VILNIAUS UNIVERSITETAS

KLIMENT OLECHNOVIČ

BALTYMŲ IR NUKLEORŪGŠČIŲ ERDVINIŲ STRUKTŪRŲ ANALIZĖS IR VERTINIMO METODAI: KŪRIMAS IR TAIKYMAS

Daktaro disertacijos santrauka
Fiziniai mokslai, informatika (09P)

Vilnius, 2017

Disertacija rengta 2012–2016 metais Vilniaus universiteto Biotechnologijos instituto Bioinformatikos skyriuje.

Mokslinis vadovas:
dr. Česlovas Venclovas (Vilniaus universitetas, biomedicinos mokslai, biologija — 01B).

Mokslinis konsultantas:
prof. habil. dr. Feliksas Ivanauskas (Vilniaus universitetas, fiziniai mokslai, informatika — 09P).

Disertacija ginama Vilniaus universiteto Informatikos mokslo krypties taryboje:

Pirmininkas:
prof. habil. dr. Antanas Žilinskas (Vilniaus universitetas, fiziniai mokslai, informatika — 09P).

Nariai:
prof. dr. Romas Baronas (Vilniaus universitetas, fiziniai mokslai, informatika — 09P),
dr. Sergio Bordel Velasco (Lietuvos sveikatos mokslų universitetas, fiziniai mokslai, informatika — 09P),
prof. dr. Saulius Gražulis (Vilniaus universitetas, fiziniai mokslai, biochemija — 04P),
dr. Andriy Kryshtafovych (Kalifornijos universitetas Devise, fiziniai mokslai, informatika — 09P).

Disertacija bus ginama viešame Informatikos mokslo krypties tarybos posėdyje 2017 m. birželio mėn. 16 d. 13 val. Matematikos ir informatikos fakultete, 211 auditorijoje.
Adresas: Didlaukio g. 47, LT-08303 Vilnius.

Disertacijos santrauka išsiuntinėta 2017 m. gegužės 16 d.
Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir Vilniaus universiteto svetainėje adresu:
http://www.vu.lt/lt/naujienos/ivykiu-kalendorius.

VILNIUS UNIVERSITY

KLIMENT OLECHNOVIČ

METHODS FOR THE ANALYSIS AND ASSESSMENT OF THE THREE-DIMENSIONAL STRUCTURES OF PROTEINS AND NUCLEIC ACIDS: DEVELOPMENT AND APPLICATIONS

Summary of doctoral dissertation
Physical sciences, informatics (09P)

Vilnius, 2017

The dissertation work was carried out at the Department of Bioinformatics, Institute of Biotechnology, Vilnius University from 2012 to 2016.

Scientific supervisor:
Dr. Česlovas Venclovas (Vilnius University, Biomedical Sciences, Biology — 01B).

Scientific consultant:
Prof. Dr. Habil. Feliksas Ivanauskas (Vilnius University, Physical Sciences, Informatics — 09P).

The dissertation will be defended at the Counsil of Scientific Field of Informatics at Vilnius University:

Chairman:
Prof. Dr. Habil. Antanas Žilinskas (Vilnius University, Physical Sciences, Informatics — 09P).

Members:
Prof. Dr. Romas Baronas (Vilnius University, Physical Sciences, Informatics — 09P),
Dr. Sergio Bordel Velasco (Lithuanian University of Health Sciences, Physical Sciences, Informatics — 09P),
Prof. Dr. Saulius Gražulis (Vilnius University, Physical Sciences, Biochemistry — 04P),
Dr. Andriy Kryshtafovych (University of California, Davis, Physical Sciences, Informatics — 09P).

The dissertation will be defended at the public meeting of the Council in the auditorium number 211 at Vilnius University Faculty of Mathematics and Informatics, on the 16th of June, 2017 at 1 PM.
Address: Didlaukio st. 47, LT-08303 Vilnius, Lithuania.

The summary of the dissertation was distributed on May 16, 2017.
The doctoral dissertacion is available at the library of Vilnius University and at VU webpage: http://www.vu.lt/lt/naujienos/ivykiu-kalendorius.

Padėka

Esu dėkingas savo vadovui Česlovui Venclovui už vadovavimą ir neįkainojamas įžvalgas apie bioinformatiką ir mokslą apskritai.

Dėkoju dabartiniams ir buvusiems bendradarbiams iš VU Biotechnologijos instituto Bioinformatikos skyriaus Justui Dapkūnui, Mindaugui Margelevičiui, Kęstučiui Timinskui, Ryčiui Dičiūnui, Albertui Timinskui, Dariui Kazlauskui, Visvaldui Kairiui, Eleonorai Kulberkytei, Nerijui Verseckui ir Mantui Marcinkui už pagalbą ir diskusijas.

Taip pat dėkoju Andriy Kryshtafovych, Jürgen Haas, Alessandro Barbato, Rimvydui Krasauskui, Kęstučiui Karčiauskui, Severinui Zubei, Janusz Bujnicki ir Andriui Merkiui už unikalias ekspertines įžvalgas. Esu dėkingas moksliniam konsultantui Feliksui Ivanauskui ir disertacijos recenzentams Romui Baronui ir Sauliui Gražuliui.

Galiausiai, dėkoju savo tėvams už supratimą ir palaikymą.

Turinys

Įvadas

Tyrimų sritis

Gyvybės pagrindą molekuliniame lygmenyje sudaro baltymai ir nukleorūgštys.[1] Norint visapusiškai suprasti šių biologinių makromolekulių veikimo mechanizmus būtina žinoti jų erdvines struktūras. Tokių struktūrų nustatymas eksperimentiniais metodais dažnai užtrunka arba iš viso nepavyksta, todėl biologinių makromolekulių modeliavimui ir analizei gyvybės mokslai pasitelkia informatikos metodus.[2] Biopolimerų erdvinių struktūrų nusakymas pagal jų sekas toli gražu nėra išspręsta problema, tačiau kai kurie sukurti metodai jau yra efektyviai taikomi praktikoje.[3] Dauguma šiuolaikinių struktūros nusakymo metodų veikia dviem etapais:

1. Sugeneruojama modelių-kandidatų aibė.
2. Išrenkamas geriausias modelis, t. y. nusakoma, kuris sugeneruotas modelis yra realistiškiausias.

Pristatoma disertacija skirta antrajai iš pirmiau paminėtų stadijų testuoti ir tobulinti; joje nagrinėjamos struktūrinių modelių analizės ir vertinimo problemos.

Tyrimų tikslai

Šio darbo tikslas yra sukurti naujus geresnius metodus šioms tarpusavyje susijusioms problemoms spręsti:

1. Molekulinių struktūrų geometrinių ypatybių analizė, galinti sudaryti pagrindą struktūriniams modeliams vertinti.
2. Struktūrinio modelio vertinimas lyginant jį su etalonu (realia struktūra).
3. Struktūrinio modelio tikslumo nusakymas, kai reali struktūra nėra žinoma, t. y. modelio vertinimas neturint etalono.

Tyrimų rezultatai, jų mokslinis naujumas ir reikšmė

Pristatoma disertacija parengta remiantis šešiais darbais, įvykdytais 2012-2016 metais ir publikuotais tarptautiniuose recenzuojamuose žurnaluose.[4–9] Didžiausias dėmesys skirtas CAD-score (angl. „Contact Area Difference Score“) metodą ir jo taikymo aspektus pristatantiems darbams.[4,6,7,9] CAD-score metodas skaitiškai įvertina struktūrinio modelio panašumą į etaloninę struktūrą. Modelių vertinimo lyginant su etalonu problema yra kritiškai svarbi testuojant ir tobulinant baltymų ir RNR struktūrų nusakymo metodus. CAD-score metodas pagrįstas tarpatominių kontaktų, gaunamų naudojant atomų Voronojaus diagramą, plotų naudojimu; tuo jis ženkliai skiriasi nuo kitų struktūrų lyginimo metodų. CAD-score turi aibę naudingų savybių, dėl kurių jis buvo teigiamai įvertintas baltymų struktūrų modeliavimo ekspertų ir tapo vienu iš standartinių metodų, naudojamų CASP (Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction)[10,11] ir CAMEO (Continuous Automated Model EvaluatiOn)[12] baltymų struktūrų modeliavimo eksperimentuose. CAD-score taip pat buvo pritaikytas efektyviai nukleorūgščių struktūrų ir įvairių makromolekulinių kompleksų struktūrų analizei.

CAD-score metode naudojami tarpatominiai kontaktai gaunami remiantis atomų Voronojaus diagrama, kur atomus atstovauja van der Valso spindulio rutuliai. Rutulių Voronojaus diagrama yra vienas iš palyginti sudėtingai skaičiuojamų Voronojaus diagramos variantų:[13] nebuvo jokių viešai prieinamų programinių įrankių, su kuriais būtų galima visiškai įgyvendinti CAD-score idėją. Todėl buvo sukurtas specializuotas algoritmas ir programinis įrankis Voronota,[5] skirtas atomų rutulių Voronojaus diagramos viršūnėms skaičiuoti. Voronota metodas prieinamas net tik kaip CAD-score dalis, bet ir kaip atskiras universalus įrankis, skirtas tiek eksperimentiškai nustatytoms, tiek nusakytoms makromolekulių struktūroms analizuoti.

Paskutinis pristatomas metodas, VoroMQA[8] (angl. „Voronoi tessellation-based Model Quality Assessment“), skirtas baltymų struktūrų modelių tikslumui nusakyti neturint etalono, t. y. nežinant atitinkamos eksperimentiškai nustatytos struktūros. VoroMQA didžiąja dalimi pagrįstas Voronota ir CAD-score metodų kūrimo metu išvystyta struktūrų analizės metodika. Naujame metode tarpatominių kontaktų ir tirpikliui prieinamų paviršių plotai naudojami apibrėžiant ir taikant empirinį (žiniomis paremtą) statistinį potencialą. Išsamūs testai su CASP eksperimentų duomenimis parodė, kad VoroMQA ypač sėkmingai

taikytinas sprendžiant geriausio modelio pasirinkimo užduotį. Be to VoroMQA gali įvertinti struktūrinio modelio kokybę absoliučiojoje skalėje ir padėti nuspręsti, ar modelis yra realistiškas.

Pagrindinis šio darbo mokslinio naujumo aspektas yra būdai tarpatominių kontaktų plotams apskaičiuoti ir panaudoti analizuojant ir vertinant biologinių makromolekulių struktūras tiek turint etaloną, tiek jo neturint.

Voronota, CAD-score ir VoroMQA metodų programiniai paketai bei CAD-score ir VoroMQA interneto serveriai yra pasiekiami toliau nurodytasi adresais:

- `http://bioinformatics.lt/software/voronota`
- `http://bioinformatics.lt/software/cad-score`
- `http://bioinformatics.lt/software/voromqa`

Ginamieji teiginiai

- Pasiūlytas rutulių Voronojaus diagramos viršūnių skaičiavimo metodas efektyviai taikytinas apdorojant makromolekulių struktūras pasinaudojant žiniomis apie dažnai pasitaikančias atomų rutulių erdvinio išsidėstymo konfigūracijas. Metodas yra paprastai lygiagretinamas. Metodas yra efektyvus įrankis tarpatominėms sąveikoms identifikuoti.
- Pasiūlytas biologinių makromolekulių struktūrų palyginimo metodas yra efektyvus sprendžiant struktūrinių modelių vertinimo esant etalonui užduotį. Taikant metodą išvengiama tradiciniams etaloninio vertinimo metodams būdingų problemų naudojant kontaktų plotus, išvestus iš atomų rutulių Voronojaus diagramos. Taikant šį metodą galima efektyviai analizuoti ir lyginti visų pagrindinių biologinių makromolekulių (baltymų, nukleorūgščių ir jų kompleksų) struktūras.
- Pasiūlytame baltymų struktūrų modelių tikslumo nusakymo neturint etalono metode efektyviai derinama empirinio statistinio potencialo idėja ir tarpatominių kontaktų plotų, gaunamų iš atomų rutulių Voronojaus diagramos, naudojimas. Metodas sistemingai sprendžia modelių kokybės vertinimo užduotis geriau negu kiti statistiniais potencialais paremti metodai.

1 Voronota: atomų rutulių Voronojaus diagramos viršūnių skaičiavimo metodas

1.1 Trumpas metodo aprašymas

3D rutulių Voronojaus diagrama ir atitinkamos Voronojaus viršūnės

Tarkime, $B = \{b_1, b_2, \ldots, b_n\}$ yra rutulių aibė, kur $b_i = \langle c_i, r_i \rangle$ yra rutulys, kurio centras $c_i \in \mathbb{R}^3$ ir spindulys $r_i \in \mathbb{R}_0^+$. Atstumas $d(p, b_i)$ nuo taško $p \in \mathbb{R}^3$ iki rutulio b_i apibrėžiamas taip:

$$d(p, b_i) = \|p - c_i\| - r_i \tag{1.1}$$

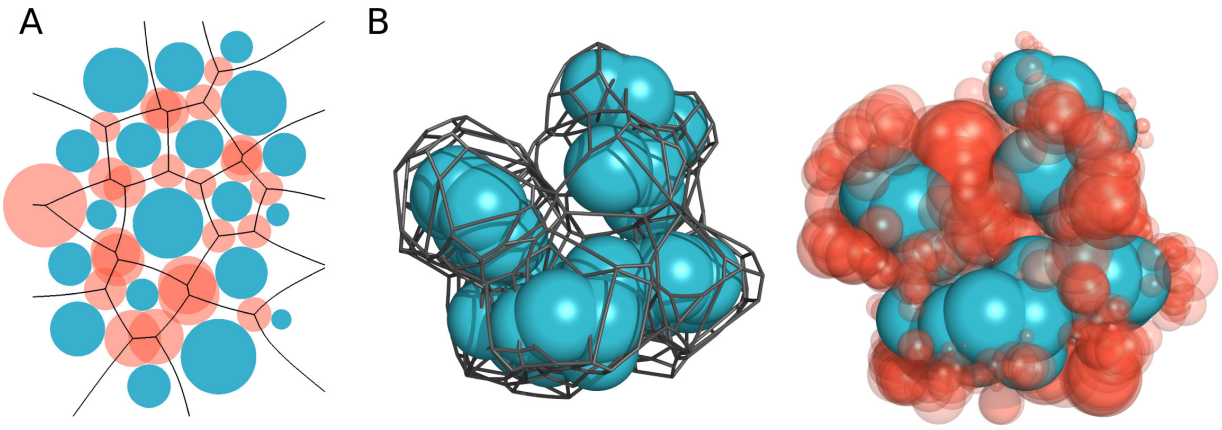
Rutulio b_i Voronojaus ląstelė[14] V_i yra erdvės regionas, apimantis visus taškus, nuo kurių atstumas iki b_i yra ne didesnis nei atstumai iki kitų rutulių:

$$V_i = \left\{p \in \mathbb{R}^3 | d(p, b_i) \leq d(p, b_j), \forall b_j \in B \setminus b_i\right\} \tag{1.2}$$

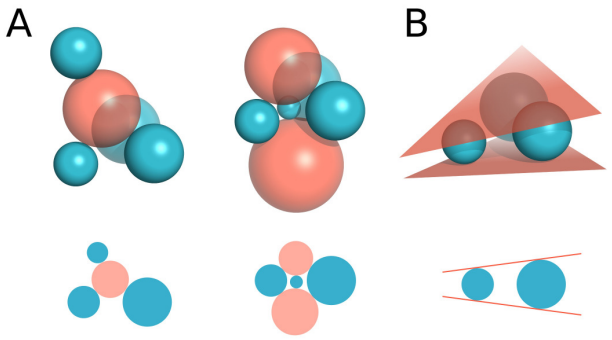
Voronojaus ląstelių aibė $\{V_1, V_2, \ldots, V_n\}$ vadinama Voronojaus diagrama.[13] 1.1 pav. pavaizduoti rutulių Voronojaus ląstelių pavyzdžiai. Du rutuliai vadinami kaimynais, jei jų Voronojaus ląstelės susikerta.

Keturių Voronojaus ląstelių susikirtimas apibrėžia tašką, vadinamą Voronojaus viršūne. Šis taškas yra keturių kaimyninių rutulių tuščiosios liestinės sferos centras (1.2 pav., A). Kai kurios Voronojaus ląstelės gali neturėti viršūnių — tokios situacijos analizuojamos atskirai.

Liestinėms sferoms konstruoti naudojame Gavrilovos ir Rokne[15] pasiūlytą metodą. Šio metodo principus naudojame ir trijų 3D rutulių liestinėms plokštumoms (1.2 pav., B) skaičiuoti. Tokios plokštumos atitinka be galo didelio spindulio liestines sferas.



Pav. 1.1: (A) 2D rutulių (pavaizduotų mėlynai) Voronojaus ląstelės ir tuščiosios liestinės sferos (pavaizduotos raudonai), atitinkančios Voronojaus viršūnes. (B) 3D rutulių Voronojaus ląstelių briaunos (kairėje) ir tuščiosios liestinės sferos, atitinkančios Voronojaus viršūnes (dešinėje).



Pav. 1.2: (A) 3D rutulių ketvertai, turintys vieną (kairėje) arba dvi (dešinėje) liestines sferas. (B) 3D rutulių trejetas, turintis dvi liestines plokštumas. Po kiekvienu 3D pavyzdžiu pateiktas atitinkamas panašus atvejis 2D erdvėje.

