LATVIJAS
UNIVERSITĀTE

Promocijas darba
kopsavilkums

Gatis Melkus

UZ GRAFIEM BALSTĪTAS METODES BIOMOLEKULĀRO DATU MODELĒŠANAI

Rīga 2025

LATVIJAS
UNIVERSITĀTE

EKSAKTO ZINĀTŅU UN TEHNOLOĢIJU FAKULTĀTE

Gatis Melkus

UZ GRAFIEM BALSTĪTAS METODES BIOMOLEKULĀRO DATU MODELĒŠANAI

Promocijas darba kopsavilkums

Iesniegts inženierzinātņu un tehnoloģiju doktora grāda iegūšanai
Nozare: datorzinātne un informātika
Apakšnozare: bioinformātika

Rīga 2025

Promocijas darbs izstrādāts Latvijas Universitātes Eksakto zinātņu un tehnoloģiju fakultātē Datorzinātņu katedrā laika posmā no 2019. gada līdz 2025. gadam

INVESTING IN YOUR FUTURE

Eiropas Sociālā fonda projekts "LU doktorantūras kapacitātes stiprināšana jaunā doktorantūras modeļa ietvarā" Nr. 8.2.2.0/20/I/006

Darbs sastāv no ievada, 3 sadaļām (tostarp 5 zinātniskām publikācijām), nobeiguma un literatūras saraksta.

Darba forma: publikāciju kopa datorzinātnes nozarē, bioinformātikas apakšnozarē.

Darba zinātniskais vadītājs: *Dr. sc. comp.* **Juris Vīksna**, prof.

Recenzenti:

1. *Dr. sc. comp.* **Ģirts Karnītis**, prof., Latvijas Universitāte;
2. *Dr. sc. ing.* **Egils Stalidzāns**, prof., Rīgas Stradiņa universitāte;
3. *Dr. sc. comp.* **Alvis Brāzma**, European Bioinformatics Institute (EMBL-EBI), Scientist Emeritus.

Promocijas darba aizstāvēšana notiks 2025. gada 24. oktobrī Latvijas Universitātes Datorzinātnes un informātikas un Elektrotehnikas, elektronikas un komunikāciju tehnoloģiju nozares promocijas padomes atklātā sēdē.

Ar promocijas darbu un tā kopsavilkumu var iepazīties Latvijas Universitātes Bibliotēkā Rīgā, Kalpaka bulvārī 4.

Promocijas padomes priekšsēdētajs *Dr. sc. comp.*, prof. **Guntis Bārzdiņš**

Promocijas padomes sekretāre **Sintija Siliņa**

ISBN 978-9934-36-420-4
ISBN 978-9934-36-421-1 (PDF)

ANOTĀCIJA

Biomolekulāro datu analīze un interpretācija ir aktuāls temats bioloģijas un arī citu zinātņu kontekstā, jo tajā iekļaujas vairākas plašākas problēmas, tostarp lieldatu apsaimniekošana un analīze, kā arī ilgtspējīgas zinātnes prakse. Bioloģisko datu kopas kļūst arvien sarežģītākas gan izmēru, gan metodoloģisko variāciju, gan salīdzināmības un savietojamības ziņā. Šajā disertācijā šo problēmu risināšanā pielietota grafu teorija tās bioloģiskajā kontekstā – tīklu bioloģija –, lai analizētu biomolekulārus datus un izstrādātu jaunas metodes to analīzei. Disertācija ir noformēta kā piecu zinātnisku publikāciju krājums, no kurām viena publikācija ir salīdzinoši vecākajā un plašāk pētītajā gēnu regulatoro tīklu modelēšanas jomā, bet atlikušās četras ir veltītas hromatīna mijiedarbību tīklu veidošanai no augstas caurlaides hromatīna konformācijas analīžu datiem, kas ir nesenāka un mazāk pētīta nozare tīklu bioloģijā. Attiecīgi disertācijas tematikai pievienots arī ir īss literatūras apskats par gēnu regulatoro tīklu un hromatīna arhitektūras modelēšanas galvenajiem jēdzieniem, lai sniegtu nepieciešamo kontekstu disertācijas laikā paveiktajam darbam. Iegūtie rezultāti liecina par plašu grafu teorijas pamatu pielietojamību hromatīna arhitektūras modelēšanā, ar zīmīgiem panākumiem gan hromatīna mijiedarbību datu analīzē, gan to interpretācijā. Publikācijās arī demonstrētas vairākas jaunas metodes gēnu regulatoro tīklu un hromatīna mijiedarbību tīklu modelēšanai, kā arī plašāka metožu kopa hromatīna mijiedarbību datu topoloģijas salīdzināšanai ar citām bioloģisko datu kopām.

Atslēgvārdi: tīkla bioloģija, gēnu regulatorie tīkli, Hi-C, tīklu topoloģija, integratīvā bioloģija

SATURS

IEVADS

Pētījuma aktualitāte un novitāte

Grafu teorijā balstītu metodoloģiju ieviešana molekulārās bioloģijas datu analīzē 21. gs. sākumā ir tīklu bioloģijas (*network biology*) disciplīnas sākotne. Tīklu bioloģijai ir raksturīgs uzsvars uz biomolekulu mijiedarbību analīzi, apvienojot plašu mijiedarbību klāstu lielākos tīklos, lai sistemātiski pētītu to struktūras pamatprincipus, atšķirībā no sistēmbioloģijas (*systems biology*), kur lielāks uzsvars ir uz šo tīklu funkcionālo dabu. Matemātiskā izpratnē šie tīkli (*networks*) ir grafi, ar kuriem var modelēt lielas mijiedarbību datu kopas, piemēram, proteīnu-proteīnu saistību, ģenētiskas mijiedarbības, gēnu regulāciju un hromatīna arhitektūru. Ar grafu palīdzību šādus mijiedarbību tīklus var pētīt ar dažādām pieejām, pielietojot grafu topoloģiju un līdzīgus jēdzienus, lai veidotu lielizmēra modeļus, kas potenciāli varētu izskaidrot plašākas cēloņsakarības bioloģiskās sistēmās (Barabási and Oltvai 2004; Sorrells and Johnson 2015).

