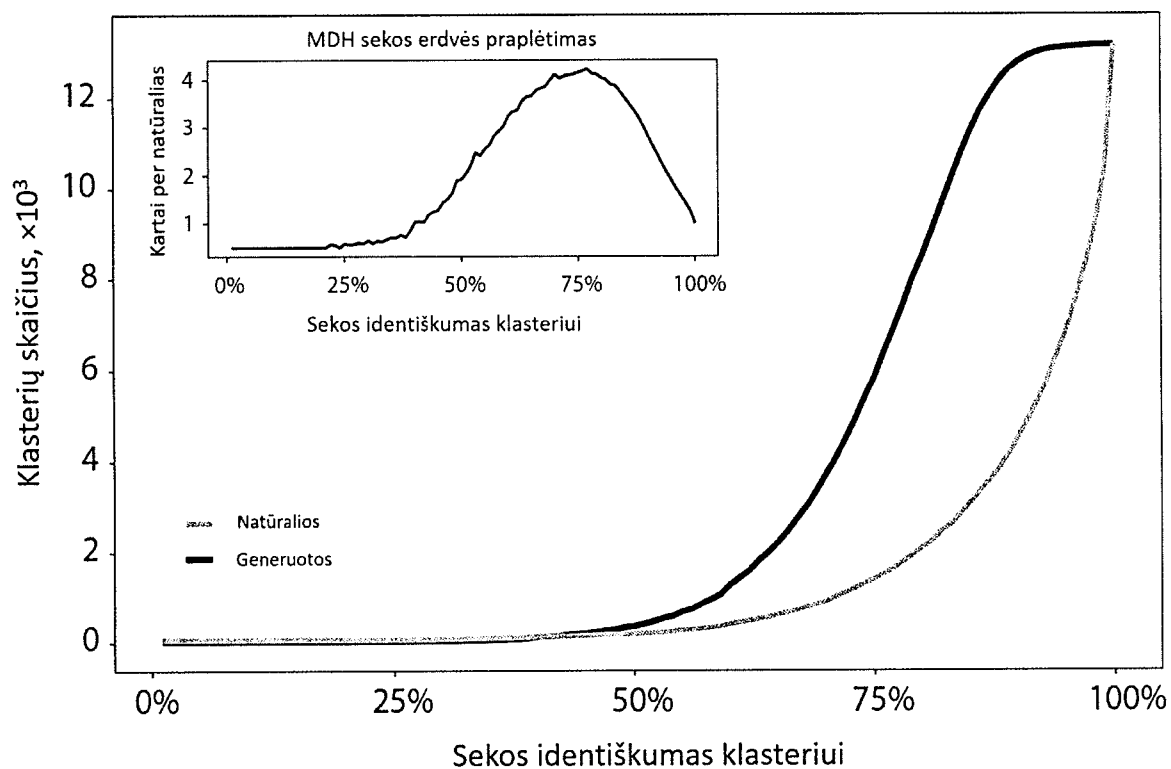




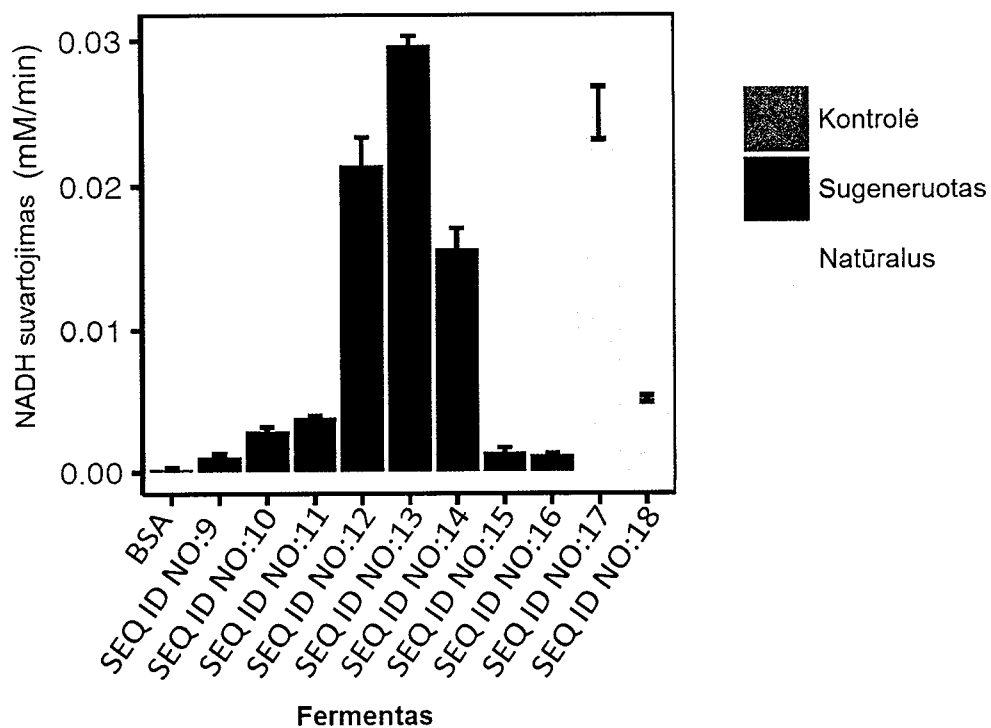
15 pav.



16 pav.



17 pav.



18 pav.