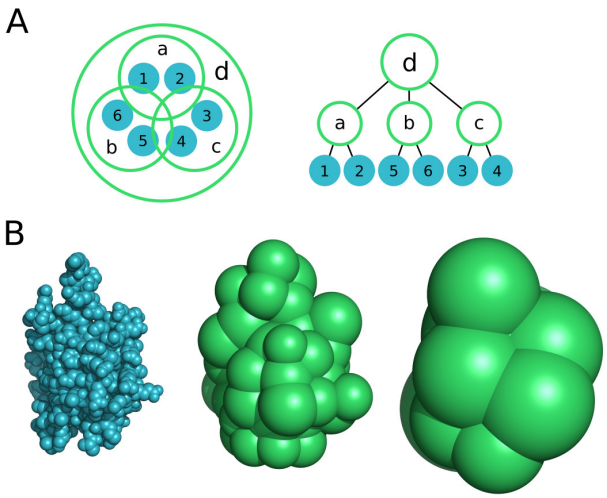
Voronojaus viršūnių paieškos algoritmo bendroji schema

Mūsų tikslas — rasti duotos rutulių aibės B rutulių ketvertus, kurie apibrėžia visas Voronojaus diagramos viršūnes. Kitaip tariant, ieškome rutulių ketvertų, kurie turi bent vieną liestinę sferą, nesikertančią su jokiais B elementais. Tokius ketvertus vadiname priimtinaisiais. Kiekvienas priimtinasis ketvertas yra trijų priimtinųjų trejetų sąjunga. Pradėję nuo vieno priimtinojo trejeto galime atrasti priimtinuosius ketvertus pridėdami priimtinuosius kaimynus prie aptiktų priimtinųjų trejetų. Toks principas, žinomas kaip „dovanos vyniojimas“ (angl. „gift wrapping“),[16] yra naudojamas taškų Delone trianguliacijai[17,18] ir rutulių kvazi-trianguliacijai[19,20] konstruoti. Mes realizuojame „dovanos vyniojimo“ strategiją atsižvelgdami į tai, kad priimtinųjų ketvertų tinklas gali būti nejungus.[21] Tam naudojame du ciklus, įdėtus vienas į kitą: išoriniame cikle ieškoma pirmo pasitaikusio priimtinojo trejeto, turinčio rutulių, nesančių jau atrastuose priimtinuosiuose ketvertuose; vidiniame cikle randami priimtinieji ketvertai pradedant nuo priimtinojo trejeto. Detalus algoritmo pseudokodas pateikiamas disertacijos pagrindiniame tekste, šioje santraukoje toliau paaiškinsime dvi jo esmines subprocedūras, skirtas pirmajam priimtinajam trejetui aptikti ir visiems priimtinojo trejeto kaimynams rasti.

Pirmojo priimtinojo trejeto paieškos procedūra

Mes darome prielaidą, kad erdvėje artimesni rutuliai yra labiau linkę suformuoti priimtinuosius ketvertus. Pradėdami nuo pirmo rutulio $b_0 \in B$ randame jo artimiausių kaimynų aibę $B_0 \subset B$. Toliau perrenkame ketvertus iš B_0 elementų. Jei ketvertas turi liestinę sferą, nesikertančią su jokiu rutuliu iš B, tai ketvertas yra priimtinasis ir jo trejetai yra grąžinami kaip rezultatas. Jei priimtinųjų ketvertų nerandama, aibė B_0 išplečiama įtraukiant daugiau b_0 kaimynų. Maksimalus B_0 dydis yra ribojamas: jei jis viršijamas, pasirenkamas kitas pradinis rutulys ir procedūra pradedama iš naujo.

Kritinis šios procedūros realizacijos aspektas yra sferų ir rutulių susikirtimų paieška. Artimiausių kaimynų paiešką irgi galima interpretuoti kaip sferos (aplink b_0) ir rutulių susikirtimų aptikimo problemą. Efektyviai susikirtimų paieškai naudojame medžio tipo duomenų struktūrą — gaubiančių sferų hierarchiją[22] (1.3 pav.). Jos konstravimas, kuris iš dalies remiasi k-d medžio erdvės dalijimo algoritmu,[23] ir panaudojimas detaliai aprašomi disertacijos pagrindini-



Pav. 1.3: (A) Gaubiančių sferų hierarchijos pavyzdys dvimatėje erdvėje. (B) Gaubiančių sferų hierarchijos pritaikymas baltymo struktūrai: kairėje pavaizduoti atomų rutuliai, viduryje ir dešinėje — pirmojo ir antrojo lygių gaubiančiosios sferos.

ame tekste.

Visų priimtinojo trejeto kaimynų paieškos procedūra

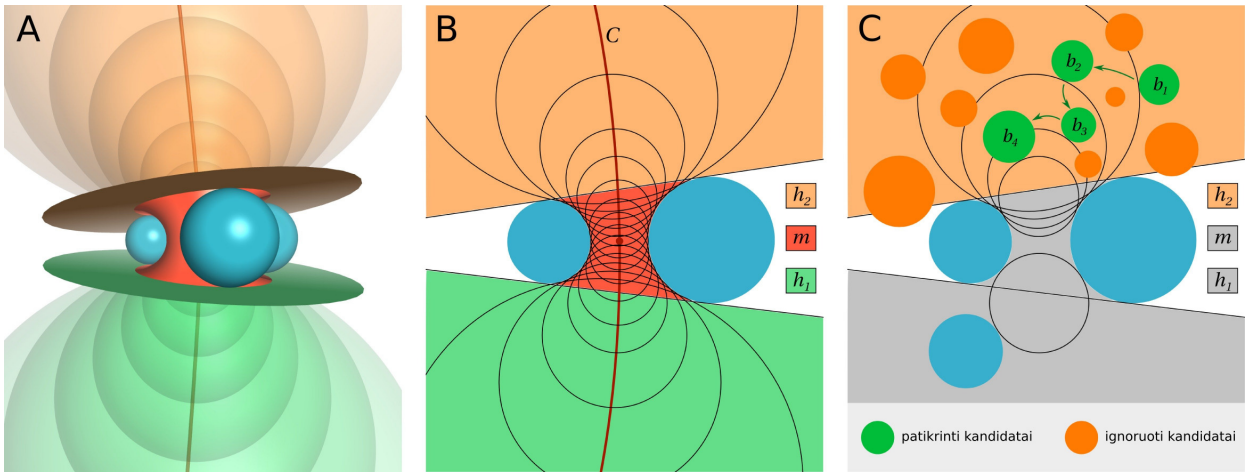
Suvaržytieji ir nesuvaržytieji trejetai

Paimkime priimtinąjį rutulių trejetą $t = \{a, b, c\} \subset B$. Bendrai t gali turėti be galo daug liestinių sferų, o rutulys d gali turėti bendrą liestinę sferą su t tada ir tik tada, kai d kerta erdvės sritį, kurią apibrėžia visų įmanomų t liestinių sferų sąjunga. Jei t turi lygiai dvi be galo didelio spindulio liestines sferas, t. y. dvi liestines plokštumas, tai mes vadiname t suvaržytuoju trejetu. Toliau apibrėžiame du atskirus kaimynų paieškos algoritmus: suvaržytiesiems ir nesuvaržytiesiems trejetams.

Suvaržytojo trejeto kaimynų paieška

Suvaržytojo trejeto t dvi liestinės plokštumos dalija visų įmanomų t liestinių sferų gaubiamą sritį į tris dalis:

- puserdvė h_1, atkirsta pirmosios plokštumos;
- puserdvė h_2, atkirsta antrosios plokštumos; h_1 ir h_2 gali kirstis;
- sritis m tarp dviejų plokštumų.



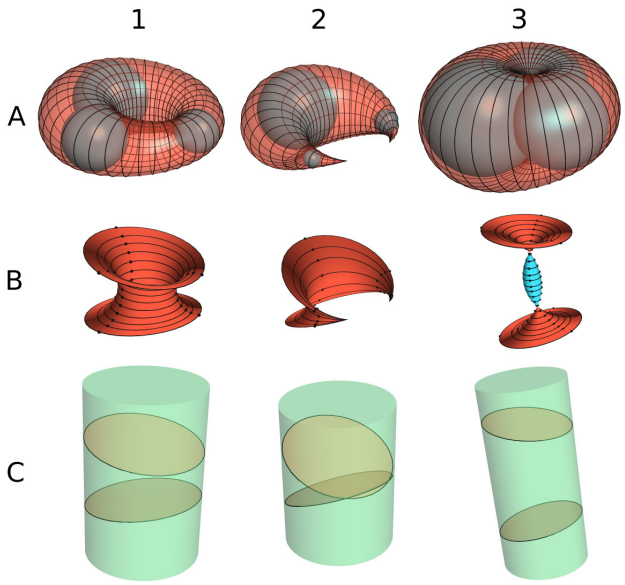
Pav. 1.4: (A) Trimatės erdvės padalijimas dviem liestinėmis plokštumomis. (B) Dvimatis (A) atitikmuo. (C) Suvaržytojo trejeto kaimyno paieškos algoritmo puserdvėje h_2 iliustracija: rezultatas yra rutulys b_4, kurį atitinka h_2 priklausanti tuščioji liestinė sfera.

Toks padalijimas iliustruotas 1.4 pav. (A, B), ten taip pat pavaizduota kreivė C, ant kurios yra išsidėlioję visų t liestinių sferų centrai, t. y. potencialios Voronojaus viršūnės.

Tarkime, trejetas t turi rutulį d_i, kuris kerta $h_x \in \{h_1, h_2\}$. Jei t ir d_i turi vieną liestinę sferą s_i, tai vadiname ją h_x priklausančia t ir d_i liestine sfera. Jei t ir d_i turi dvi liestines sferas, tai h_x priklausančia vadiname tą iš jų, kuri yra arčiau h_x. Kitą trejeto t liestinę sferą s_j galima gauti judinant s_i centrą C kreive (atitinkamai keičiant sferos spindulį). Jei judėjimas nukreiptas h_x link, tai s_j kirs d_i, t. y. s_j nebus tuščia. Judant kita kryptimi (nuo h_x) s_j nelies d_i: tokiu atveju, jei s_i tuščia, tai s_j nelies jokio rutulio iš h_x. Todėl, jei s_i tuščia, tai s_i yra vienintelė h_x priklausanti trejeto t liestinė sfera.

Remdamiesi tokiomis h_x priklausančios sferos savybėmis, apibrėžiame suvaržytojo trejeto t kaimyno h_x puserdvėje paieškos procedūrą. Jos detalus pseudokodas pateikiamas disertacijos pagrindiniame tekste, toliau yra aprašoma supaprastinta schema:

1. Pradedame nuo bet kurio rutulio, kertančio h_x ir turinčio h_x priklausančią liestinę sferą kartu su t.
2. Randame bet kokį rutulį, kuris kerta h_x bei prieš tai gautą h_x priklausančią liestinę sferą ir kuris turi kitą h_x-priklausančią liestinę sferą kartu su t.
3. 2 žingsnis kartojamas tol, kol pavyksta gauti vis naują liestinę sferą.



Pav. 1.5: (A) Trys Dupeno ciklidžių, kurias galima apibrėžti suvaržytiesiems rutulių trejetams, pavyzdžiai. (B) Suvaržytojo trejeto vidurinė sritis atitinka Dupeno ciklidės paviršiaus dalį. (C) Vidurinę sritį galima aproksimuoti cilindru, gaubiančiu apskritimus, kurie apibrėžia vidurinės srities sąlytį su liestinėmis plokštumomis.

4. Jei paskutinė gauta liestinė sfera tuščia, tai atitinkamas rutulys yra priimtinasis t kaimynas.

Šio algoritmo iliustracija pateikta 1.4 pav. (C). Algoritmas yra godusis, jis netikrina visų rutulių, kertančių h_x: jo vykdomų iteracijų skaičių galima sumažinti, jei 1 ir 2 žingsniuose pasirenkami rutuliai iš artimos t kaimynystės. Procedūra realizuojama naudojant pirmiau pristatytą gaubiančių sferų hierarchiją geometrinių objektų (sferų, rutulių, puserdvių) susikirtimams greitai aptikti.

Radus priimtinuosius t kaimynus h_1 ir h_2 puserdvėse, lieka paieškoti kaimynų vidurinėje srityje m. Žinoma, kad šios srities paviršius atitinka dalį geometrinio objekto, vadinamo Dupeno ciklide (1.5 pav., A, B).[19,24] Pasinaudoję šiuo faktu aproksimuojame m sritį cilindru (1.5 pav., C) ir taikome gaubiančiųjų sferų hierarchiją rutulių, kertančių m, paieškai ir atitinkamų liestinių sferų patikrai.

Nesuvaržytojo trejeto kaimynų paieška

Jei priimtinasis trejetas t neturi tiksliai dviejų liestinių plokštumų, jo priimtinųjų kaimynų paieška atliekama tikrinant kiekvieną rutulį $b \in B \setminus t$. b laikomas priimtinuoju kaimynu, jei b ir t turi tuščiąją liestinę sferą: ši sąlyga tikrinama naudo-

jant gaubiančių sferų hierarchiją. Galima būtų apibrėžti greitesnę nesuvaržytojo trejeto kaimynų paieškos procedūrą, tačiau tai tik nežymiai pagreitintų bendro Voronojaus viršūnių paieškos algoritmo greitį: mūsų testai (pateikiami toliau šiame skyriuje) rodo, kad makromolekulių struktūrose nesuvaržytieji trejetai pasitaiko itin retai.

Algoritmo lygiagretinimas

Lygiagretiname algoritmą realizuodami tokią strategiją:

1. Suskirstome įvesties rutulių aibę B į mažesnius poaibius $B_1, B_2, \ldots B_k$: naudojame rekursyvų k-d medžio padalijimo algoritmą[23] taip, kad gauti rutulių poaibiai būtų panašaus dydžio.
2. Lygiagrečiai: kiekvienam $B_i \in \{B_1, B_2, \ldots B_k\}$ surandame priimtinų ketvertų aibę Q_i, kurioje kiekvienas ketvertas turi bent vieną rutulį iš B_i.
3. Grąžiname pilną priimtinųjų ketvertų aibę — 2 žingsnyje rastų aibių sąjungą $Q = Q_1 \cup Q_2 \cup \ldots \cup Q_k$.

2 žingsnyje naudojame bendrąją Voronojaus viršūnių paieškos algoritmo schemą, aprašytą 1.1 dalyje, bet laikomės papildomos sąlygos: kiekvienas apdorojamas priimtinasis trejetas turi turėti bent vieną rutulį iš B_i. Lygiagrečiosios ir nuosekliosios algoritmo versijų korektiškumo įrodymas pateikiamas pagrindiniame disertacijos tekste.

Programinė realizacija

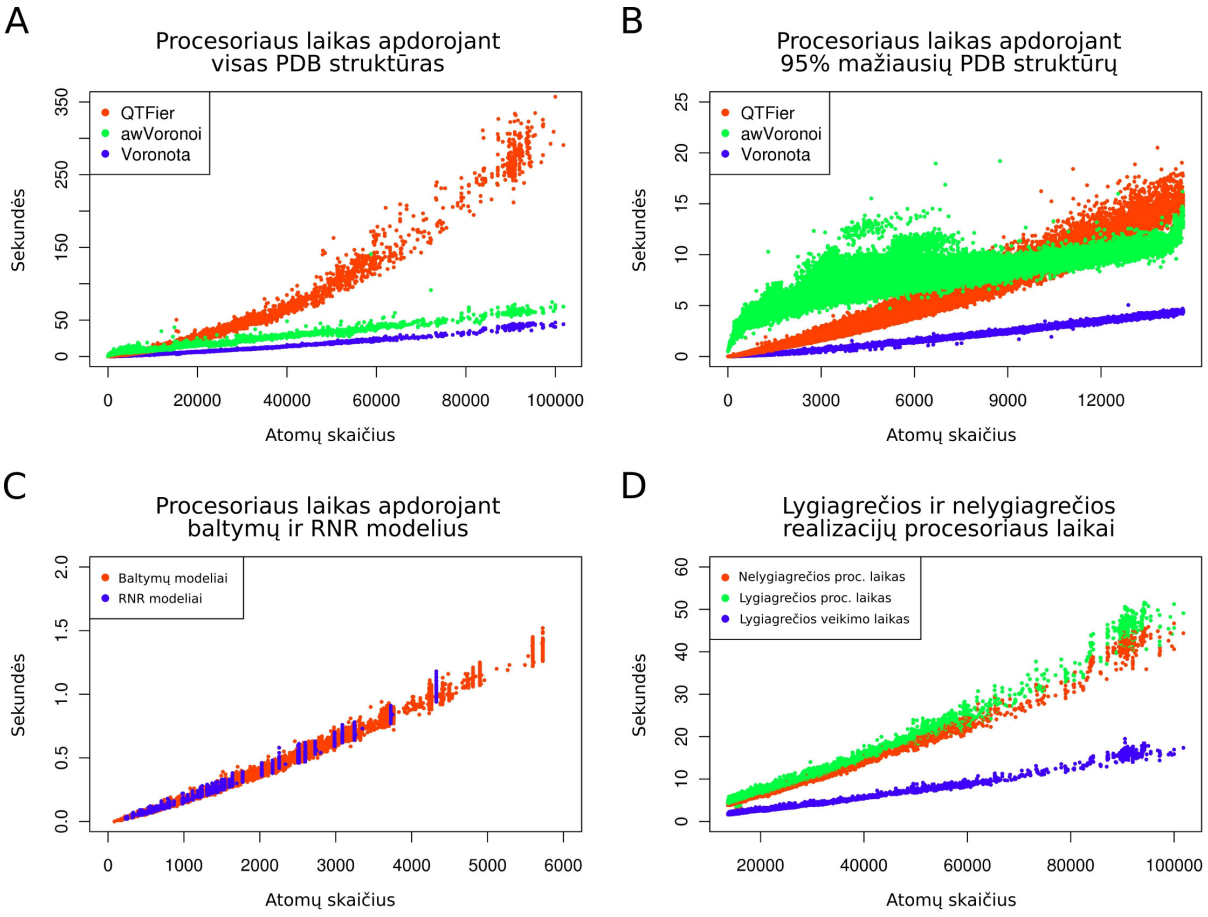
Algoritmas implementuotas C++ kalba, lygiagretinimas realizuotas naudojant OpenMP ir MPI technologijas. Siekiant sumažinti galimų skaičiavimo paklaidų įtaką, lygčių sistemų sprendimuose naudojamas Kahano sumavimo algoritmas.[25] Taip pat naudojamas mažesnes paklaidas užtikrinantis kvadratinių lygčių sprendimo algoritmas.[26] Ieškant liestinių sferų ir plokštumų susikirtimų su rutuliais atsižvelgiama į liestinių sferų ir plokštumų skaičiavimo paklaidas.