Molekulārās bioloģijas pētījumi daudzu gadu gaitā ir radījuši plašu un daudzveidīgu lielizmēra datu kopu klāstu, kas ir publiski pieejamas un brīvi izmantojamas tīklu modeļu izstrādē, pārbaudē un attīstīšanā. Šajā klāstā ir ne tikai mijiedarbību un regulāciju dati, kas ir tiešā veidā izmantojami tīklu veidošanā, bet arī genoma sekvenču dati, eksperimentu apkopojumi un informatīvas datubāzes, kuras var izmantot šo modeļu validācijai un precizēšanai. Protams, šie dati visbiežāk ir heterogēni gan pēc to izpētes metodoloģijas, izcelsmes un kvalitātes, kas lielā mērā ir skaidrojams ar tīklu bioloģijas disciplīnas salīdzinoši īso vēsturi, bet diemžēl bieži vien nozīmē, ka tīklu bioloģijas metodēm ir grūti izvirzīt galīgus secinājumus par to efektivitāti un optimizēšanas iespējām. Šāda situācija ir novērojama, piemēram, gēnu regulatoro tīklu gadījumā (Sorrells and Johnson 2015), bet ir īpaši izplatīta tādās mazāk pētītās jomās kā hromatīna telpiskās organizācijas pētījumos, kur datu integrācija un validācija joprojām ir atvērts jautājums un kur jaunu metožu izstrāde var sniegt ievērojamu devumu nozares attīstībā un bioloģisku secinājumu gūšanā (Pancaldi 2021; 2023).

Ņemot vērā, ka tīklu bioloģijā pieejamais datu klāsts ir ļoti plašs un galīgu secinājumu par tīklu struktūras pamatprincipiem ir salīdzinoši maz, ir būtiski izstrādāt jaunas metodoloģijas grafu veidošanai no biomolekulāriem datiem, kas varētu sniegt uzticamus rezultātus. Īpaši nozīmīga ir metožu izstrāde, kas varētu būt noderīga integratīvai datu analīzei, kas varētu apkopot ne tikai kombinētus multiomikas eksperimentus, bet arī individuālu šūnu eksperimentu datus (Vandereyken et al. 2023), veidojot plašāku izpratni par šo datu būtību un variācijām dažādos bioloģiskos objektos.

Darba mērķis un uzdevumi

Šī promocijas darba mērķis bija izstrādāt jaunas metodes biomolekulāru datu analīzei ar grafu teorijā balstītu metožu palīdzību, īpašu vērību veltot grafu topoloģiskām metodēm. Lai šo mērķi sasniegtu, bija jāizpilda šādi uzdevumi:

1. Jāpaplašina un jāpielāgo esošas topoloģiskas metodes tīkla motīvu un grafletu pētījumiem, lai pētītu grafu topoloģijas līdzību starp gēnu pāriem regulatora tīkla ietvaros.
2. Jāveido jaunas topoloģiskas pieejas hromatīna mijiedarbību tīklu pētīšanai ar saistības komponenšu un no tām atvasinātu elementu palīdzību.
3. Jāizmanto grafu topoloģijas elementi, lai hromatīna mijiedarbību tīklos atrastu funkcionāli nozīmīgus apakšgrafus, piemēram, gēnu regulācijas moduļus.
4. Jāpiesaista papildus biomolekulāri dati, lai veiktu rezultātu validāciju un pārliecinātos par modeļos novēroto sakarību atbilstību bioloģiskai realitātei.
5. Attīstot pielietotās topoloģiskās metodes, jāatrod efektīvākais topoloģiskais kritērijs funkcionāli nozīmīgu apakšgrafu atrašanai hromatīna mijiedarbību tīklā.

Galvenās tēzes

Šī darba ietvaros tika pārbaudīta hipotēze, ka biomolekulāros datos ir iespējams atrast bioloģiskas īpašības, vadoties tikai pēc topoloģiskiem grafu teorētiskiem kritērijiem. Šie kritēriji var būt, piemēram, noteiktu tīkla motīvu klātbūtne vai dažādu topoloģiski definētu apakšgrafu (piemēram, kliķu) skaitu atšķirības. Šīs hipotēzes pārbaudes ietvaros darba galvenos secinājumus var aprakstīt ar šīm tēzēm:

- *Bi-fan* motīvi, ja tos vispārina kā *bi-fan* vienības, ir vienkārši pielietojami un pielāgojami rīki gēnu regulatoro tīklu simetrijas pētījumiem starp gēnu pāriem dažādos organismos, un tie ir īpaši piemēroti paralogu gēnu izpētei.
- Augstas caurlaides hromatīna konformācijas notveres (Hi-C) dati ir topoloģiski raksturojami ar saistības komponenšu palīdzību, jo tās ļauj izšķirt un analizēt funkcionālas grupas bez papildus bioloģiskas informācijas.
- Pielietojot šaurāk definētus topoloģiskus elementus, hromatīna mijiedarbību datos ir iespējams nošķirt funkcionāli nozīmīgus apakšgrafus, kuros ir biežāk sastopamas aktīvas gēnu ekspresijas iezīmes.
- Grafu balstīti modeļi vienkāršo informācijas apkopošanu un integrāciju, jo tajos var efektīvi kombinēt dažādas iezīmes, piemēram, pozicionālu gēnu ekspresiju, hromatīna anotācijas un līdzīga veida informāciju.
- No apskatītajiem grafu topoloģiskajiem elementiem visnoderīgākās ir kliķes, kas hromatīna mijiedarbību datos atbilst hromatīna sakopojumiem, kuros notiek intensīva gēnu regulācija, un šo gēnu regulāciju ir vērtīgi turpmāk analizēt.