1.2 Testavimo rezultatai

Sukurtą programinę įrangą, kurią pavadinome Voronota, pirmiausia ištestavome panaudoję visus „Protein Data Bank“ (PDB) duomenų bazės[27] įrašus. Panaudojome 90365 asimetrinių vienetų struktūras, atsisiųstas 2013.05.15. Ligandai, vandens molekulės ir vandenilio atomai buvo netraukiami į skaičiavimus. Tuos pačius duomenis padavėme QTFier (`voronoi.hanyang.ac.kr/software.htm`, 1.0 versija) ir awVoronoi (`sourceforge.net/projects/awvoronoi`, 1.0.0 versija) programoms: mūsų žiniomis, QTFier ir awVoronoi buvo vieninteliai laisvai prieinami rutulių Voronojaus diagramos skaičiavimo įrankiai. Voronota veikė žymiai greičiau už kitas programas (1.6 pav., A, B). Be to, Voronota, skirtingai negu kitos programos, sugebėjo sėkmingai apdoroti visas įvestas struktūras.

PDB apdorojimo rezultatai parodė, kad priimtinųjų ketvertų skaičius tiesiškai koreliuoja su atomų skaičiumi (Pirsono koreliacijos koeficientas didesnis už 0,99). Priimtinųjų ketvertų vidutiniškai yra 6,6 karto daugiau negu atomų. Tik mažiau negu 0,005 % visų priimtinųjų trejetų buvo nesuvaržytieji ir tik 18 iš $4,5 \cdot 10^8$ atomų rutulių turėjo Voronojaus ląsteles be viršūnių. Vidutiniškai tik 11 ketvertų turėjo būti perrinkti ieškant pirmojo priimtinojo trejeto.

Panašūs testai buvo sėkmingai atlikti ir su visomis prieinamomis BMR struktūromis, turinčiomis vandenilio atomus: šiuo atveju rasta palyginti daugiau nesuvaržytųjų trejetų — apie 0,5 %. Taip pat buvo atlikti testai su baltymų CASP9[10,28] ir RNR[29] modeliais (1.6 pav., C) bei lygiagretinimo testai (1.6 pav., D): Voronota sėkmingai veikė visais atvejais, detalesni rezultatai pateikiami pagrindiniame disertacijos tekste.



Pav. 1.6: (A) Procesoriaus laiko reikšmės, gautos apdorojant visas PDB struktūras. (B) Procesoriaus laiko reikšmės, gautos apdorojant 95 % mažiausių PDB struktūrų. (C) Testavimo su baltymų ir RNR modeliais rezultatai. (D) Voronota lygiagretinimo rezultatai naudjant 4 procesorius. Visi testai atlikti naudojant „Intel Core i7-2600“ 3,40 GHz procesorius.

2 CAD-score: kontaktų plotais pagrįstas metodas makromolekulių erdvinių struktūroms palyginti

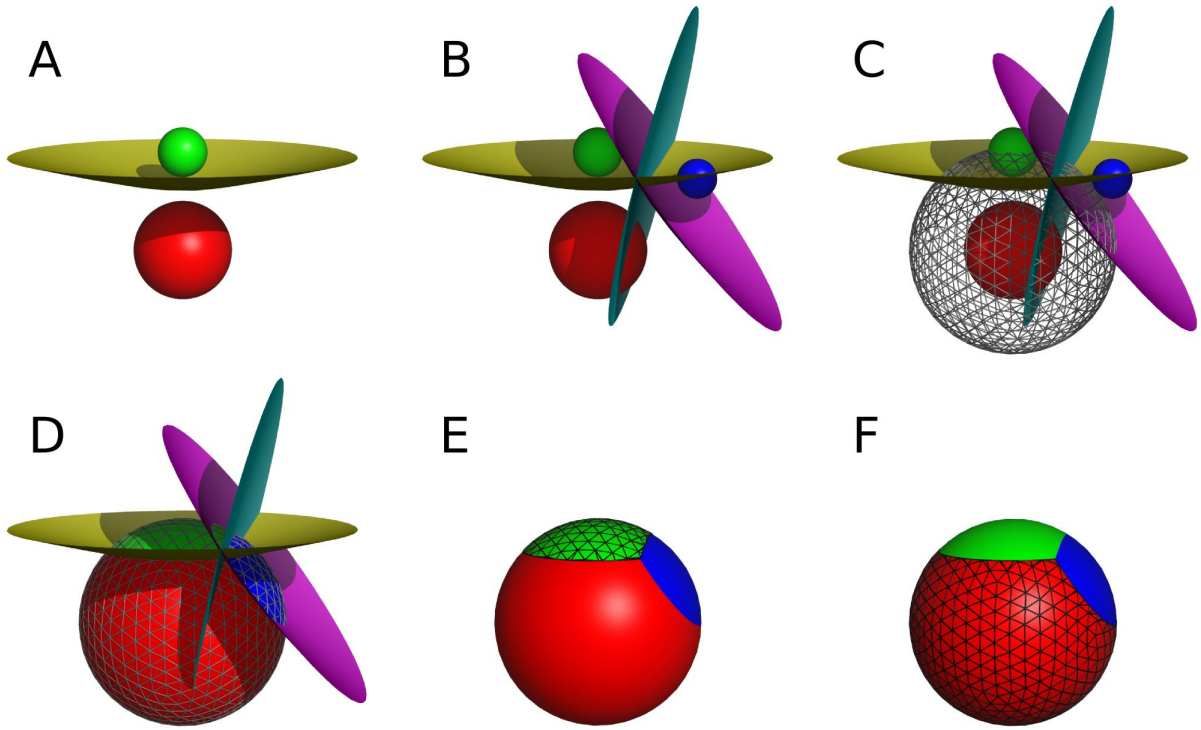
2.1 Trumpas metodo aprašymas

Kontaktų konstravimas

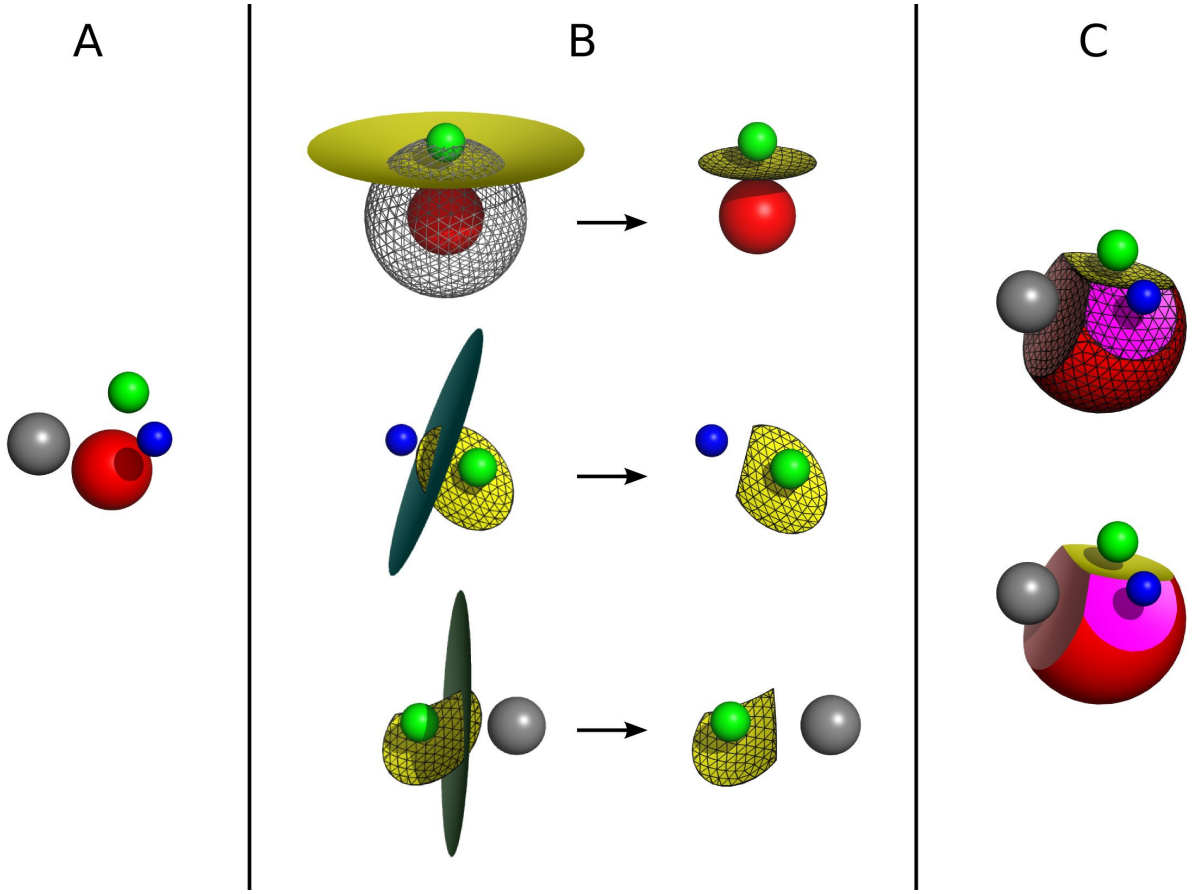
Tarpatominius kontaktus konstruojame naudodami baltymo struktūros atomų rutulių Voronojaus diagramą. Rutulių spinduliai yra atomų van der Valso spinduliai, naudojame Li ir Nussinov spindulių rinkinį.[30] Remiamės McConkey et al.[31] idėja, kad kontaktų paviršius galima konstruoti ant sferos aplink atomą. Laikome, kad du atomai kontaktuoja, jeigu tarp jų negali prasisprausti vandens molekulė. Todėl atomo kontaktų sferos spindulys lygus atomo van der Valso spindulio ir standartinio vandens molekulės spindulio (1,4 Å) sumai.

Taikome ikosaedro padalijimo metodiką[32] kontaktų sferos paviršiaus trianguliuotai reprezentacijai gauti. Kontaktus gauname kirsdami sferos paviršiaus trikampius hiperboloidais, kurie atitinka atomo rutulio Voronojaus ląstelės sieneles (2.1 pav.). Naudojame Kim et al.[19] pasiūlytą hiperboloidų analitinę reprezentaciją. Trikampių „pjaustymo“ hiperboloidais metodas gali būti naudojamas ir kontaktams, tiesiogiai atitinkantiems Voronojaus ląstelės sieneles, konstruoti (2.2 pav.): tokius kontaktus vadiname sukaustytosiomis Voronojaus sienelėmis. Mūsų tyrimų pradžioje naudojome kontaktus ant sferos, nes jų konstravimo realizacija buvo paprastesnė ir veikė greičiau.

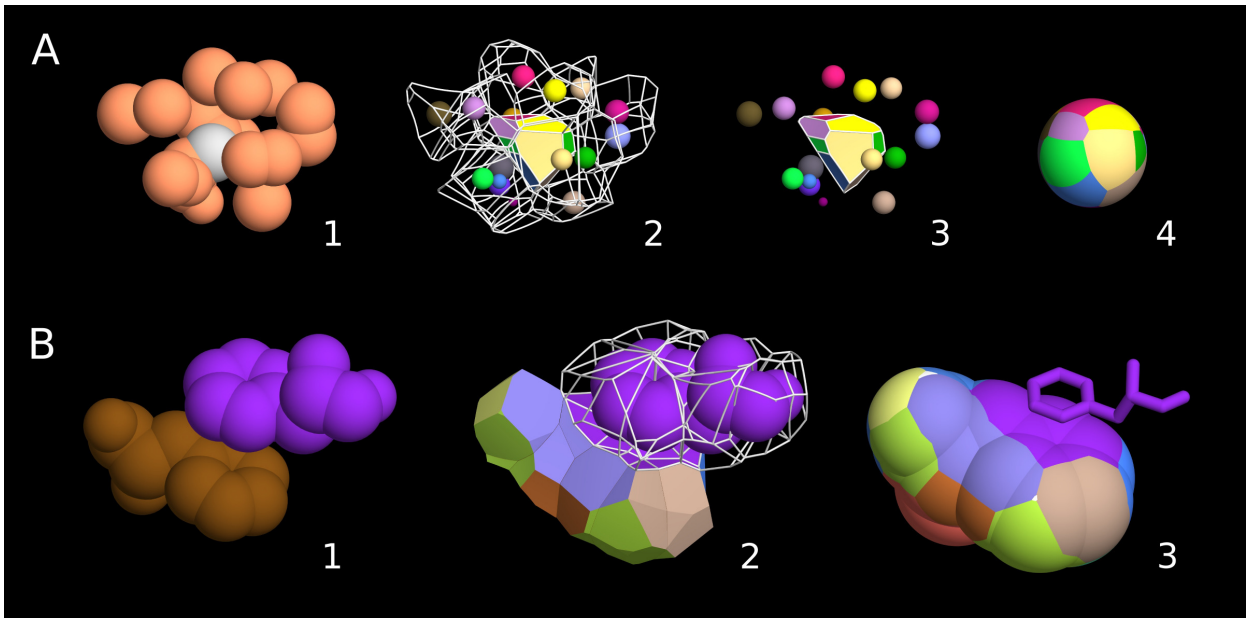
Kontaktai tarp aminorūgščių ar nukleotidų liekanų konstruojami tiesiog grupuojant tarpatominius kontaktus, tai iliustruota 2.3 pav. Grupuojant kontaktus galima atsižvelgti į standartinius liekanų atomų poaibius: pagrindinę grandinę ir šoninę grandinę. Kontaktus tarp nukleobazių galima papildomai suskirstyti į stekingo ir ne stekingo kontaktus naudojant plokštumas, atitinkančias bazes (2.4 pav.). 2.5 pav. parodyti kontaktų tarp liekanų variantai, atitinkantys įvairius tarpatominių kontaktų grupavimo būdus.



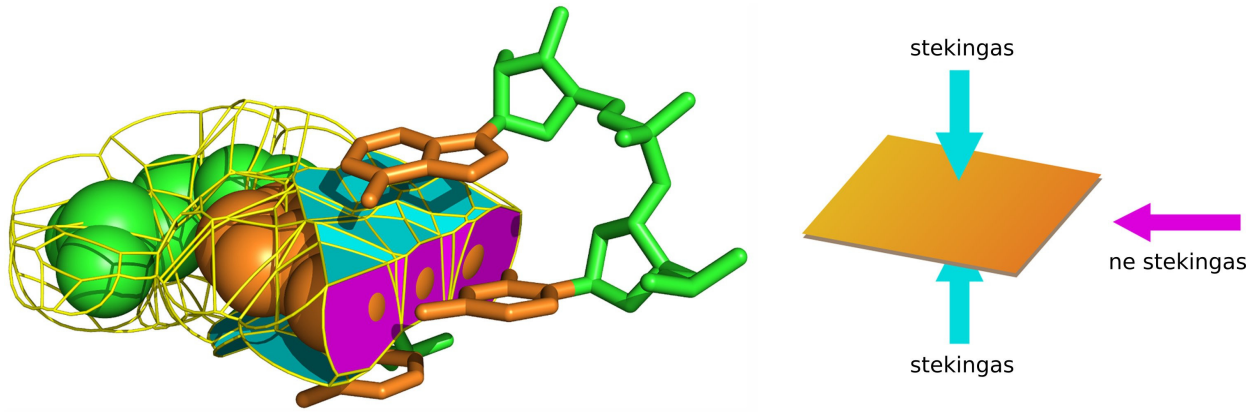
Pav. 2.1: (A) Hiperboloidas tarp dviejų rutulių. Kiekvienas šio hiperboloido taškas yra vienodai nutolęs nuo abiejų rutulių. (B) Hiperboloidai trims rutuliams. (C) Hiperboloidai kerta trianguliuotąją raudonojo atomo kontaktų sferos paviršiaus reprezentaciją. (D) Hiperboloidai padalina kontaktų sferą į tris kontaktų paviršius. (E) Kontaktai apibrėžti (D) dalyje, papildomai išryškinta kontakto su žaliuoju rutuliu paviršiaus trianguliacija. (F) Panašu į (E), tik išryškinta raudonojo rutulio kontakto su tirpikliu paviršiaus trianguliacija.



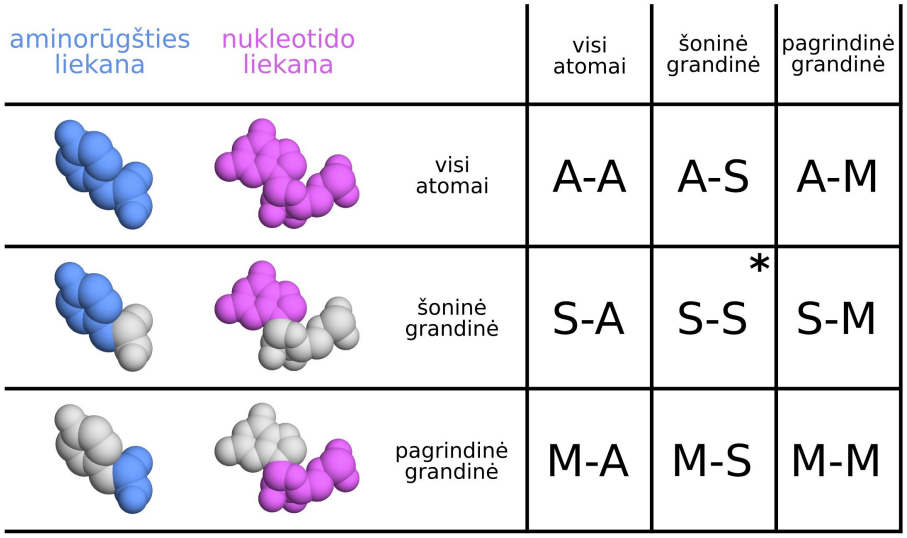
Pav. 2.2: Sukaustytųjų Voronojaus sienelių konstravimas. (A) Keturi kaimyniniai rutuliai. (B) Trys kontakto tarp raudonojo ir žaliojo rutulių konstravimo žingsniai: hiperboloido tarp raudonojo ir žaliojo rutulių dalies, esančios raudonojo atomo kontaktų sferos viduje, trianguliacijos konstravimas; inicializuotos trianguliacijos „pjovimas“ hiperboloidu, apibrėžtu tarp žaliojo ir mėlynojo rutulių; tolimesnis modifikuotos trianguliacijos „pjovimas“ hiperboloidu, apibrėžtu tarp žaliojo ir pilkojo rutulių. (C) Sukaustytosios Voronojaus sienelės, gautos taikant (B) pademonstruotą metodiką, ir atitinkamas raudonojo rutulio kontakto su tirpikliu sferinis paviršius.



Pav. 2.3: (A) Vieno atomo kaimynai (1) ir kontaktai su jais, pavaizduoti keliais būdais (2–4). (B) Dvi baltymo aminorūgščių liekanos (1) ir kontaktas tarp jų (2, 3), gautas sugrupavus tarpatominius kontaktus. Paveikslas sugeneruotas naudojant mūsų sukurtą programinį įrankį Voroprot.[33]



Pav. 2.4: Nukleotidų kontaktų skirstymo į stekingo ir ne stekingo kontaktus iliustracija.



Pav. 2.5: Kontaktų tarp liekanų (aminorūgščių ir nukleotidų) skirstymas pagal tai, kokios atomų grupės kontaktuoja. Atomų grupės koduojamos raidėmis: „A“ (visi atomai), „S“ (šoninės grandinės atomai), „M“ (pagrindinės grandinės atomai).

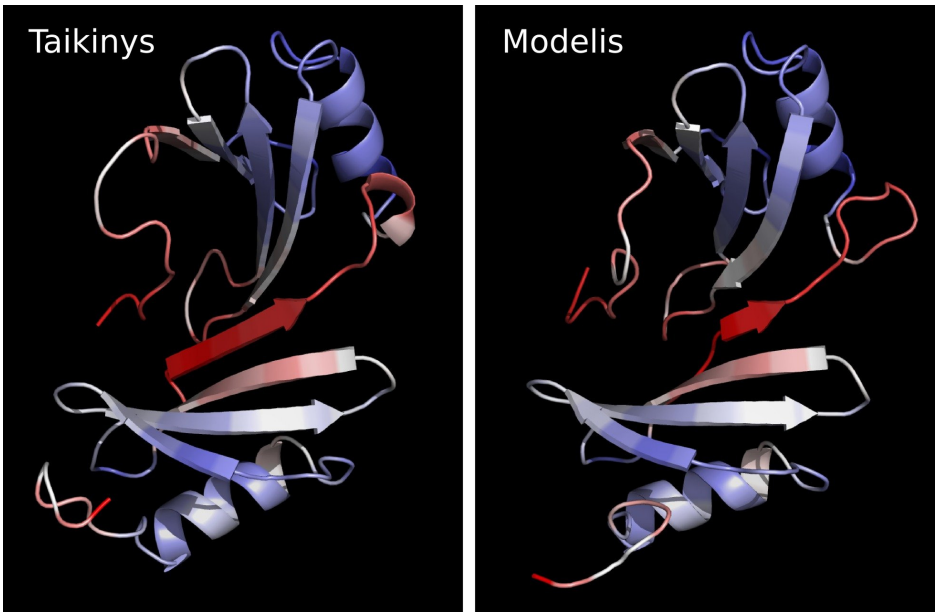
CAD-score apibrėžimas

CAD-score apibrėžiame atsižvelgdami į tris pagrindinius reikalavimus: 1) kontaktai modelyje turi būti vertinami pagal kontaktus, esančius etalono (taikinio) struktūroje; 2) jei modelyje trūksta liekanų, turi būti laikoma, kad atitinkami kontaktai nebuvo teisingai nusakyti; 3) modelio kontaktai, turintys žymiai didesnius plotus nei atitinkami kontaktai taikinyje, turi būti ekvivalentūs visiškai nenusakytiems kontaktams.

Tarkime, G yra aibė visų liekanų porų (i, j), turinčių nenulinio ploto kontaktus $T_{(i,j)}$ taikinio struktūroje. Tada skaičiuojame kiekvienos poros $(i, j) \in G$ kontaktų plotus modelyje, žymime juos $M_{(i,j)}$. $M_{(i,j)}$ lygus nuliui, jei (i, j) modelyje nėra. Toliau galime apibrėžti kiekvienos poros $(i, j) \in G$ kontaktų taikinyje ir modelyje plotų skirtumą:

$$\mathrm{CAD}_{(i,j)} = |T_{(i,j)} - M_{(i,j)}| \tag{2.1}$$

Norėdami simetriškai traktuoti nenusakytusius ir pernelyg didelius nusakytuosius plotus, apibrėžiame apribotą kontaktų plotų skirtumą:



Pav. 2.6: Eksperimentiškai nustatyta struktūra (taikinys) ir nusakyta struktūra (modelis) nuspalvintos pagal lokaliuosius CAD-score įverčius. Raudona spalva žymi blogai nusakytas dalis, mėlyna — gerai nusakytas dalis.

$$\mathrm{CAD}_{(i,j)}^{\mathrm{bounded}} = \min(\mathrm{CAD}_{(i,j)}, T_{(i,j)}) \tag{2.2}$$

Tada viso modelio CAD-score skaičiuojamas taip:

$$\text{CAD-score} = 1 - \frac{\sum_{(i,j)\in G} \mathrm{CAD}_{(i,j)}^{\mathrm{bounded}}}{\sum_{(i,j)\in G} T_{(i,j)}} \tag{2.3}$$

CAD-score reikšmė visada priklauso [0,1] intervalui. Jei modelio ir taikinio struktūros identiškos, tai CAD-score=1. Jei joks taikinio kontaktas neatkartotas modelyje, t. y. nėra atvejų, kai $\mathrm{CAD}_{(i,j)} < T_{(i,j)}$, tai CAD-score=0.

Galima skaičiuoti ne tik visų struktūros kontaktų aibės, bet ir mažesnių poaibių CAD-score vertę. Pagrindiniame disertacijos tekste detaliai aprašomi papildomi CAD-score variantai pavienių liekanų kontaktams ar kontaktams tarp skirtingų grandinių skaičiuoti. Lokaliąsias CAD-score reikšmes galima vizualizuoti spalvinant taikinių ir modelių struktūras (2.6 pav.).

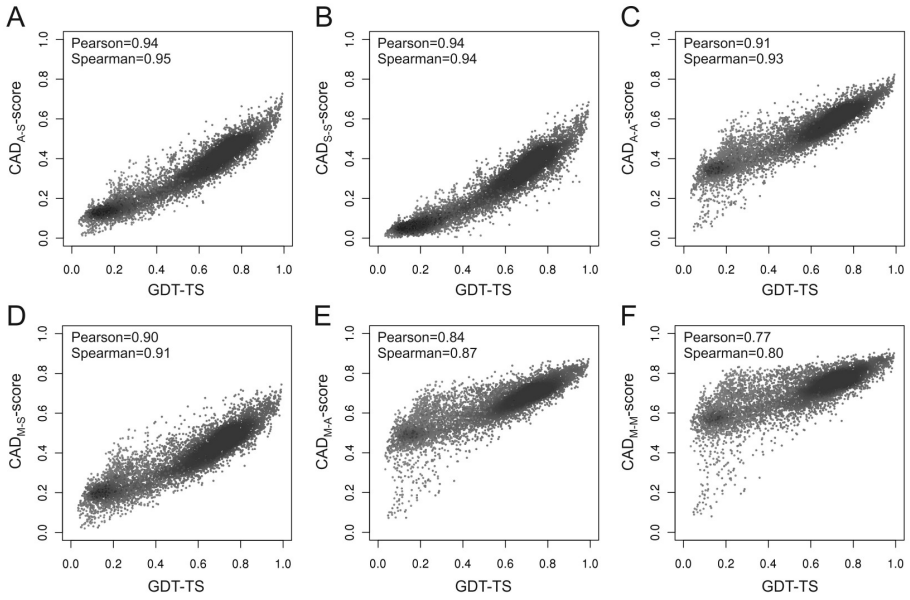
2.2 Testavimo rezultatai

Testavimo naudojant baltymų struktūrų modelius rezultatai

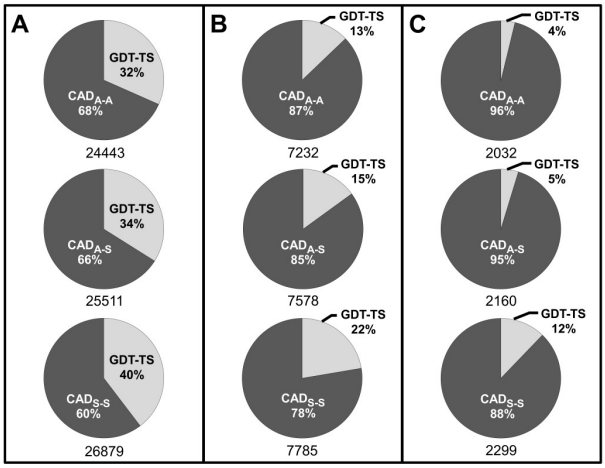
Testavimui panaudojome CASP9 (angl. „The Ninth Community-wide Assessment of Protein Structure Prediction“)[28,34] tarptautinio eksperimento duomenis. CASP modeliai sugeneruoti naudojant dešimtis įvairiausių metodų, modelių kokybė yra labai įvairi, o modelių vertinimo bei rikiavimo (pagal modelio panašumą į taikinį) uždavinys yra sudėtingas.[35]

Pirmiausiai palyginome CAD-score su pagrindiniu CASP naudojamu įverčiu — GDT-TS (angl. „Global Distance Test Total Score“).[36] GDT-TS efektyviai vertina atskirus baltymų domenus, t. y. mažiau lanksčias, nepriklausomai stabilias substruktūras. Domenų lygyje bet koks prasmingas struktūrų palyginimo metodas turėtų koreliuoti su GDT-TS. CAD-score tenkina šį reikalavimą, tai iliustruota 2.7 pav., kas yra gana stebėtina turint omenyje visiškai skirtingą CAD-score ir GDT-TS prigimtį. Tačiau taikant CAD-score ir GDT-TS metodus nustatytas geresnis iš dviejų modelis ne visada sutampa. Tokiems nesutapimams tirti buvo pasitelktas trečiasis nepriklausomas MolProbity metodas,[37] kuris vertina baltymo struktūros modelio fizinį realizmą nelygindamas modelio su taikiniu. Remiantis MolProbity vertinimu, taikant CAD-score žymiai dažniau išskiriamas realistiškesnis modelis negu GDT-TS, tai iliustruota 2.8 pav.

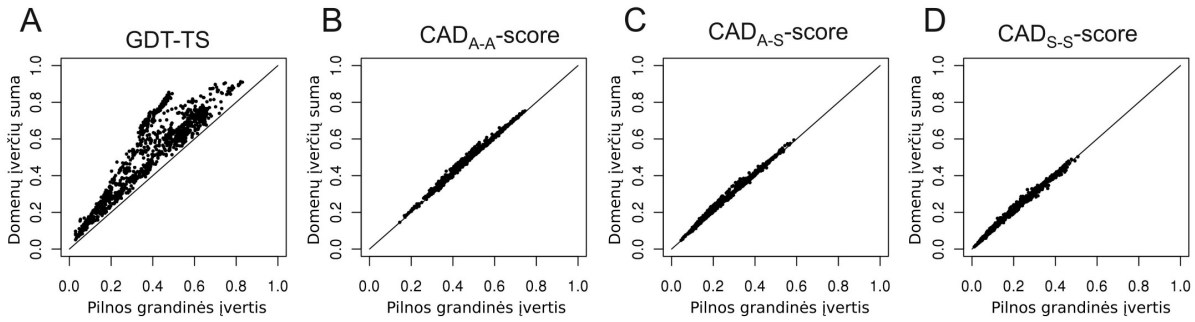
CAD-score, skirtingai negu GDT-TS, nesiremia struktūrų superpozicija, todėl teoriškai turėtų efektyviai analizuoti multidomenines struktūras. Patikrinome tai palyginę CASP9 pilnų multidomeninių modelių įverčius (CAD-score ir GDT-TS) su atitinkamų atskirų domenų įverčių svertinėmis sumomis. Rezultatai pavaizduoti 2.9 pav., jie parodo, kad CAD-score efektyviai veikia ir neskaidant baltymų į domenus. Bet tai nereiškia, kad CAD-score neatsižvelgia į domenų tarpusavio sąveikas: tokių sąveikų indėlis į bendrą įvertį atitinka sąveikų kontaktų plotus. Be to, CAD-score geba tiesiogiai vertinti sąveikas tarp domenų ar grandinių (2.10 pav.). Apibendrinant, CAD-score nebūdingos pagrindinės GDT-TS problemos, bet būdingi pagrindiniai privalumai: išsamią diskusiją apie tai galima rasti pagrindiniame disertacijos tekste.



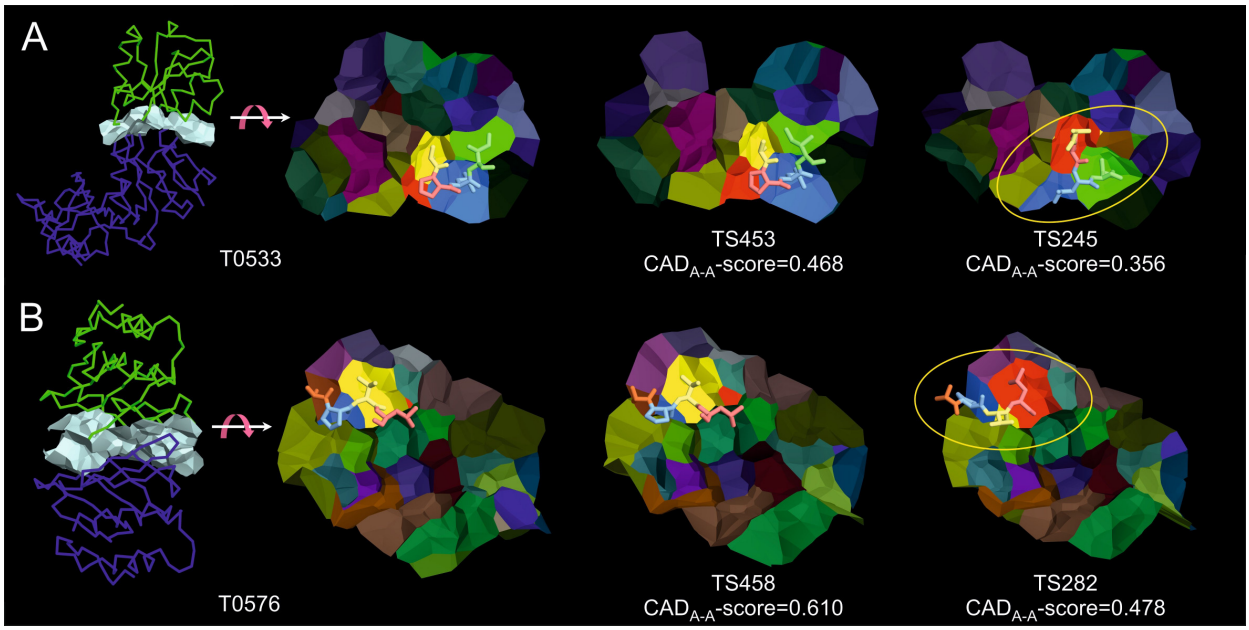
Pav. 2.7: GDT-TS ir kelių CAD-score variantų koreliacijos grafikai.



Pav. 2.8: Nagrinėjamos CASP9 modelių poros, kuriose konfliktuoja GDT-TS ir CAD-score rangai. Paimti tik modeliai su didesne už 0,6 (60 %) GDT-TS verte. Skritulinės diagramos rodo, kaip dažnai MolProbity vertinimas sutampa su GDT-TS ir su pagrindiniais CAD-score variantais. Analizuotų porų skaičius nurodytas po kiekviena diagrama. (A) Visos poros. (B) Poros, kur MolProbity skirtumas didesnis už empirinį standartinį nuokrypį. (C) Poros, kur tiek MolProbity, tiek GDT-TS ar CAD-score skirtumai didesni už atitinkamus empirinius standartinius nuokrypius.



Pav. 2.9: Koreliacijos tarp visos struktūros įverčių ir atitinkamų domenų įverčių svertinių sumų. Parodyti kelių metodų rezultatai: (A) GDT-TS, (B) CAD_{A-A}, (C) CAD_{A-S} ir (D) CAD_{S-S}.