Literatūras apskats

Šī promocijas darba ietvaros tika pētīti tīklu motīvi, kas vēsturiski pētīti gēnu regulatorajos tīklos, kā arī hromatīnu mijiedarbību tīkli. Abos gadījumos pētījumu pamatmetodes ir topoloģiskas, lai gan gēnu regulatoro tīklu topoloģija literatūrā ir plašāk pētīta.

Lai gan tīklu motīvi bioloģijā tika izmantoti jau senāk ekoloģijas pētījumos, biomolekulu mijiedarbību aprakstīšana ar motīvu palīdzību sākās ar Alona laboratorijas pētījumiem 21. gs. sākumā (Babu et al. 2004; Alon 2007; Stone, Simberloff, and Artzy-Randrup 2019). Tīklu motīvi sākotnēji bija definēti kā lielāka gēnu regulatorā tīkla (tīkla, kas sastāv no transkripcijas faktoru gēniem un tālākiem gēniem, ko tie regulē) apakšgrafi, kas šajā tīklā bija sastopami daudz biežāk nekā matemātiski modelētā tīklā ar līdzīgām topoloģiskām īpašībām. Šiem motīviem, pamatojoties uz to struktūru, pētnieki piedēvēja dažādas hipotētiskas funkcijas, piemēram, atbildes reakcijas palēninášanu vai pielāgošanu. Kopš tīkla motīvu (un tīklu bioloģijas) pētījumu pirmsākumiem daudzas no sākotnējām hipotēzēm (piemēram, *scale-free* tīklu izplatība bioloģiskos objektos) nav apstiprinājušās praksē (Lima-Mendez and van Helden 2009; Broido and Clauset 2019), tīklu motīviem līdzīgi jēdzieni vēl joprojām tiek lietoti tīklu bioloģijas nozarē ar noteiktiem pielāgojumiem, no kuriem spilgtākais piemērs ir grafletu analīzes (Pržulj 2007; Sarajlić et al. 2016).

Šajā promocijas darbā konkrētāk apskatīti ir *bi-fan* motīvi (Ward and Thornton 2007), kuru zīmīgākā īpašība ir to tiešā sasaiste ar gēnu dubultošanos sugas evolūcijas gaitā, atšķirībā no tādiem motīviem kā *feed-forward* cilpas, kuru funkcijas signālu modulācijā ir galvenokārt hipotētiskas. Ņemot vērā, ka gēnu dubultošanās ir notikusi daudzos zināmos gadījumos dažādu sugu ietvaros, piemēram, raugā *Saccharomyces cerevisiae.* Rauga genoms ir pilnībā dubultojies vismaz vienu reizi, un gan visu gēnu dublikātu anotēts saraksts, gan raugam raksturīgie gēnu regulatorie tīkli ir brīvi pieejami publisku resursu formā (Byrne and Wolfe 2005; Teixeira et al. 2023). Šīs darba daļas mērķis bija attiecīgi izmantot šo informāciju, lai pārbaudītu, vai bi-fan tīklu motīviem ir tieša saistība ar gēnu pāra filoģenētisko vēsturi.

Salīdzinot ar pārsvarā netieši novērojamo informāciju, kas atspoguļota jebkurā gēnu regulatorajā tīklā, hromatīna mijiedarbību tīkli atveido hromatīna molekulu telpisko organizāciju. Hromatīns sastāv no proteīnu, nekodējošo ribonukleīnskābju (RNS) un dezoksiribonukleīnskābes (DNS) kompleksa. Gēnu regulācija nosaka hromatīna struktūru (un līdzīgā kārtā hromatīna struktūra nosaka gēnu regulāciju), un hromatīna telpiskā izvietojuma principi eikariotisku šūnu kodolos ir nozīmīgi gēnu funkcionalitātes noskaidrošanai, bet galvenokārt ir ļoti nepilnīgi izzināti. Visplašākais informācijas klāsts par hromatīna izvietojumu jeb konformāciju ir iegūts no hromatīna konformācijas notveres (*chromatin conformation capture*) eksperimentiem. Šajos eksperimentos, īpaši

augstas caurlaides jeb Hi-C eksperimentos (Dekker et al. 2002; Lieberman-Aiden et al. 2009), parasti iegūst hromatīna kontaktu kartējumu (*contact maps*), kur ar genomisku koordinātu palīdzību ir norādīti hromatīni reģioni, kas atrodas savstarpēji tuvā kontaktā. Šo hromatīna kontaktu savietojums, kas izsakāms kā kontaktu matrica, dažādos mērogos var norādīt uz funkcionālu elementu – topoloģiski asociējošo domēnu, promoteru-enhānseru mijiedarbību, hromatīna nodalījumu – klātbūtni un struktūru. Daudzus no šiem jēdzieniem, kas vēlāk ir pārbaudīti ar citām metodēm, pētnieki oriģināli konstatēji no kontaktu matricām ar statistisku analīžu palīdzību (Lajoie, Dekker, and Kaplan 2015; Pancaldi 2023). Šādas skaitļošanas metodes ir ne tikai absolūti nepieciešamas Hi-C datu apstrādei, bet arī to interpretācijai, jo metožu attīstība ļauj noteikt arvien jaunas bioloģiskas funkcijas un to iepriekš nezināmas nianses, kas papildina pieejamo informāciju par gēnu regulāciju šūnu kodolos.