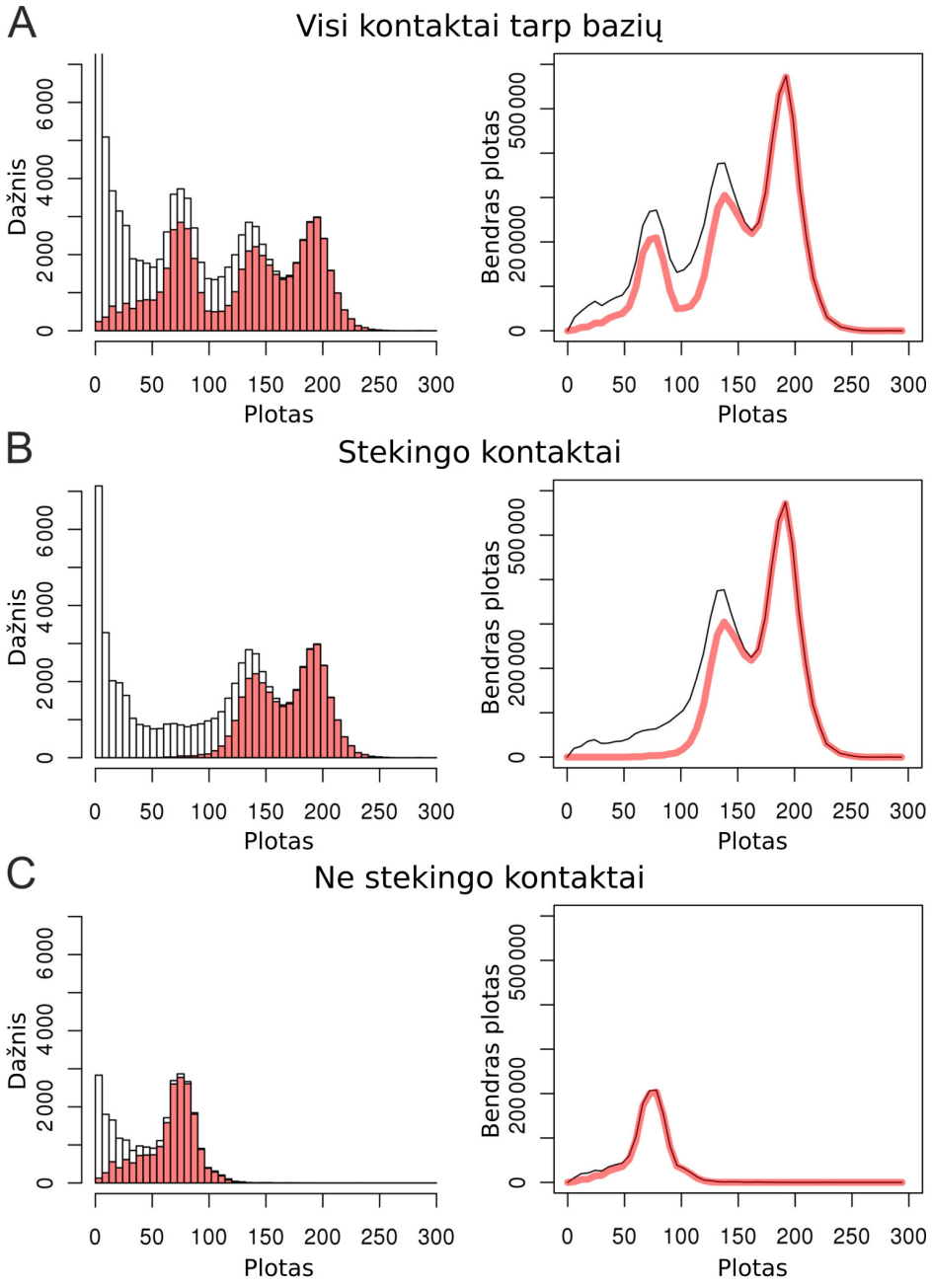


Pav. 2.10: CAD-score naudojimas sąveikoms tarp domenų (A) ir tarp grandinių (B) vertinti. Sąveikos pavaizduotos kaip atitinkamos Voronojaus ląstelių sienelės ir nuspalvintos pagal kontaktuojančių aminorūgščių liekanų identifikatorius. Dideli skirtumai blogiau nusakytuose modeliuose pažymėti geltonai.

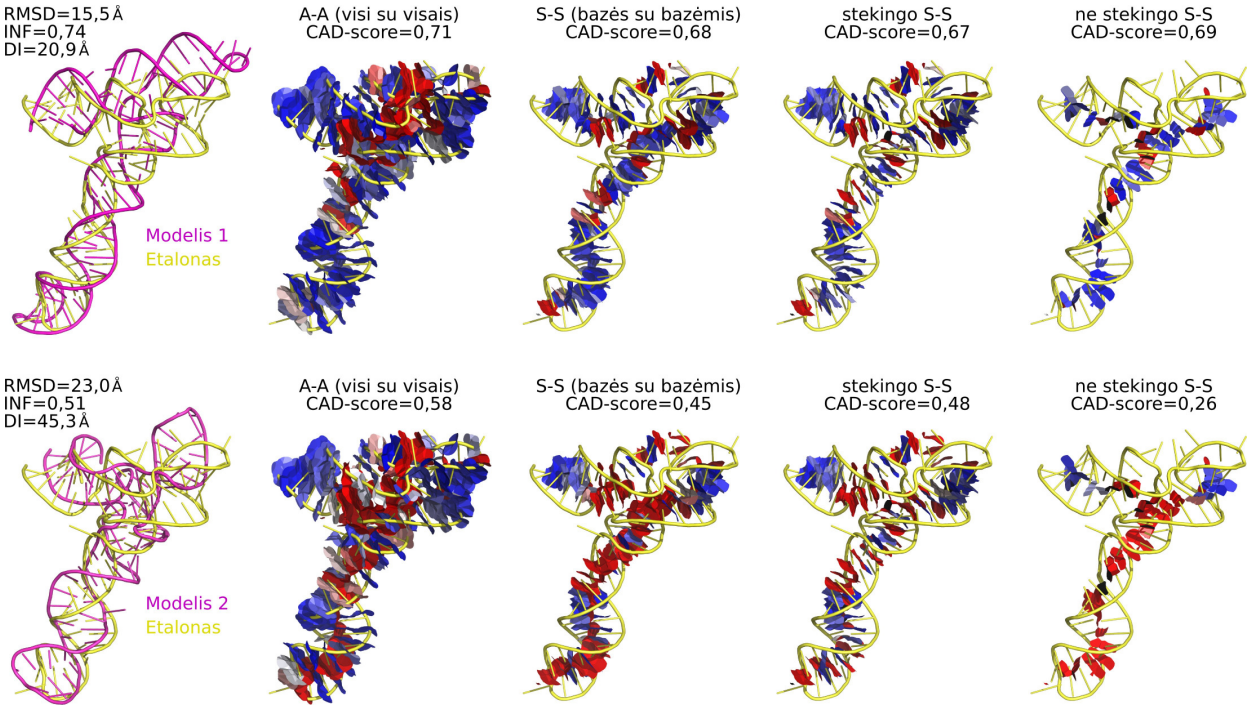
Testavimo naudojant RNR struktūrų modelius rezultatai

Prieš taikydami CAD-score metodą RNR modeliams vertinti, panagrinėjome, kokią įtaką turi šoninės grandinės atomų kontaktai RNR struktūrose lyginant su analogiškų kontaktų įtaka baltymuose. Išanalizavę aukštos kokybės struktūras iš PDB gavome, kad kontaktų tarp nukleobazių plotas vidutiniškai sudaro net 49 % visų kontaktų ploto RNR struktūrose. Baltymuose kontaktai tarp šoninės grandinės atomų sudaro 30 % bendrojo kontaktų ploto. Tokie rezultatai rodo, kad kontaktus tarp nukleobazių verta paanalizuoti detaliau. Visų pirma, panagrinėjome atskirų kontakų plotų pasiskirstymo histogramą ir atitinkamą svertinę histogramą, kurioje dažniai padauginti iš atitinkamų plotų. Nagrinėjome ne tik CAD-score aptiktus kontaktus, bet ir kontaktus, kuriuos atpažino ir suklasifikavo specialus RNR struktūrų anotavimo įrankis MC-Annotate.[38] MC-Annotate įrankis koncentruojasi į patikimai klasifikuojamas sąveikas, todėl aptiko mažiau mažo ploto kontaktų (2.11 pav., A, kairėje). Tačiau svertinė histograma (2.11 pav., A, dešinėje) rodo, kad galimai nesvarbūs, „triukšmo“ kontaktai sudaro palyginti nedidelę bendrojo kontaktų ploto dalį. Taigi galima laikyti kontakto plotą savotišku kontakto svarbumo matu: tai yra vienas iš pagrindinių CAD-score privalumų. Taip pat dėl kontaktų plotų paprasto CAD-score stekingo ir ne stekingo sąveikų klasifikavimo metodo rezultatai ne tik gerai atitinka MC-Annotate išvestį, bet tuo pačiu aprėpia daugiau kontaktų (2.11 pav., B, C).

Toliau testavome CAD-score naudodami taikinių ir modelių struktūras iš RNA-puzzles[39] eksperimento, kuris idėjiškai panašus į CASP. Lyginome mūsų metodo rezultatus su RNA-puzzles organizatorių naudojamais RMSD (angl. „Root Mean Square Deviation“),[40] INF (angl. „Interaction Network Fidelity“)[41] ir DI (angl. „Deformation Index“).[41] RMSD paremtas struktūrų superpozicija, INF — MC-Annotate anotacijų palyginimu, o DI yra RMSD ir INF kombinacija. Panašiai kaip ir baltymų modelių atveju, parodėme, kad CAD-score sugeba efektyviai išrikiuoti RNR modelius pagal kokybę, o MolProbity vertinimo rezultatai labiau atitinka CAD-score, kai jo vertinimas nesutampa su DI, INF ar RMSD. Be to, CAD-score, skirtingai nei RMSD ir DI, atsižvelgia į modelių nepilnumą: jeigu vienas modelis yra kito modelio dalinė versija, mažesnis modelis vertinamas kaip blogesnis. Detalūs testų rezultatai pateikiami disertacijos pagrindiniame tekste. Čia norėtume pabrėžti, kad CAD-score leidžia pamatyti tiesioginį ryšį tarp modelio struktūros lokaliųjų klaidų ir globaliojo to modelio įverčio. 2.12 pav. pavaizduoti dviejų modelių vertinimo pavyzdžiai: pirmojo modelio atveju stekingo ir ne stekingo kontaktų nusakymo lygis panašus, antrajame modelyje ne stekingo kon-



Pav. 2.11: Kontaktų tarp nukleobazių dažnio (kairioji histograma) ir bendrojo ploto (dešinioji, svertinė histograma) priklausomybė nuo kontakto ploto dydžio. Rodomi visų kontaktų tarp bazių (A), stekingo (B) ir ne stekingo kontaktų (C) duomenys. Balti stulpeliai ir juodos linijos atitinka CAD-score aptiktus kontaktus, raudoni stulpeliai ir storos raudonos linijos — MC-Annotate aptiktus kontaktus.

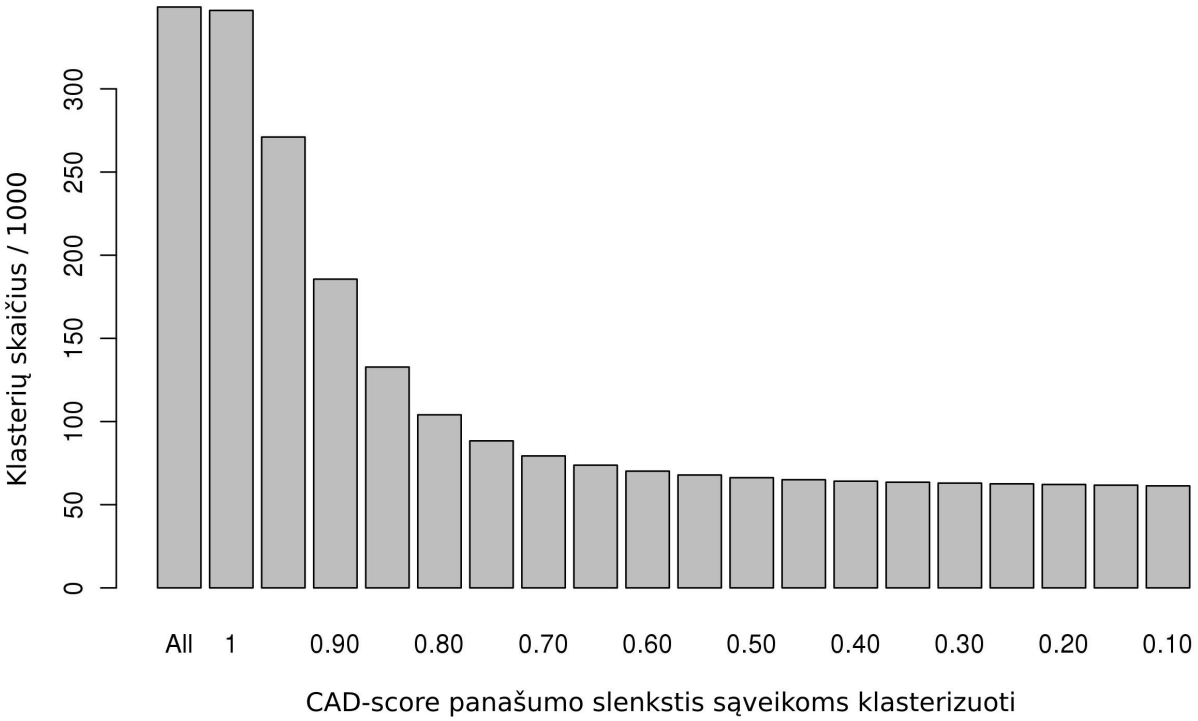


Pav. 2.12: Dviejų skirtingos kokybės modelių vertinimo pavyzdžiai. Kontaktai pavaizduoti kaip sukaustytosios Voronojaus ląstelių sienelės ir nuspalvinti pagal nusakymo tikslumą naudojant mėlyna-balta-raudona spalvų gradientą (mėlyna — tiksliai, raudona — netiksliai).

taktai nusakyti žymiai blogiau. Be to, 2.12 pav. parodyta, kad CAD-score leidžia susikoncentruoti ne tik į kontaktų grupes, bet ir į atskirus kontaktus. Apibendrinant, parodėme, kad CAD-score iš tikrųjų yra universalus: jis sėkmingai veikia ne tik su baltymų, bet ir su nukleorūgščių struktūromis. Pagrindinėmis CAD-score funkcijomis galima pasinaudoti CAD-score serveryje, kuriame taip pat realizuotas ir baltymų-nukleorūgščių kompleksų modelių vertinimas.

Baltymų sąveikų klasterizavimo rezultatai

Panaudojome CAD-score duomenis baltymų kompleksų struktūrų sąveikoms klasterizuoti PPI3D (angl. „Protein-Protein Interactions in 3D“) serveryje.[9] PPI3D bendroji veikimo schema pateikiama pagrindiniame disertacijos tekste, čia susikoncentruojame į visų žinomų eksperimentiškai nustatytų baltymų struktūrų sąveikų klasterizavimą ir jo rezultatus. Naudojome Teiloro-Butinos[42] klasterizavimo algoritmą ir du panašumo kriterijus: baltymų grandinių sekų panašumą (pirminiam klasterizavimui) ir CAD-score globalųjį įvertį (klas-



Pav. 2.13: PDB sąveikų klasterizavimo rezultatai gauti 2015.12.04 naudojant 95 % sekų panašumo slenkstį ir įvairius CAD-score slenksčius.

terizavimui pirminių klasterių viduje). Kadangi reikėjo lyginti skirtingos sekos struktūras, naudojome MAFFT[43] metodu sugeneruotus sekų palyginius aminorūgščių liekanų numeracijai sutapatinti. Rezultatai, kurių dalis iliustruota 2.13 pav., parodė, kad vien sekų panašumo kriterijaus neužtektų sąveikų struktūroms efektyviai suklasterizuoti: CAD-score taikymas buvo visiškai pagrįstas ir leido atskleisti atvejus, kai evoliuciškai gimininingi baltymų kompleksai turi nepanašius sąveikos paviršius.

3 VoroMQA: tarpatominių kontaktų plotais pagrįstas metodas baltymų struktūrų modelių tikslumui nusakyti nežinant etalono

3.1 Trumpas metodo aprašymas

Kontaktų konstravimas ir kategorizavimas

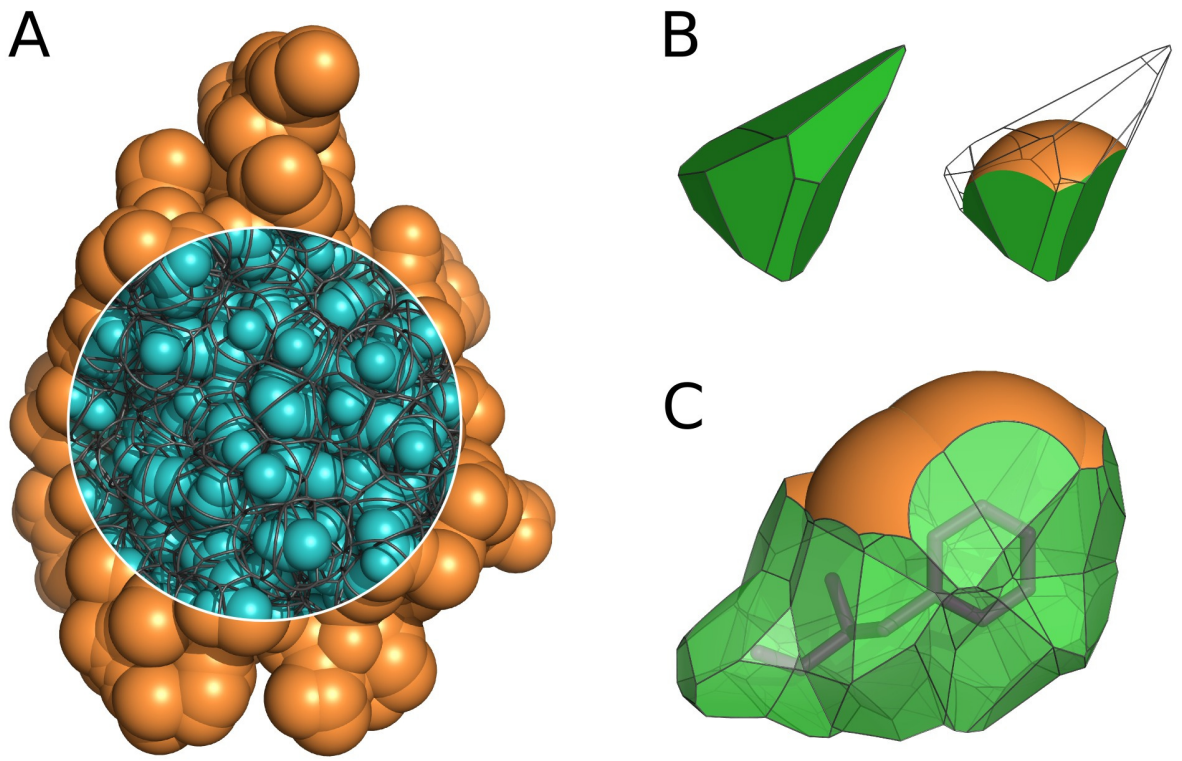
Konstruojame duotos baltymo struktūros atomų rutulių Voronojaus diagramą ir apribojame Voronojaus ląsteles baltymo tirpikliui prieinamu paviršiumi: tokiu būdu gaunamų kontaktų paviršių pavyzdžiai pavaizduoti 3.1 pav. Tokių kontaktų konstravimą ir jų plotų skaičiavimą realizavome Voronota[5] programiniame pakete.

Pastebėjome, kad Voronojaus kontaktų įvairovėje galima išskirti dvi paprastas kontaktų klases; pavadinome jas centrinių ir necentrinių kontaktų klasėmis. Jei tiesė, einanti per dviejų kontaktuojančių atomų centrus, kerta tų atomų kontakto paviršių, tas kontaktas vadinamas centriniu, jei nekerta — necentriniu: tai iliustruota 3.2 pav., A. Aprašinėdami kontaktus taip pat atsižvelgiame į kontaktuojančių aminorūgščių atstumą baltymo sekoje (3.2 pav., B–F).

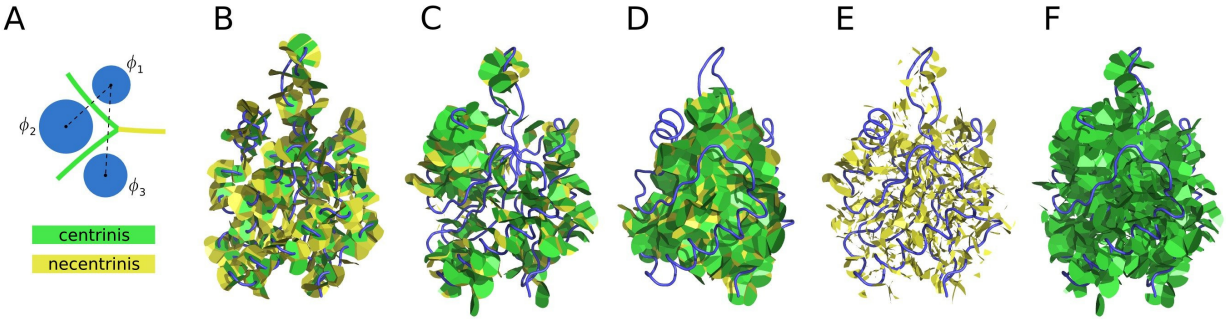
Kokybės įverčio apibrėžimas

Atomų kontaktų plotai gali būti naudojami baltymų struktūrinių modelių kokybei vertinti pasitelkiant žiniomis paremto statistinio potencialo idėją;[44,45] pirmą tokį bandymą atliko McConkey et al.[46] Mūsų metodas remiasi tuo pačiu principu, bet naudoja detalesnius kontaktų aprašus ir kitaip skaičiuoja kontaktų tikimybes.

Pirmiausia apibrėžkime galimų kontaktų tipų aibę. Tarkime, $A = \{a_0, a_1, ..., a_n\}$ yra atomų tipų aibė, o $C = \{c_0, c_1, ..., c_m\}$ kontaktų kategorijų aibė (pavyzdžiui, kontakto kategorija gali rodyti, ar jis centrinis ir kokiu atstumu sekoje yra atitinkamos aminorūgščių liekanos). Kontaktų tipas apibrėžiamas trejetu $(a_i, a_j, c_k) \in A \times A \times C$, kuris ekvivalentus (a_j, a_i, c_k), nes mes laikome, kad kontaktai neturi krypčių.



Pav. 3.1: (A) Voronojaus ląstelių, sukaustytų baltymo struktūros paviršiaus viduje, briaunos. (B) Nesukaustytoji ir sukaustytoji Voronojaus ląstelės ir atitinkamas tirpikliui (vandeniui) pasiekiamas paviršius. (C) Vienos aminorūgšties liekanos kontaktų su kitomis liekanomis ir su vandeniu integralusis paviršius, gautas kombinuojant pavienių atomų paviršius.



Pav. 3.2: (A) Centrinių ir necentrinių kontaktų 2D iliustracija: kontaktas tarp ϕ_1 ir ϕ_3 rutulių yra necentrinis, kiti kontaktai yra centriniai. (B) Centriniai (žali) ir necentriniai (geltoni) kontaktai, kai atstumas sekoje lygus 1. (C) Centriniai ir necentriniai kontaktai, kai atstumas sekoje yra nuo 2 iki 6. (D) Centriniai ir necentriniai kontaktai, kai atstumas sekoje yra didesnis už 6. (E) Tik necentriniai kontaktai, kai atstumas sekoje yra didesnis už 1. (F) Tik centriniai kontaktai, kai atstumas sekoje yra didesnis už 1. Naudotos struktūros PDB ID yra 1T3Y.

Atomo tipas a_0 reprezentuoja tirpiklį, o kontakto kategorija c_0 reprezentuoja tirpikliui pasiekiamą paviršių, todėl a_0 ir c_0 visada pasirodo kartu. Visų galimų kontaktų aibė yra $T = ([A \setminus a_0] \times [A \setminus a_0] \times [C \setminus c_0]) \cup ([A \setminus a_0] \times \{a_0\} \times \{c_0\})$.

Kontakto tipui gali būti priskiriama pseudo-energijos reikšmė $\mathrm{E}(a_i, a_j, c_k)$, kuri išvedama iš stebimų ($\mathrm{P_{obs}}$) ir laukiamų ($\mathrm{P_{exp}}$) tikimybių reikšmių:

$$\mathrm{E}(a_i, a_j, c_k) = \log \frac{\mathrm{P_{exp}}(a_i, a_j, c_k)}{\mathrm{P_{obs}}(a_i, a_j, c_k)} \tag{3.1}$$

Stebimos tikimybės skaičiuojamos empiriškai, naudojant aukštos kokybės eksperimentiškai nustatytų baltymų struktūrų kontaktų plotų reikšmes. Tarkime, $\mathrm{S}(a_i, a_j, c_k)$ yra visų (a_i, a_j, c_k) tipo kontaktų plotų, stebimų mokymosi aibėje, suma. Taip pat pažymėkime, kad jei $(a_i, a_j, c_k) \notin T$, tai $\mathrm{S}(a_i, a_j, c_k) = 0$. Tarkime, $\mathrm{S_{sol}}$ ir $\mathrm{S_{int}}$ yra, atitinkamai, kontaktų su tirpikliu ir tarpatominių kontaktų plotų sumos:

$$\mathrm{S_{sol}} = \sum_{1 \le i \le n} \mathrm{S}(a_i, a_0, c_0) \tag{3.2}$$

$$\mathrm{S_{int}} = \sum_{1 \le i \le n} \sum_{1 \le j \le i} \sum_{1 \le k \le m} \mathrm{S}(a_i, a_j, c_k) \tag{3.3}$$

Tada (a_i, a_j, c_k) tipo stebima tikimybė yra plotų sumų santykis:

$$\mathrm{P_{obs}}(a_i, a_j, c_k) = \frac{\mathrm{S}(a_i, a_j, c_k)}{\mathrm{S_{int}} + \mathrm{S_{sol}}} \tag{3.4}$$

Atitinkama laukiama tikimybė turi atitikti tai, kaip dažnai tokio tipo kontaktai būtų sutinkami atsitiktinai sulankstytų baltymų struktūrų aibėje. Tokia tikimybė skaičiuojama naudojant kontakto tipo (a_i, a_j, c_k) izoliuotųjų komponenčių tikimybes:

$$\mathrm{P_{exp}}(a_i, a_j, c_k) = \begin{cases} \mathrm{P_{obs}}(a_i) \cdot \mathrm{P_{obs}}(c_0) & \text{if } j = 0 \\ \mathrm{P_{obs}}(a_i) \cdot \mathrm{P_{obs}}(a_j) \cdot \mathrm{P_{obs}}(c_k) & \text{if } j \ge 1, i = j \\ \mathrm{P_{obs}}(a_i) \cdot \mathrm{P_{obs}}(a_j) \cdot 2 \cdot \mathrm{P_{obs}}(c_k) & \text{if } j \ge 1, i \ne j \end{cases} \tag{3.5}$$

$$\mathrm{P_{obs}}(a_i) = \frac{\sum_{0 \le j \le n} \sum_{0 \le k \le m} \mathrm{S}(a_i, a_j, c_k)}{2\,\mathrm{S_{int}} + \mathrm{S_{sol}}} \tag{3.6}$$

$$\mathrm{P}_{\mathrm{obs}}(c_k) = \frac{\sum_{0 \le i \le n} \sum_{0 \le j \le i} \mathrm{S}(a_i, a_j, c_k)}{\mathrm{S}_{\mathrm{int}} + \mathrm{S}_{\mathrm{sol}}} \tag{3.7}$$

Pseudo-energijos reikšmės, apibrėžtos lygtimis 3.1–3.7, naudojamos atskirų atomų aplinkos realistiškumui vertinti. Skaičiuojame atomo ϕ ir jo bei jo kaimynų kontaktų aibės Ω_ϕ normalizuotą pseudo-energiją naudodami informaciją apie kiekvieno kontakto $\omega \in \Omega_\phi$ plotą (area_ω) ir tipą ($\mathrm{type}_\omega \in T$):

$$\mathrm{E_n}(\Omega_\phi) = \frac{\sum_{\omega \in \Omega_\phi} \mathrm{E}(\mathrm{type}_\omega) \cdot \mathrm{area}_\omega}{\sum_{\omega \in \Omega_\phi} \mathrm{area}_\omega} \tag{3.8}$$

Atomo kokybės įvertį $\mathrm{Q_a}(\Omega_\phi) \in [0, 1]$ apibrėžiame naudodami Gauso klaidos funkciją:

$$\mathrm{Q_a}(\Omega_\phi) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\mathrm{E_n}(\Omega_\phi) - \mu_{\mathrm{type}_\phi}}{\sigma_{\mathrm{type}_\phi}\sqrt{2}}\right)\right) \tag{3.9}$$

Įvertinami kiekvieno atomo tipo μ (vidurkis) ir σ (standartinis nuokrypis) parametrai naudojant mokymosi aibės struktūras.

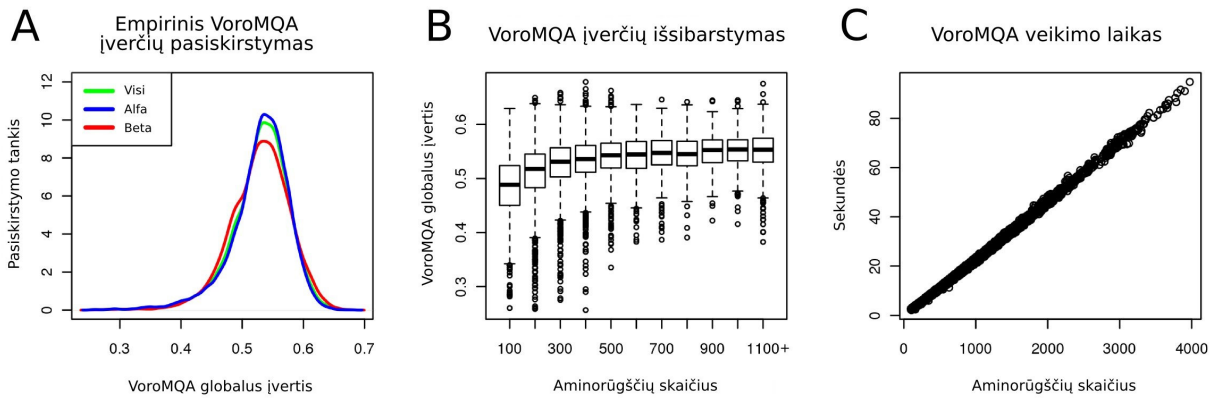
Baltymo struktūros, kurios visų atomų aibė yra Φ, globalųjį kokybės įvertį $\mathrm{Q_g}(\Phi)$ skaičiuojame kaip atskirų atomų kokybės įverčių svertinį aritmetinį vidurkį (svoriai atitinka atomų topologinį atstumą nuo tirpiklio):

$$\mathrm{Q_g}(\Phi) = \frac{\sum_{\phi \in \Phi} \mathrm{Q_a}(\Omega_\phi) \cdot \mathrm{weight}_\phi}{\sum_{\phi \in \Phi} \mathrm{weight}_\phi} \tag{3.10}$$

Panašiai apibrėžiami ir lokalesni, aminorūgščių liekanų lygio įverčiai.

3.2 Testavimo rezultatai

Aukščiau apibrėžtą metodą, pavadintą VoroMQA, realizavome programiškai ir pirmiausiai išbandėme panaudoję aukštos kokybės eksperimentiškai nustatytas baltymų struktūras. Rezultatai, pavaizduoti 3.3 pav., parodė, kad realių struktūrų VoroMQA globalieji įverčiai daugiausia koncentruojasi intervale nuo 0,4 iki 0,65, jie mažai priklauso nuo struktūroje dominuojančio antrinės struktūros tipo ir nuo struktūros dydžio.

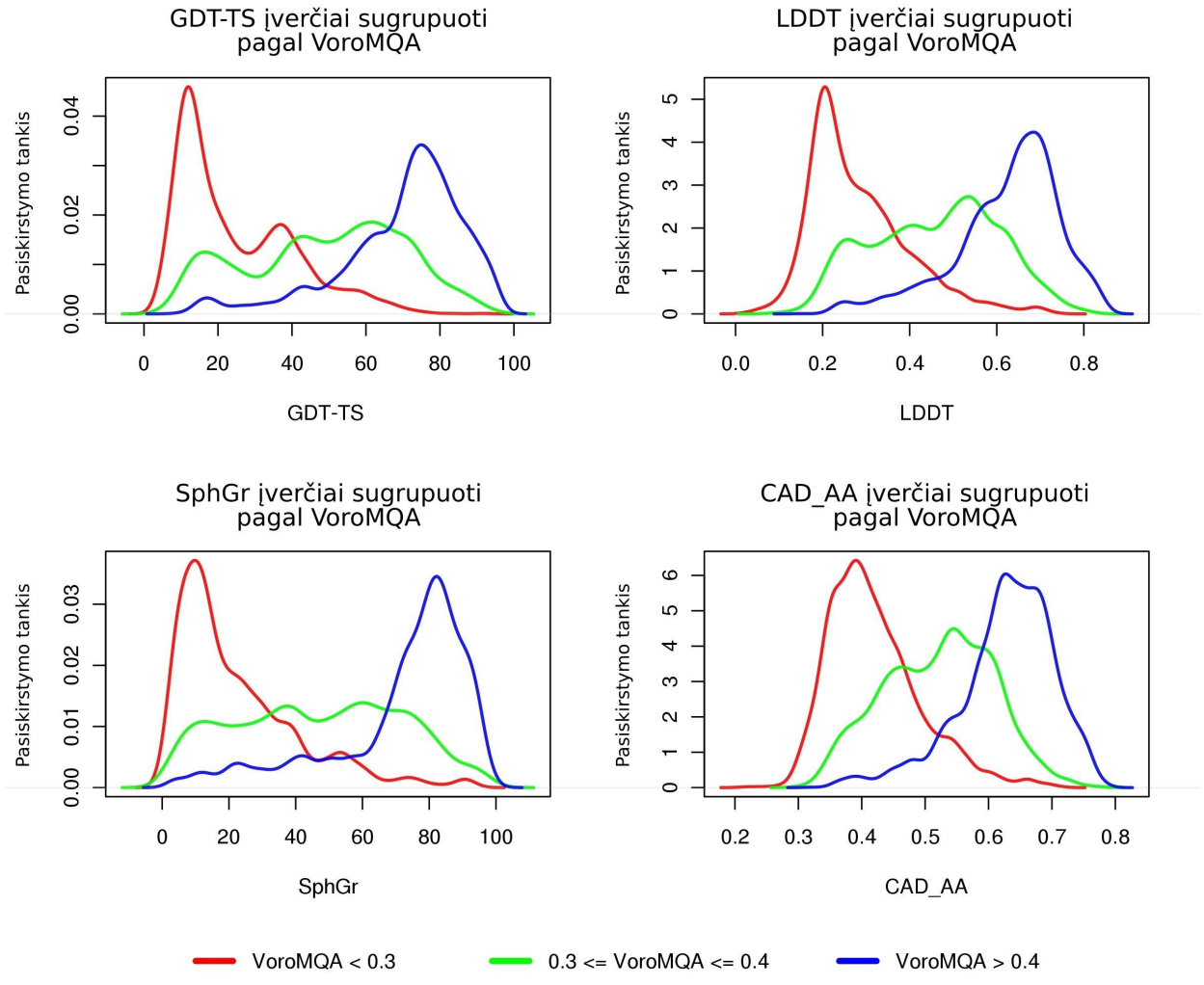


Pav. 3.3: (A) VoroMQA globaliojo įverčio empiriniai pasiskirstymo tankiai, kai struktūrose dominuoja įvairūs antrinės struktūros tipai. (B) VoroMQA globaliojo įverčio priklausomybė nuo struktūros dydžio. (C) VoroMQA veikimo laiko priklausomybė nuo struktūros dydžio (naudotas Intel® Xeon® E5-2670 v3 @ 2.30GHz procesorius).

Toliau metodą testavome naudodami paskutinių keturių CASP eksperimentų (CASP8-11) duomenis. Patikrinome, ar dažnai VoroMQA sugeba atpažinti taikinio struktūrą jo modelių aibėje; tam naudojome 140 monomerinių taikinių duomenis. Tokį testą atlikome ir taikydami keturis kitus metodus, paremtus statistiniais potencialais: DOOP,[47] GOAP,[48] dDFIRE[49] ir senesnę supaprastintą VoroMQA versiją (toliau ją vadinsime VoroMQA-old, o dabartinę versiją — VoroMQA-new arba tiesiog VoroMQA). VoroMQA metodas, nesugebėjęs atpažinti 8 taikinių iš 140, pasirodė geriau už kitus metodus.

Kitame testavimo etape naudojome CASP11[50] duomenis siekdami nustatyti, kaip VoroMQA globalieji įverčiai atitinka vertinimus, gautus naudojant keturis metodus, kurie lygina modelius su atitinkamais etalonais (taikiniais): GDT-TS,[36,51] LDDT,[52] SphGr (SphereGrinder)[53] ir CAD_AA (CAD-score A-A variantas).[4] Rezultatai, iliustruoti 3.4 pav., parodė, kad VoroMQA globaliuosius įverčius galima interpretuoti paprastai: jei VoroMQA>0,4, tai tikėtina, kad modelis panašus į taikinį (geras); jei VoroMQA<0,3, tai tikėtina, kad modelis nepanašus į taikinį (blogas).