Metožu apskats

Promocijas darbā iekļautie zinātniskie pētījumi tika veikti, izmantojot biomolekulārus datus no jau pabeigtiem zinātniskiem pētījumiem, kas atrodami publiskās datubāzēs. Šo datu plašā pieejamība un daudzveidība nodrošināja lielu potenciālo informācijas klāstu gan gēnu regulatoro tīklu izveidošanai, gan individuālu eksperimentu daudzumu, no kuriem bija iespējams veidot hromatīna mijiedarbību tīklus. Publikācijās veidotie gēnu regulatorie tīkli tika iegūti no *YEASTRACT* un *RegulonDB* (Teixeira et al. 2023; Tierrafría et al. 2022) resursiem, kamēr hromatīna dati tika iegūti no to respektīvajām publikācijām un papildināti ar *3DIV* datubāzes informāciju (Javierre et al. 2016; Jung et al. 2019; Kim et al. 2021), papildinot šos oriģināldatus ar publiski pieejamām genoma anotācijām no *Ensembl* un *ENCODE* (Harrison et al. 2024; Abascal et al. 2020) validācijas nolūkiem, kā arī piesaistot gēnu ekspresijas datus no *Genotype Tissue Expression (GTEx)* un *FANTOM5* gēnu ekspresijas atlasiem (Lonsdale et al. 2013; Noguchi et al. 2017), lai kvantificētu mūsu topoloģisko iezīmju ietekmi uz gēnu ekspresiju.

Iegūtos oriģināldatus bija iespējams tālāk apstrādāt, lai iegūtu topoloģiskās analīzēs lietojamus grafus. Gan hromatīna mijiedarbību, gan gēnu regulācijas gadījumā grafu veidošanas process ir pamatā identisks: genoma lokusi vai gēni veido grafa virsotnes, un mijiedarbības start tiem, ja tādas eksistē, veido grafa šķautnes. Šie grafi ir tieši veidoti no datiem, atlasot mijiedarbības pēc katrai datu kopai specifiskiem kritērijiem, piemēram, pēc mijiedarbību p-vērtības vai ChiCAGO skaitļa (Cairns et al. 2016) hromatīna mijiedarbību datos – katrā publikācijā ir norādīti specifiski lietotie kritēriji. Iegūtajos grafos tad tiek meklēti interesējošie topoloģiskie elementi (konkrēta veida apakšgrafi) atbilstoši pētījuma mērķiem. Šie elementi var būt tīkla motīvi jeb grafleti (publikācijā I), saistības komponentes (publikācijās II–IV) un tādi sīkāki elementi kā cikli un

kliķes (publikācijās III–V). Galvenokārt iekļautajās publikācijās šie elementi tika skaitīti un to normalizētās vērtības savstarpēji salīdzinātas dažādu metriku formā, lai gūtu vispārīgu priekšstatu par topoloģiskām atšķirībām hromosomas vai organisma līmenī. Lai paveiktu šo uzdevumu, darba gaitā bija jārisina vairākas algoritmiskas problēmas un jāveido specializētas metodes topoloģijas izvērtēšanai, bet algoritmu optimizācija nav šī promocijas darba centrālā tematika, tāpēc lielāks uzsvars publikācijās veltīts uz jauno metodoloģiju noderīgumu bioinformātikas kontekstā. Papildus tam, ārpus promocijas darba publikāciju sērijas publicēti tika blakus rezultāti par hromatīna mijiedarbību tīklu randomizāciju un vizualizāciju, kas var būt noderīgi interesentiem par metodoloģijas detaļām un sarežģījumiem darba gaitā (Sizovs, Silina, et al. 2024; Sizovs, Melkus, et al. 2024).

Metožu izstrādes gaitā tika izmēģinātas un pārbaudītas vairākas grafu topoloģijas metodes bioloģiski nozīmīgu pazīmju atrašanai, kas tika validētas ar neatkarīgi iegūtu bet bioloģiski saistītu datu (piemēram, gēnu ekspresijas datu pētītajos audu tipos) palīdzību. Šīs bioloģiskās sakarības tad tika pārbaudītas ar statistisku analīžu palīdzību (piemēram, neparametrisku variācijas analīzi), lai apstiprinātu rezultātu ticamību. Konkrētu bioloģisku pazīmju saistība ar topoloģiskiem elementiem tika formulēta kā "bagātinājums", parasti saistībā ar palielinātu ekspresiju vai citādi mainītu bioloģisko aktivitāti. Šie pamatprincipi publikāciju izstrādes gaitā tika pielāgoti dažādos veidos attiecīgi pētāmajām problēmām un datiem. Detalizētāki metodoloģijas apraksti būs sniegti tālākajās sadaļās, un arī ir atrodami attiecīgajās publikācijās un tajās norādītajās *GitHub* saitēs.

Rezultātu apskats

Lai gan pielietotās grafu teorijā balstītās modelēšanas metodes šī darba publikāciju klāstā ir cieši saistītas, gan to pielāgojumi, gan to rezultāti darba gaitā nozīmē, ka visefektīvāk darba sadaļas izklāstīt ir pēc individuālām publikācijām. Visas piecas darba publikācijas ir iekļautas promocijas darba pamattekstā.

Publikācijā I ir nosegts darbs ar gēnu regulatorajiem tīkliem. Tajā ir pielietots grafos balstīts gēnu regulācijas modelis kombinācijā ar grafletu metodēm, lai analizētu simetriskus tīkla motīvus paralogos gēnu pāros. Darbā ir definēts *bi-fan* vienības jēdziens, un vadoties pēc šī jēdziena, papildus definēts kompleksu motīvu jēdziens, kurā ir dubultotas un simetriskas gēnu pāru pozīcijas. Šo komplekso motīvu klātbūtne ir pārbaudīta raugā *Saccharomyces cerevisiae,* nematodē *Caenorhabditis elegans* un zarnu nūjiņā *Escherichia coli*. Lai gan *E. coli* regulatorajā tīklā nav atrodama zīmīga motīvu simetrija, gan nematodes, gan rauga regulatorajos tīklos paralogiem gēniem ir novērojama ievērojama iekšsugas paralogu simetrija un regulatoro tīklu pārklājums, kas nav raksturīgs ārpussugas paralogiem vai nejauši izvēlētiem gēniem.