Toliau patikrinome, kaip efektyviai mūsų metodas sugeba pasirinkti geriausią modelį iš dviejų. Tokiam testui ėmėme tik tas CASP11 modelių poras, kuriose vieną modelį geresniu už kitą vienbalsiai pripažino GDT-TS, LDDT, SphGr ir CAD_AA. Lentelėje 3.1 pateikti pagrindiniai rezultatai. VoroMQA rezultatus lyginome su prieš tai minėtų DOOP, GOAP, dDFIRE ir VoroMQA-old metodų rezultatais bei su rezultatais, kuriuos parodė CASP11 modelių kokybės vertin-



Pav. 3.4: GDT-TS, LDDT, SphGr ir CAD_AA įverčių pasiskirstymai kai VoroMQA-new įverčiai yra intervaluose $(0; 0, 3)$, $[0, 3; 0, 4]$ ir $(0, 4; 1)$.

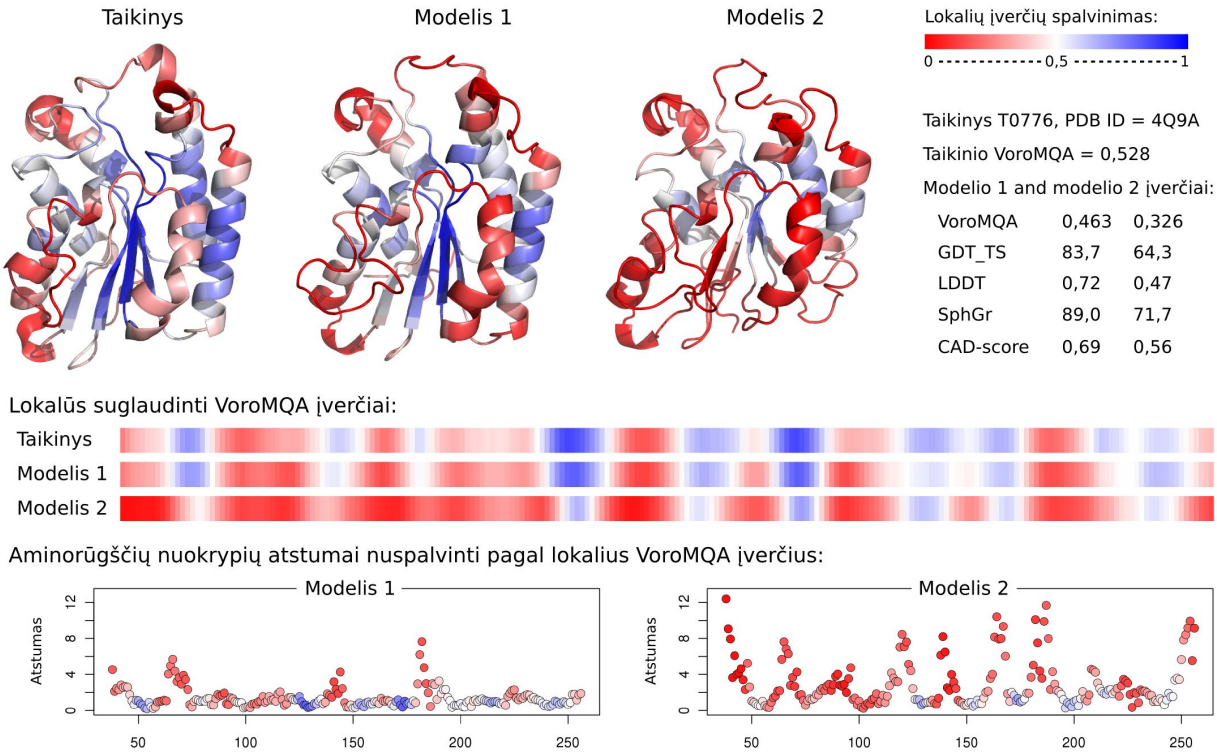
Metodas	Bendras %	Vidutinis %
VoroMQA-new-sr	82,50	82,70
VoroMQA-new	81,80	82,16
GOAP	80,11	80,57
VoroMQA-old	79,70	80,48
MULTICOM-NOVEL	79,66	80,25
ProQ2-refine	78,69	79,40
MULTICOM-CLUSTER	78,76	79,21
ProQ2	78,13	78,86
dDFIRE	77,73	78,43
DOOP	76,09	76,57
Wang_SVM	74,42	75,24
Wang_deep_2	72,12	72,83
Wang_deep_3	71,65	72,30
Wang_deep_1	71,57	72,19

Lent. 3.1: Geriausio modelio iš dviejų atpažinimo dažniai: visiems CASP11 modeliams kartu (antrasis stulpelis) ir vidutiniškai kiekvienam taikiniui (trečiasis stulpelis). Surikiuota pagal trečiąjį stulpelį.

imo kategorijoje dalyvavę metodai, kurie sugeba analizuoti pavienes struktūras, nors ir naudoja papildomą informaciją iš sekų duomenų bazių: MULTICOM-CLUSTER,[54] MULTICOM-NOVEL,[55] ProQ2,[56] ProQ2-refine,[57] Wang_SVM,[58] Wang_deep_{1,2,3}.[58] VoroMQA naudojome dviem režimais: nemodifikuojant įvesties struktūros ir pakeičiant jos šonines grandines SCWRL4 metodu.[59] Antrasis variantas (VoroMQA-new-sr) parodė šiek tiek geresnius rezultatus negu pirmasis (VoroMQA-new). Buvo atlikti ir testai, kuomet geriausią modelį reikėjo pasirinkti iš didesnių modelių aibių: jie detaliai aprašomi pagrindiniame disertacijos tekste; VoroMQA pasiekė geriausius rezultatus daugumoje iš jų.

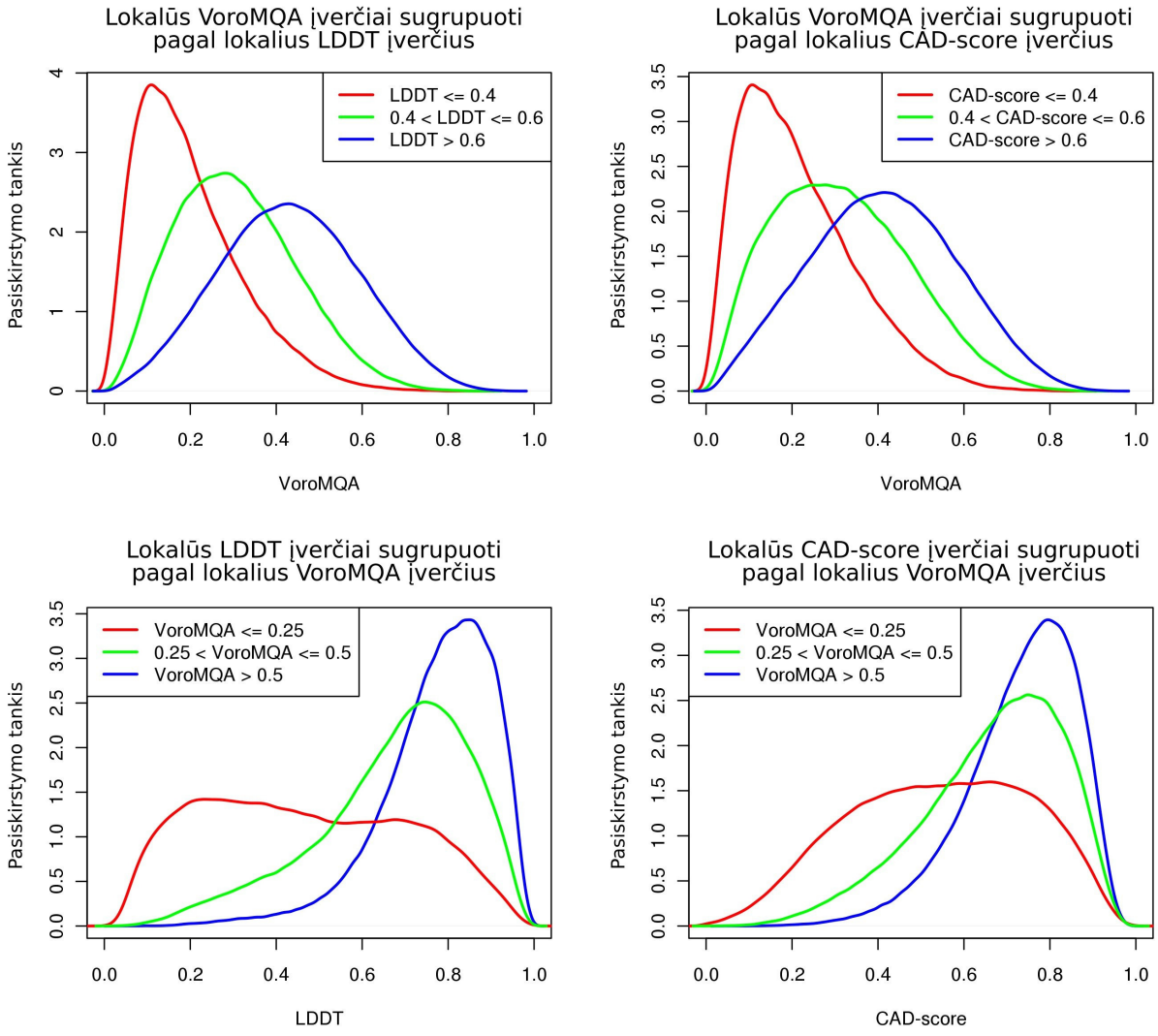
VoroMQA lokalieji įverčiai visų pirma skirti vizualiai struktūrų kokybės analizei (3.5 pav.). Supaprastintą jų interpretacijos taisyklę galima išvesti iš VoroMQA testavimo CAMEO[12] eksperimente rezultatų (3.6 pav.): jei VoroMQA lokalusis įvertis aukštas (didesnis už 0,5), tai atitinkamas regionas modelio struktūroje tikriausiai panašus į tos pozicijos regioną taikinyje; jei VoroMQA lokalusis įvertis nėra aukštas, tai dar nereiškia, kad atitinkamas regionas blogai sumodeliuotas (galimas paaiškinimas yra tas, kad ir realios struktūros nėra energetiškai homogeniškos, kai kurios jų dalys gali būti mažiau stabilesnės už kitas).

VoroMQA taip pat pritaikėme dalyvaudami 2016 metų CASP12 ir CAPRI eksperimentuose. Tarp 128 grupių, registruotų CASP12 struktūrų nusakymo



Pav. 3.5: VoroMQA taikymas trijų tos pačios sekos struktūrų lokaliajam vertinimui.

kategorijoje, mūsų grupė, pavadinta „VoroMQA-select“, buvo penkta (pagal pirmuosius modelius) ir antra (pagal geriausius modelius): šie rezultatai pateikti oficialiame CASP12 puslapyje `www.predictioncenter.org/casp12/zscores_final.cgi`. Tuo pačiu metu vykusiame CAPRI baltymų kompleksų prognozavimo eksperimente, derinant mūsų modelių vertinimo (VoroMQA) ir modeliavimo šablonų parinkimo (PPI3D) metodus mūsų grupei („Venclovas“) pavyko pasiekti geriausius rezultatus: oficialus vertinimas pateiktas adresu `http://predictioncenter.org/casp12/doc/presentations/CASP12_CAPRI_Lensink.pdf`.



Pav. 3.6: CAMEO vienerių metų duomenų apibendrinimas. Viršuje: VoroMQA lokaliųjų įverčių, sugrupuotų pagal atitinkamus LDDT ir CAD-score lokaliuosius įverčius, empiriniai pasiskirstymai. Apačioje: LDDT ir CAD-score lokaliųjų įverčių, sugrupuotų pagal atitinkamus VoroMQA lokaliuosius įverčius, empiriniai pasiskirstymai.

Išvados

Disertacijoje aprašyti trys nauji metodai, skirti biologinių makromolekulių struktūroms analizuoti ir vertinti. Pristatyti metodai konstruoja ir naudoja atomų rutulių Voronojaus diagramą. Tarpatominių kontaktų plotų, išvedamų iš Voronojaus diagramos, panaudojimas struktūrų analizei yra pagrindinis bruožas, skiriantis naujus metodus nuo tradicinių, aprašančių sąveikas remiantis atstumais. Žemiau pateikiamos išvados apie kiekvieną iš pristatytų metodų.

- Pirmasis metodas, Voronota, skirtas rutulių Voronojaus diagramos viršūnėms konstruoti. Jis ypač gerai tinka biologinių makromolekulių struktūroms apdoroti. Voronota algoritmas išnaudoja faktą, kad makromolekulėse absoliuti daugumą atomų rutulių trejetų turi dvi liestines plokštumas. Tai leidžia paprastai aprašyti trejeto kaimynų paieškos erdvę ir efektyviai panaudoti hierarchinį erdvinį indeksavimą trejeto kaimynams rasti. Trejetai be dviejų liestinių plokštumų yra labai reti makromolekulinėse struktūrose, todėl jie apdorojami paprasčiausiai perrenkant visus potencialius kaimynus; bendrą algoritmo veikimo laiką tai padidina nežymiai. Voronota metodo paprasta pirmojo priimtinojo trejeto paieškos procedūra leidžia paprastai lygiagretinti kaimynų paieškos algoritmą. Didelio masto testai parodė, kad Voronota yra greitas ir patikimas įrankis, tinkamas tiek eksperimentiškai nustatytoms, tiek nusakytoms (sumodeliuotoms kompiuteriais metodais) makromolekulių struktūroms analizuoti.
- Antrasis metodas, CAD-score, skirtas makromolekulių skirtingoms konformacijoms lyginti. Jo pagrindinė taikymo sritis — baltymų ir RNR modelių struktūrų vertinimas lyginant su etalonine struktūra. CAD-score veikia analizuodamas fizinius kontaktus, išvestus iš atomų rutulių Voronojaus diagramos, ir skaičiuodamas kontaktų plotų skirtumus. Metodas gali tiesiogiai vertinti tiek visą struktūrą, tiek sąveikas tarp jos grandinių ar domenų. Universali CAD-score prigimtis leidžia taikyti jį visų pagrindinių biologinių makromolekulių tipų (baltymų, nukleorūgščių ir jų kompleksų) struktūroms analizuoti. Išsamūs testai, atlikti naudojant baltymų ir RNR struktūrų modelius iš CASP ir RNA-puzzles modeliavimo eksperimentų, parodė, kad CAD-score yra patikimas struktūrų vertinimo metodas, turintis kelis esminius privalumus lyginant su tradiciškai naudojamais, struktūrų super-

pozicija paremtais metodais: CAD-score labiau atsižvelgia į modelių fizinį realistiškumą, gali efektyviai vertinti multidomenines struktūras, gali tiesiogiai vertinti tarpdomenines ir tarpgrandinines sąveikas. Papildomai, CAD-score suteikia galimybę efektyviai klasterizuoti makromolekulių struktūrinę informaciją: tuo buvo pasinaudota kuriant PPI3D metodą, skirtą baltymų sąveikų struktūrų paieškai ir analizei.

- Trečiasis metodas, VoroMQA, skirtas baltymų struktūrinių modelių kokybei vertinti (tikslumui nusakyti) nežinant etaloninės struktūros. VoroMQA remiasis empirinio statistinio potencialo idėja, bet, užuot tradiciškai naudojus sąveikų atstumus, yra naudojami Voronojaus diagramos pagrindu sukonstruotų tarpatominių kontaktų ir tirpikliui prieinamų paviršių plotai. VoroMQA nenaudoja papildomos išorinės informacijos, pavyzdžiui, antrinės struktūros ar tirpiklio prieinamumo nusakymo. VoroMQA išveda lokalius ir globalius įverčius intervale $(0, 1)$; šie įverčiai daugiausia nepriklauso nuo struktūros tipo ar dydžio. VoroMQA galima naudoti ne tik pasirenkant geriausią modelį, bet ir modelį priskiriant realistiškų ar nerealistiškų struktūrų klasei. VoroMQA lokalūs įverčiai gali būti naudojami patikimiems regionams modelio struktūroje išskirti. Testai, atlikti su CASP8-CASP11 duomenimis, parodė, kad VoroMQA dažniausiai veikia geriau negu kiti statistiniais potencialais paremti metodai. Be to, VoroMQA dažnai pranoksta ir tuos metodus, kuriuose naudojama papildoma informacija. VoroMQA pagrindu sukurtas modelių parinkimo protokolas buvo akluoju būdu ištestuotas per 2016 metų CASP12 eksperimentą: pasiekti aukšti rezultatai, kurių nepralenkė kiti metodai, taip pat pagrįsti automatinių serverių modelių atranka. VoroMQA taip pat buvo svarbus pasiekiant geriausius rezultatus baltymų kompleksų struktūrų modeliavimo eksperimente CAPRI 2016 metais.

Apibendrinant, svarbiausia disertacijos išvada yra ta, kad atomų Voronojaus diagramos pagrindu sukonstruoti kontaktai ir jų plotai atspindi svarbias biologinių makromolekulių ypatybes ir gali būti sėkmingai naudojami kaip pagrindas naujiems efektyviems baltymų ir nukleorūgščių struktūrų analizės ir vertinimo metodams kurti.

Bibliografinės nuorodos

[1] J. Kuriyan, B. Konforti, and D. Wemmer, *The molecules of life: physical and chemical principles*. New York, NY: GS, Garland Science, 2013. OCLC: 779577263.

[2] R. B. Altman and J. M. Dugan, "Defining Bioinformatics and Structural Bioinformatics," in *Methods of Biochemical Analysis* (P. E. Bourne and H. Weissig, eds.), pp. 1–14, Hoboken, NJ, USA: John Wiley & Sons, Inc., Jan. 2005.

[3] A. Tramontano, *Protein structure prediction: concepts and applications*. Weinheim: Wiley-VCH, 2006. OCLC: 181462508.

[4] K. Olechnovič, E. Kulberkytė, and C. Venclovas, "CAD-score: a new contact area difference-based function for evaluation of protein structural models," *Proteins*, vol. 81, pp. 149–162, Jan. 2013.