Publikācijā II apskatīti hromatīna mijiedarbību tīkli un tajos atrodamo saistības komponenšu topoloģija dažādu asins šūnu promoteru notveres Hi-C (*promoter capture Hi-C* jeb *pcHi-C*) datu ietvaros. Pētījumā ieviests un aprakstīts algoritms bioloģiski nozīmīgu saistības komponenšu atrašanai, kā galvenos kritērijus izvirzot grafa šķautņu pārklājumu dažādos audu tipos un komponenšu izmēru. Šīs komponentes tālāk tiek pārbaudītas ar vienkāršām transkripcijas faktoru bagātinājuma analīzēm, lai noteiktu kopējas regulācijas un ko-ekspresijas pazīmes.

Publikācijās III un IV, kuru tematika ievērojami pārklājas, ir turpināti pētījumi ar hromatīna mijiedarbību datiem, papildinot iepriekš aprakstīto saistības komponenšu pieeju ar topoloģisku metriku komplektu. Šīs metrikas publikācijās ir pielietotas, lai konstatētu topoloģiskas atšķirības starp dažādiem šūnu tipiem saskaņā ar to bioloģisko radniecību (kas izteikta dažādu attāluma metrikas). Pielietojot šīs metrikas asins šūnu pcHi-C datos, tika secināts, ka šīs metrikas nosacītā mērā atbilst atšķirībām starp audu tipiem. Papildus tam komponenšu bioloģiskā nozīmība ir pamatota ar datiem no *FANTOM5* promoteru līmeņa ekspresijas datiem, kas liecina par koordinētas ekspresijas klātbūtni komponentēs.

Publikācijā V tiek izvirzīts konkrēts interesējošs topoloģisks elements hromatīna mijiedarbību tīklu analīzē: kliķe izmērā 3 (triju virsotņu grafs, kurā visas virsotnes ir savstarpēji saistītas). Šādu kliķu sakopojumi dažādās Hi-C datu kopās sakrīt ar ievērojamu transkripcijas sākumpunktu bagātinājumu, kā arī būtiski palielinātu gēnu ekspresiju, īpaši tādā gadījumā, ja atlasa kliķu virsotnes ar RNS polimerāzes II saistības vietām. Hromatīna apgabali, kuros sastopams daudz šādu kliķu, teorētiski atbilst "transkripcijas rūpnīcu" jēdzienam no literatūras, kas apzīmē koordinētas transkripcijas reģionus cieši telpiski saistītos hromatīna apgabalos, tāpēc ir pamats uzskatīt, ka ar kliķu palīdzību šādus reģionus varētu būt viegli atrast datos arī bez papildus bioloģiskas informācijas. Pamatojoties uz šiem panākumiem, tālāk apspriesti tiek arī papildinājumi un pielāgojumi metodēm, lai izvērstu turpmākus pētījumus.

Rezultātu aprobācija

Promocijas darba galvenie rezultāti ir aprakstīti šajās publikācijās, kas pilnībā iekļautas promocijas darba pamattekstā:

1. Melkus, Gatis, Peteris Rucevskis, Edgars Celms, Kārlis Čerāns, Karlis Freivalds, Paulis Kikusts, Lelde Lace, Mārtiņš Opmanis, Darta Rituma, and Juris Viksna. "Network Motif-Based Analysis of Regulatory Patterns in Paralogous Gene Pairs." Journal of Bioinformatics and Computational Biology 18, no. 03 (June 18, 2020): 2040008. https://doi.org/10.1142/S0219720020400089. (autora sniegums: 75%)

2. Viksna, J., G. Melkus, E. Celms, K. Čerāns, K. Freivalds, P. Kikusts, L. Lace, M. Opmanis, D. Rituma, and P. Rucevskis. "Topological Structure Analysis of Chromatin Interaction Networks." BMC Bioinformatics 20 (2019). https://doi.org/10.1186/s12859-019-3237-z. (autora sniegums: 55%)
3. Lace, Lelde, Gatis Melkus, Peteris Rucevskis, Edgars Celms, Karlis Cerans, Paulis Kikusts, Martiņš Opmanis, Darta Rituma, and Juris Viksna. "Graph-Based Characterisations of Cell Types and Functionally Related Modules in Promoter Capture Hi-C Data." Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019), 78–89, 2019. https://doi.org/10.5220/0007390800780089 (autora sniegums: 60%)
4. Lace, Lelde, Gatis Melkus, Peteris Rucevskis, Edgars Celms, Kārlis Čerāns, Paulis Ķikusts, Mārtiņš Opmanis, Darta Rituma, and Juris Viksna. "Characteristic Topological Features of Promoter Capture Hi-C Interaction Networks." Communications in Computer and Information Science, vol. 1211, pp. 192-215, 2020. https://doi.org/10.1007/978-3-030-46970-2_10. (autora sniegums: 65%)
5. Melkus, Gatis, Andrejs Sizovs, Peteris Rucevskis, and Sandra Silina. "Transcriptional Hubs Within Cliques in Ensemble Hi-C Chromatin Interaction Networks." Journal of Computational Biology 31, no. 6 (June 1, 2024): 589–96. https://doi.org/10.1089/cmb.2024.0515. (autora sniegums: 75%)

Papildus šīm publikācijām ir arī papildus zinātniski raksti, kuru tematika ir saistīta ar promocijas darbu, bet nav iekļauta promocijas darba publikāciju kopā (bet var sniegt papildinformāciju par pētījumu gaitu):