[5] K. Olechnovič and C. Venclovas, "Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls," *J. Comput. Chem.*, vol. 35, pp. 672–681, Mar. 2014.

[6] K. Olechnovič and C. Venclovas, "The use of interatomic contact areas to quantify discrepancies between RNA 3D models and reference structures," *Nucleic Acids Res.*, vol. 42, pp. 5407–5415, May 2014.

[7] K. Olechnovič and C. Venclovas, "The CAD-score web server: contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes," *Nucleic Acids Res.*, vol. 42, pp. W259–263, July 2014.

[8] K. Olechnovic and C. Venclovas, "VoroMQA: Assessment of protein structure quality using interatomic contact areas," *Proteins*, Mar. 2017.

[9] J. Dapkunas, A. Timinskas, K. Olechnovic, M. Margelevicius, R. Diciunas, and C. Venclovas, "The PPI3D web server for searching, analyzing and modeling protein–protein interactions in the context of 3D structures," *Bioinformatics*, vol. 33, pp. 935–937, Mar. 2017.

[10] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)–round x," *Proteins*, vol. 82 Suppl 2, pp. 1–6, Feb. 2014.

[11] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction: Progress and new directions in round XI," *Proteins*, vol. 84 Suppl 1, pp. 4–14, Sept. 2016.

[12] J. Haas, S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli, and T. Schwede, "The Protein Model Portal–a comprehensive resource for protein structure and model information," *Database*, vol. 2013, p. bat031, 2013.

[13] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Wiley Series in Probability and Statistics, Wiley, 2000.

[14] A. Goede, R. Preissner, and C. Frömmel, "Voronoi cell: new method for allocation of space among atoms: elimination of avoidable errors in calculation of atomic volume and density," *J. Comput. Chem.*, vol. 18, pp. 1113–1123, 1997.

[15] M. L. Gavrilova and J. Rokne, "Updating the topology of the dynamic Voronoi diagram for spheres in Euclidean d-dimensional space," *Comput. Aided Geom. D.*, vol. 20, pp. 231–242, 2003.

[16] P. Su and R. L. S. Drysdale, "A comparison of sequential Delaunay triangulation algorithms," *Comput. Geom.*, vol. 7, pp. 361–385, 1997.

[17] T. Masaharu, T. Ogawa, and N. Ogita, "A new algorithm for three-dimensional voronoi tessellation," *J. Comput. Phys.*, vol. 51, pp. 191–207, 1983.

[18] A. Maus, "Delaunay triangulation and the convex hull ofn points in expected linear time," *BIT Numer. Math.*, vol. 24, pp. 151–163, 1984.

[19] D.-S. Kim, Y. Cho, and D. Kim, "Euclidean Voronoi diagram of 3D balls and its computation via tracing edges," *Comput. Aided Design*, vol. 37, pp. 1412–1424, Nov. 2005.

[20] N. N. Medvedev, V. P. Voloshin, V. A. Luchnikov, and M. L. Gavrilova, "An algorithm for three-dimensional Voronoi S-network," *J. Comput. Chem.*, vol. 27, pp. 1676–1692, 2006.

[21] D. S. Kim, Y. Cho, and K. Sugihara, "Quasi-worlds and quasi-operators on quasi-triangulations," *Comput. Aided Design*, vol. 42, pp. 874–888, 2010.

[22] J. Spillmann, M. Becker, and M. Teschner, "Efficient updates of bounding sphere hierarchies for geometrically deformable models," *J. Vis. Commun. Image R.*, vol. 18, pp. 101–108, 2007.

[23] J. L. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching," *Commun. ACM*, vol. 18, pp. 509–517, 1975.

[24] W. Degen, "Cyclides," in *Handbook of computer aided geometric design*, pp. 575–601, Elsevier, 2002.

[25] W. Kahan, "Pracniques: Further Remarks on Reducing Truncation Errors," *Commun. ACM*, vol. 8, p. 40, 1965.

[26] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, 2002.

[27] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235–242, Jan. 2000.

[28] J. Moult, K. Fidelis, A. Kryshtafovych, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)–round IX," *Proteins*, vol. 79 Suppl 10, pp. 1–5, 2011.

[29] E. Capriotti, T. Norambuena, M. A. Marti-Renom, and F. Melo, "All-atom knowledge-based potential for RNA structure prediction and assessment," *Bioinformatics*, vol. 27, pp. 1086–1093, Apr. 2011.

[30] A. J. Li and R. Nussinov, "A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking," *Proteins*, vol. 32, pp. 111–127, July 1998.

[31] B. J. McConkey, V. Sobolev, and M. Edelman, "Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure," *Bioinformatics*, vol. 18, pp. 1365–1373, Oct. 2002.

[32] A. Dietrich and B. Maigret, "Program for the visualization of inorganic crystals," *J. Mol. Graph.*, vol. 9, pp. 85–90, 97–99, June 1991.

[33] K. Olechnovic, M. Margelevicius, and C. Venclovas, "Voroprot: an interactive tool for the analysis and visualization of complex geometric features of protein structure," *Bioinformatics*, vol. 27, pp. 723–724, Mar. 2011.

[34] A. Kryshtafovych, J. Moult, S. G. Bartual, J. F. Bazan, H. Berman, D. E. Casteel, E. Christodoulou, J. K. Everett, J. Hausmann, T. Heidebrecht, T. Hills, R. Hui, J. F. Hunt, J. Seetharaman, A. Joachimiak, M. A. Kennedy, C. Kim, A. Lingel, K. Michalska, G. T. Montelione, J. M. Otero, A. Perrakis, J. C. Pizarro, M. J. van Raaij, T. A. Ramelot, F. Rousseau, L. Tong, A. K. Wernimont, J. Young, and T. Schwede, "Target highlights in CASP9: Experimental target structures for the critical assessment of techniques for protein structure prediction," *Proteins*, vol. 79 Suppl 10, pp. 6–20, 2011.

[35] A. Tramontano, D. Cozzetto, A. Giorgetti, and D. Raimondo, "The assessment of methods for protein structure prediction," *Methods Mol. Biol.*, vol. 413, pp. 43–57, 2008.

[36] A. Zemla, C. Venclovas, J. Moult, and K. Fidelis, "Processing and analysis of CASP3 protein structure predictions," *Proteins*, vol. Suppl 3, pp. 22–29, 1999.

[37] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography," *Acta Crystallogr. D Biol. Crystallogr.*, vol. 66, pp. 12–21, Jan. 2010.

[38] P. Gendron, S. Lemieux, and F. Major, "Quantitative analysis of nucleic acid three-dimensional structures," *J. Mol. Biol.*, vol. 308, pp. 919–936, May 2001.

[39] J. A. Cruz, M.-F. Blanchet, M. Boniecki, J. M. Bujnicki, S.-J. Chen, S. Cao, R. Das, F. Ding, N. V. Dokholyan, S. C. Flores, L. Huang, C. A. Lavender, V. Lisi, F. Major, K. Mikolajczak, D. J. Patel, A. Philips, T. Puton, J. Santalucia, F. Sijenyi, T. Hermann, K. Rother, M. Rother, A. Serganov, M. Skorupski, T. Soltysinski, P. Sripakdeevong, I. Tuszynska, K. M. Weeks, C. Waldsich, M. Wildauer, N. B. Leontis, and E. Westhof, "RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction," *RNA*, vol. 18, pp. 610–625, Apr. 2012.

[40] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallogr. Sect. A*, vol. 32, pp. 922–923, Sept. 1976.

[41] M. Parisien, J. A. Cruz, E. Westhof, and F. Major, "New metrics for comparing and assessing discrepancies between RNA 3D structures and models," *RNA*, vol. 15, pp. 1875–1885, Oct. 2009.

[42] D. Butina, "Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets," *Journal of Chemical Information and Computer Sciences*, vol. 39, pp. 747–750, July 1999.

[43] K. Katoh, K.-i. Kuma, H. Toh, and T. Miyata, "MAFFT version 5: improvement in accuracy of multiple sequence alignment," *Nucleic Acids Res.*, vol. 33, no. 2, pp. 511–518, 2005.

[44] M. J. Sippl, "Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins," *J. Mol. Biol.*, vol. 213, pp. 859–883, June 1990.

[45] M. J. Sippl, "Recognition of errors in three-dimensional structures of proteins," *Proteins*, vol. 17, pp. 355–362, Dec. 1993.

[46] B. J. McConkey, V. Sobolev, and M. Edelman, "Discrimination of native protein structures using atom-atom contact scoring," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, pp. 3215–3220, Mar. 2003.

[47] M.-H. Chae, F. Krull, and E.-W. Knapp, "Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction," *Proteins*, vol. 83, pp. 881–890, May 2015.

[48] H. Zhou and J. Skolnick, "GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction," *Biophys. J.*, vol. 101, pp. 2043–2052, Oct. 2011.

[49] Y. Yang and Y. Zhou, "Specific interactions for ab initio folding of protein terminal regions with secondary structures," *Proteins*, vol. 72, pp. 793–803, Aug. 2008.

[50] A. Kryshtafovych, A. Barbato, B. Monastyrskyy, K. Fidelis, T. Schwede, and A. Tramontano, "Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11," *Proteins*, vol. 84 Suppl 1, pp. 349–369, Sept. 2016.

[51] A. Zemla, "LGA: A method for finding 3D similarities in protein structures," *Nucleic Acids Res.*, vol. 31, pp. 3370–3374, July 2003.

[52] V. Mariani, M. Biasini, A. Barbato, and T. Schwede, "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests," *Bioinformatics*, vol. 29, pp. 2722–2728, Nov. 2013.

[53] P. L. M. Antczak, T. Ratajczak, J. Blazewicz, P. Lukasiak, and J. Blazewicz, "SphereGrinder - reference structure-based tool for quality assessment of protein structural models," pp. 665–668, IEEE, Nov. 2015.

[54] J. Li, R. Cao, and J. Cheng, "A large-scale conformation sampling and evaluation server for protein tertiary structure prediction and its assessment in CASP11," *BMC Bioinformatics*, vol. 16, p. 337, 2015.

[55] R. Cao and J. Cheng, "Protein single-model quality assessment by feature-based probability density functions," *Sci. Rep.*, vol. 6, p. 23990, 2016.

[56] A. Ray, E. Lindahl, and B. Wallner, "Improved model quality assessment using ProQ2," *BMC Bioinformatics*, vol. 13, p. 224, 2012.

[57] K. Uziela and B. Wallner, "ProQ2: estimation of model accuracy implemented in Rosetta," *Bioinformatics*, vol. 32, pp. 1411–1413, May 2016.

[58] T. Liu, Y. Wang, J. Eickholt, and Z. Wang, "Benchmarking Deep Networks for Predicting Residue-Specific Quality of Individual Protein Models in CASP11," *Sci. Rep.*, vol. 6, p. 19301, 2016.

[59] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack, "Improved prediction of protein side-chain conformations with SCWRL4," *Proteins*, vol. 77, pp. 778–795, Dec. 2009.

Autoriaus publikacijų sąrašas

Straipsniai disertacijos tema

1. Kliment Olechnovič, Eleonora Kulberkytė and Česlovas Venclovas. *CAD-score: a new contact area difference-based function for evaluation of protein structural models*. Proteins (2013) 81 (1): 149-162.

2. Kliment Olechnovič and Česlovas Venclovas. *Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls*. J Comput Chem (2014) 35 (8): 672-681.

3. Kliment Olechnovič and Česlovas Venclovas. *The use of interatomic contact areas to quantify discrepancies between RNA 3D models and reference structures*. Nucl Acids Res (2014) 42 (9): 5407-5415.

4. Kliment Olechnovič and Česlovas Venclovas. *The CAD-score web server: contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes*. Nucleic Acids Res (2014) 42 (W1): 259-263. *NAR Breakthrough Article*.

5. Justas Dapkūnas, Albertas Timinskas, Kliment Olechnovič, Mindaugas Margelevičius, Rytis Dičiūnas and Česlovas Venclovas. *The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures*. Bioinformatics (2017) 33 (6): 935-937.

6. Kliment Olechnovič and Česlovas Venclovas. *VoroMQA: Assessment of protein structure quality using interatomic contact areas*. Proteins (2017) 10.1002/prot.25278.

Kiti straipsniai

1. Kliment Olechnovič, Mindaugas Margelevičius and Česlovas Venclovas. *Voroprot: an interactive tool for the analysis and visualization of complex geometric features of protein structure*. Bioinformatics (2011) 27 (5): 723-724.

Tarptautinės konferencijos 2012–2016

1. „EMBO Conference on Critical Assessment of Protein Structure Prediction“ (Gaeta, Italija, 2012.12.9-12). Stendinis pranešimas: *CAD-score: a new method for the evaluation of protein structural models*. Apdovanotas už geriausią stendinį pranešimą, išrinktas žodiniam pranešimui.

2. „SocBiN: Society for Bioinformatics in Northern European countries“ (Torunė, Lenkija, 2013.06.26-29). Stendinis pranešimas: *The use of interatomic contact areas for the assessment of RNA 3D structural models*. Apdovanotas už geriausią stendinį pranešimą.

3. „Intelligent Systems for Molecular Biology“ (Berlynas, Vokietija, 2013.07.19-23). Stendinis pranešimas: *The use of interatomic contact areas for the assessment of RNA 3D structural models*.

4. „European Conference on Computational biology“ (Strasbūras, Prancūzija, 2014.09.7-10). Stendinis pranešimas: *The CAD-score webserver: contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes*.

5. „11$^{\text{th}}$ Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction“ (Riviera Maja, Meksika, 2014.12.7-10). Stendinis pranešimas: *Quality assessment of single protein structure models using inter-atom contact areas derived from the Voronoi diagram of atomic balls*.

6. „European Conference on Computational biology“ (Haga, Niderlandai, 2016.09.3-7). Stendinis pranešimas: *Estimation of protein structure quality using contact areas derived from the Voronoi tessellation of atomic balls*.

7. „12$^{\text{th}}$ Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction“ (Gaeta, Italija, 2016.12.10-13). Stendinis pranešimas: *VoroMQA: assessment of protein structure quality using interatomic contact areas derived from the Voronoi tessellation of atomic balls*. Apdovanotas už geriausią stendinį pranešimą, išrinktas žodiniam pranešimui.

Autoriaus gyvenimo aprašymas

Asmeninė informacija

Vardas ir pavardė:	Kliment Olechnovič
Gimimo data ir vieta:	1987.08.18, Vilnius, Lietuva
El. paštas:	`kliment@ibt.lt`

Išsilavinimas

2012 – 2016	informatikos doktorantas, Vilniaus universitetas.
2010 – 2012	informatikos magistras, *Magna Cum Laude*, Vilniaus universitetas.
2005 – 2009	informatikos bakalauras (bioinformatikos specialybė), Vilniaus universitetas.
1993 – 2005	L. Karsavino vidurinė mokykla, Vilnius.

Darbo patirtis

2016 – dabar	Jaunesnysis mokslo darbuotojas, Vilniaus universiteto Biotechnologijos instituto Bioinformatikos skyrius.
2010 – 2016	Inžinierius tyrėjas, Vilniaus universiteto Biotechnologijos instituto Bioinformatikos skyrius.
2009 – 2010	Laborantas, Biotechnologijos instituto Bioinformatikos skyrius.
2007 – 2009	C++ programuotojas, 4Team Corporation, Vilnius.

Publikacijos

7 straipsniai recenzuojamuose žurnaluose, įtrauktuose į Clarivate Analytics Web of Science duomenų bazę.

12 pranešimų tarptautinėse konferencijose (3 apdovanojimai už geriausią stendinį pranešimą).

Apdovanojimai

2015	Lietuvos Mokslų Akademijos apdovanojimas už geriausius jaunųjų mokslininkų darbus 2014 metais.
2013–2015	Lietuvos Mokslo Tarybos stipendiją doktorantams už aktyvią mokslinę veiklą.
2013	INFOBALT skatinamoji stipendija jauniesiems mokslininkams.

Santrauka anglų kalba (Abstract)

This dissertation describes three novel effective methods for the analysis and evaluation of biomolecular structures. The presented methods construct and utilize the Voronoi tessellation of atomic balls. The usage of the tessellation-derived interatomic contact areas to analyze structural models is the main feature that sets the presented methods apart from the traditional distance-based structure analysis methods. The first method, Voronota, is a method for computing the vertices of the Voronoi diagram of balls. It is capable of processing macromolecular structures efficiently by exploiting common patterns of atomic spatial arrangements. Voronota serves as an effective tool for defining interatomic interactions, it is also easily parallelizable. The second method, CAD-score (Contact Area Difference Score), is a method for the comparison of different conformations of macromolecules, for example, native and modeled structures. It uses Voronoi tessellation-derived contact areas to avoid common problems of traditionally used reference-based assessment methods. CAD-score is universally applicable for the comparison of structures of all the major types of macromolecules (proteins, nucleic acids and their complexes). The third method, VoroMQA (Voronoi diagram-based Model Quality Assessment), is a method for the evaluation of predicted protein structures when the native structure is unknown. It efficiently combines the idea of knowledge-based statistical potential with the concept of interatomic contact areas derived from the Voronoi tessellation of atomic balls. VoroMQA consistently outperforms other statistical potential-based protein structure quality assessment methods. The main conclusion of the presented studies is that Voronoi tessellation-derived contact areas capture important structural features of biological macromolecules and are useful as a foundation for new effective methods for the analysis and assessment of three-dimensional structures of proteins and nucleic acids.

Kliment Olechnovič

BALTYMŲ IR NUKLEORŪGŠČIŲ ERDVINIŲ STRUKTŪRŲ
ANALIZĖS IR VERTINIMO METODAI: KŪRIMAS IR TAIKYMAS

Daktaro disertacijos santrauka
Fiziniai mokslai, informatika (09P)
Redaktorė Rūta Šiaučiulytė

Kliment Olechnovič

METHODS FOR THE ANALYSIS AND ASSESSMENT OF THE
THREE-DIMENSIONAL STRUCTURES OF PROTEINS
AND NUCLEIC ACIDS: DEVELOPMENT AND APPLICATIONS

Summary of doctoral dissertation
Physical sciences, informatics (09P)
Editor Neringa Slapikevičiūtė