- Gatis Melkus, Pēteris Ručevskis, Edgars Celms, Kārlis Čerans, Karlis Freivalds, Paulis Kikusts, Lelde Lāce, Mārtiņš Opmanis, Dārta Rituma, Juris Vīksna. Graph-based network analysis of transcriptional regulation pattern divergence in duplicated yeast gene pairs. ACM International Conference Proceeding Series, 3365954, 2019. https://doi.org/10.1145/3365953.3365954
- Gatis Melkus, Sandra Silina, Andrejs Sizovs, Peteris Rucevskis, Lelde Lace, Edgars Celms, and Juris Viksna. "Clique-Based Topological Characterization of Chromatin Interaction Hubs," 476–86, 2023. https://doi.org/10.1007/978-981-99-7074-2_38.
- Andrejs Sizovs, Gatis Melkus, Peteris Rucevskis, Sandra Silina, Lelde Lace, Edgars Celms, and Juris Viksna. "A Technique for Preserving Network Structure in Randomized Hi-C Data." Journal of Bioinformatics and Computational Biology 22, no. 05 (October 24, 2024). https://doi.org/10.1142/S0219720024400018.
- Andrejs Sizovs, Sandra Silina, Gatis Melkus, Peteris Rucevskis, Lelde Lace, Edgars Celms, and Juris Viksna. "Exploration and Visualization Methods

for Chromatin Interaction Data." edited by Wei Peng, Zhipeng Cai, and Pavel Skums, 101–13. Singapore: Springer Nature Singapore, 2024.

- Melkus, Gatis, Karlis Cerans, Karlis Freivalds, Lelde Lace, Darta Zajakina, and Juris Viksna. "Analysis of Dynamics and Stability of Hybrid System Models of Gene Regulatory Networks." In The 12th International Conference on Computational Systems-Biology and Bioinformatics, 1–10. New York, NY, USA: ACM, 2021. https://doi.org/10.1145/3486713.3486727.
- Lace, Lelde, Karlis Cerans, Karlis Freivalds, Gatis Melkus, and Juris Viksna. "Hybrid Gene Regulation Models of Mammalian Circadian Cycles." In Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies, 130–37. SCITEPRESS – Science and Technology Publications, 2022. https://doi.org/10.5220/0010834400003123.
- Melkus, Gatis, Karlis Cerans, Karlis Freivalds, Lelde Lace, Darta Zajakina, and Juris Viksna. "Behavioral Dynamics of Bacteriophage Gene Regulatory Networks." Journal of Bioinformatics and Computational Biology 20, no. 05 (October 14, 2022). https://doi.org/10.1142/S0219720022500214.
- Viksna, Juris, Karlis Cerans, Lelde Lace, and Gatis Melkus. "Characterizing Behavioural Differentiation in Gene Regulatory Networks with Representation Graphs." NAR Genomics and Bioinformatics 6, no. 3 (July 2, 2024). https://doi.org/10.1093/nargab/lqae102.

Saistībā ar iepriekšminētajām publikācijām, divi studenti no autora pētniecības grupas izstrādāja kvalifikācijas, kursa vai bakalaura darbus promocijas darba autora vadībā:

- Sandra Siliņa, 2023. "Vizualizācijas hromatīna interakciju datu analīzei". Kvalifikācijas darbs.
- Andrejs Sizovs, 2023. "Sistēmas izstrāde Hi-C datu analīzei". Kvalifikācijas darbs.
- Andrejs Sizovs, 2024. "Metodes Hi-C datu randomizācijai, saglabājot to topoloģisko struktūru". Kursa darbs.
- Andrejs Sizovs, 2024. "Metode tīkla struktūras saglabāšanai randomizētajos hromatīna interakciju datos". Bakalaura darbs.

Papildus norādītajām publikācijām saistībā ar promocijas darbu, darba rezultātus autors prezentēja vairākās starptautiskās konferencēs plakātu un mutisku prezentāciju formā:

- "Graph-based network analysis of transcriptional regulation pattern divergence in duplicated yeast gene pairs" – CSBio 2019, Nica, Francija (mutiska prezentācija)
- "Topological features of chromatin interaction networks" – RECOMB 2020, attālināti (plakāts)
- "Structural comparison of chromatin interaction networks generated from Hi-C data" – ECCB 2022, Sidžesa, Spānija (plakāts)

- "Clique-based identification of functional modules in Hi-C graphs" – RECOMB 2023, Stambula, Turcija (plakāts)
- "The utility of cliques in topological characterization of Hi-C data" – ISMB/ECCB 2023, Liona, Francija (plakāts)
- "Clique-based topological characterization of chromatin interaction hubs" – ISBRA 2023, Vroclava, Polija (mutiska prezentācija)
- "Gene Expression Variability Linked to Chromatin Clique Configurations and cis-Regulatory Elements" – ICCBB 2024, Kioto, Japāna (mutiska prezentācija)

Galvenās publikācijas ir iekļautas promocijas darbā kā atsevišķas nodaļas, un gan kopsavilkums, gan promocijas darbs ir attiecīgi organizēts, lai sniegtu šo publikāciju apskatu. Darbs ir 144 lappušu garumā un satur atsauces uz 183 avotiem (neskaitot atsauces, kas iekļautas rakstu krājuma publikācijās, kas norādītas individuāli skaidrības labad).

1. PUBLIKĀCIJA I – *NETWORK MOTIF-BASED ANALYSIS OF PARALOGOUS GENE PAIRS*

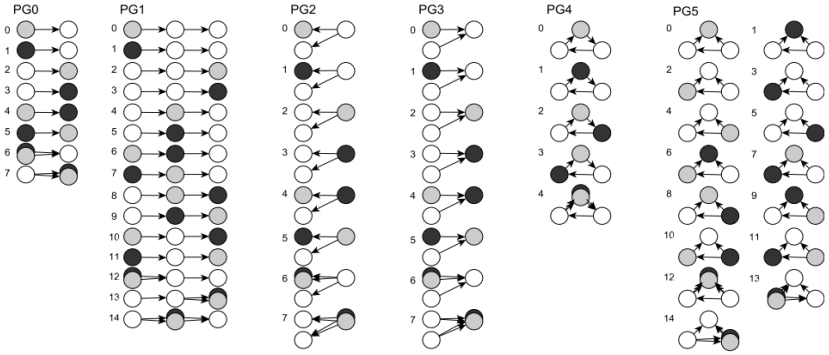
Šajā nodaļā tiek izskatīta pirmā kopas publikācija, kurā pētīta *bi-fan* tīkla motīvu sastopamība un topoloģija gēnu regulatorajos tīklos. Tīkla motīvu sākotne ir meklējama tīklu bioloģijas pirmsākumos, kur pētnieki novēroja, ka daži topoloģiski veidojumi gēnu regulatorajos tīklos ir daudz biežāk sastopami nekā līdzīgos matemātiski modelētos tīklos ar identiskām virsotņu pakāpēm un šķautņu sadalījumiem (Milo et al. 2002). Dažiem no šiem motīviem, piemēram, *feed-forward* un atgriezeniskajām (*feedback*) cilpām, ir veltīts plašs literatūras klāsts, bet šajā publikācijā uzmanība ir vērsta tieši uz *bi-fan* motīvu un tā topoloģisko saistību ar gēnu dublikāciju.

Maizes rauga *Saccharomyces cerevisiae* pilna genoma dublikācija sākotnēji tika pierādīta ar gēnu sakārtojuma pētījumu un filoģenētiskas izpētes palīdzību (Kellis, Birren, and Lander 2004). Papildus genoma sekvencē konstatējamām dublikācijas pēdām, tīklu bioloģijas literatūrā iepriekš ir parādīts, ka simetriski *bi-fan* motīvi gēnu regulatorajos tīklos var liecināt par gēnu radniecību (Ward and Thornton 2007). Šīs simetrijas teorētiskais pamats ir gēnu dublikācijas mehānisms – pieņemot, ka dublētais gēns ir transkripcijas faktors are vairākiem regulatoriem mērķiem, tad gēnu dublikācijas gadījumā abi jaunie gēni visticamāk gan regulēs vienus un tos pašus gēnus, gan arī tiks regulēti galvenokārt identiski citu transkripcijas faktoru ietekmē. Pēc sākotnējās dublikācijas gēnu sekvences savstarpēji diverģē mutāciju un citu evolucionāru notikumu ietekmē, un attiecīgi saistības vietas ar regulatoriem (kas ir atkarīgas no sekvences) laika gaitā izzūd un mainās arī gēnu produktu struktūra. Tas nozīmē, ka ir teorētiski ticami, ka paralogiem (gēniem, kuri radušies dublikācijas procesā un kādreiz bijuši tikai viens gēns) vienas sugas ietvaros vajadzētu būt kopīgām regulatorām mijiedarbībām ne tikai palielināta *bi-fan* motīvu skaita ziņā, bet arī citu tīklu motīvu vai grafletu sastāvā, kas arī ir sastopami gēnu regulatorajos tīklos.

Rauga gēnu regulatorā tīklu izveidei tika izmantoti dati no YEASTRACT regulatoru datubāzes (Teixeira et al. 2023). Pamatojoties uz šo informāciju, tika izveidots grafs, kur katra virsotne bija gēns (un tā attiecīgais produkts, ja tas regulē tālākus gēnus), bet katra šķautne bija regulatora mijiedarbība, neizšķirot aktivējošas un represējošas mijiedarbības, jo tām var būt neviennozīmīga ietekme uz gēnu ekspresiju (Larsen et al. 2019). Lai pārbaudītu metodes efektivitāti citos organismos, pārbaudīts arī tika zarnu nūjiņas *Escherichia coli* regulatorais tīkls no RegulonDB (Tierrafría et al. 2022) un publicēts proteīnu-DNS mijiedarbību tīkls no nematodes *Caenorhabditis elegans* (Reece-Hoyes et al. 2013). Visiem trijiem tīkliem ir pieejami labi zināmi saraksti ar iekšsugas paralogiem

(gēniem, kuru dublikācija notikusi nesenā sugas evolūcijas vēsturē), ārpussugas paralogiem (jeb gēniem, kas ir vienā gēnu dzimtā bet ir dublējušies pirms sugas nošķiršanās no senčiem), kā arī onologiem (*ohnologs*, gēniem, kuru dublikācija notikusi visa genoma dublikācijas ietvaros). Šīs kategorijas pētijumā izmantotas, lai nošķirtu dažādas izcelsmes paralogus vienu no otra, lai iegūtu precīzāku priekšstatu par šo tīklu evolūciju.

Lai izvērtētu motīvu simetriju mūsu gēnu regulatorajos tīklos, pētījumā tika ieviests *bi-fan* vienības jēdziens. *Bi-fan* vienība ir mazs apakšgrafs (graflets), ko veido trīs virsotnes, starp kurām ir divas orientētas šķautnes, no kurām abas ieiet vienā un tajā pašā virsotnē vai iziet no vienas un tās pašas virsotnes. Salīdzinot ar *bi-fan* motīvu, šī vienība ir vieglāk skaitāma un vienkāršo *bi-fan* lielāka izmēra *bi-fan* struktūru aprēķināšanu, kur gēnu pārim var būt kopīgs plašs regulatoru vai regulācijas mērķu skaits. Lai apskatītu motīvu simetriju plašākā nozīmē, pētījumā bi-fan vienības ir iesaistītas lielākos motīvos, radot dubultotas regulatoras pozīcijas vienkāršos motīvos (sk. 1. attēlu), kas veido kompleksus motīvus ar simetriskām pozīcijām un dažādiem motīvu pārklājuma variantiem. Ar kompleksu motīvu palīdzību ir iespējams izmērīt *bi-fan* simetriju konkrētam gēnu pārim, aprēķinot motīvu pārklājumu starp abiem pāra gēniem salīdzinājumā ar to individuālo motīvu skaitu.



1. attēls. Visas iespējamās grafletu (jeb tīkla motīvu) variācijas, kas publikācijā ir skaitītas un izmantotas simetrijas mērījumiem. Melnā un pelēkā krāsā iekrāsotās pozīcijas norāda gēnus, kas veido pētāmo paralogu pāri.

Tīklu motīvu sākotnējā formulējumā to bioloģiskais nozīmīgums ir pamatots ar to biežāku sastopamību bioloģiskos tīklos salīdzinājumā ar to, kas būtu sagaidāms modelētā grafā ar tādu pašu virsotņu sadalījumu (Alon 2007). Tas nozīmē ka, piemēram, rauga regulatorajā tīklā vajadzētu būt būtiski vairāk *feed-forward* cilpu nekā būtu tipiski novērojamas nejauši izveidotā tīklā ar

identiskām virsotņu pakāpēm un kopumā līdzīgām īpašībām (klasiski šos nejaušos tīklus modelē kā Erdēša-Renji tīklus, bet ir arī citi iespējami varianti). Šajā gadījumā grafletu palielināts skaits tika pārbaudīts, nosakot, vai tad, ja datiem pievieno noteiktu trokšņa proporciju (apmainot tīkla virsotņu šķautnes, saglabājot virsotņu pakāpi tīklā kopumā), samazinās vienkāršu un kompleksu motīvu skaits datos.

Rezultāti

Ar kompleksu motīvu metriku palīdzību bija iespējams izšķirt zīmīgas atšķirības tīklu topoloģijas izmaiņās trokšņa pielikšanas gadījumā. Pirmkārt, kompleksu motīvu skaits rauga datos bija par pakāpi zemāks nekā vienkāršo motīvu skaits, un daži kompleksie motīvi bija ievērojami plašāk izplatīti nekā citi. Galvenā novērojamā atšķirība bija dubultoto pozīciju biežāka sastopamība gala jeb Z pozīcijās mūsu izvēlētajos grafletos, kas ir izskaidrojams ar to, ka šajās pozīcijās nav nepieciešams specifiski transkripcijas faktora gēns, bet gan jebkurš gēns, kas ir ievērojami plašāka potenciālo gēnu kopa. Otrkārt, kompleksiem motīviem bija novērojama daudz intensīvāka skaita samazināšanās trokšņa ietekmē nekā vienkāršiem motīviem. Ņemot vērā to, ka motīvu sastopamības biežums klasiski ir viens no galvenajiem hipotētiskas bioloģiskās nozīmības kritērijiem, tas nozīmē, ka šie kompleksie motīvi patiešām varētu būt veidojušies bioloģiska procesa ietekmē un nav izskaidrojami tikai ar tīkla īpašībām. Treškārt, neviena no mūsu topoloģiskajām metrikām nekorelēja ar gēnu, promoteru vai proteīnu sekvenču līdzību, kas nav pārsteidzoši, jo gēnu radniecību var būt grūti mērīt tikai no sekvences salīdzinājuma.

Individuālu motīvu metrikas galvenokārt ir spēcīgi sasaistītas rauga datos, īpaši viena gēnu pāra ietvaros. No tām vērts izcelt spēcīgo korelāciju starp 2-metrikām (gēnu pāra otrā un parasti mazākas virsotņu pakāpes gēna motīvi) un 12-metrikām (abu gēnu pāru kopīgajiem motīviem), kur 2-metrikas noteica augšējo robežu attiecīgo 12-metriku skaitļiem. No šī novērojuma tika izveidota jauna metrika, *simetrija*, kas ir attiecība starp 2-metriku un tās atbilstošo 12-metriku vienā gēnu pārī. Šī metrika darbojās kā efektīvs līdzeklis gēnu pāra topoloģiskās līdzības mērs, kas reizē arī vienkāršā veidā normalizēja motīvu skaitus, atvieglinot to salīdzinājumu. Simetrijas salīdzinājumi rauga, nematodes un *E. coli* regulatorajos tīklos sniedza ieskatu šo organismu regulatoro tīklu topoloģijas līdzībai paralogu starpā. Lai gan *E. coli* gadījumā iekšsugas paralogiem nebija novērojama ievērojami palielināta simetrija salīdzinot ar nejauši izvēlētiem gēniem, gan nematodes, gan rauga gadījumā iekšsugas paralogiem (un rauga gadījumā arī onologiem) bija novērojama stipri palielināta motīvu simetrija salīdzinājumā ar ārpussugas paralogiem, kā arī ar nejaušiem gēniem. Rezultātu atšķirība *E. coli* var būt dažādi skaidrojama, bet visticamāk ir saistīta ar šī regulatorā tīkla salīdzinošo skrajumu, kā arī anotāciju atšķirībām un pašu

baktēriju atšķirīgo genomiskās evolūcijas tempu un niansēm salīdzinājumā ar eikariotiem. Lai tālāk vispārinātu šī pētījuma rezultātus, visticamāk būtu produktīvi izsmeļošāk pētīt vajadzīgos tīkla izmērus un grafu topoloģiskās atšķirības vispārīgākā ziņā.

Jebkurā gadījumā šajā publikācijā tika veiksmīgi pārbaudīta ne tikai *bi-fan* motīva saistība ar gēnu radniecību, bet arī ieviestas metodes potenciāliem tālākiem grafletu pētījumiem. Šajā gadījumā promocijas darba ietvaros pētījums netika turpināts, tā vietā pievēršoties hromatīna mijiedarbību tīkliem, kur topoloģijas atšķirības starp dažādiem datu tipiem arī turpinās būt aktuāls jautājums līdz ar citām hromatīna pētījumiem specifiskām problēmām.

2. PUBLIKĀCIJA II – *TOPOLOGICAL STRUCTURE ANALYSIS OF CHROMATIN INTERACTION NETWORKS*

Hromatīna konformācijas notveres tehnoloģijas ir viegli savietojamas ar grafos balstītām analīzes metodēm, un šādas metodes sniedz plašas iespējas jaunu secinājumu iegūšanā no jau eksistējošiem datiem. Šī iemesla dēļ atlikušās promocijas darba publikācijas ir saistītas tieši ar Hi-C (augstas caurlaides hromatīna konformācijas notveres, *high-throughput chromatin conformation capture*) datu apstrādi, analīzi un integrāciju ar grafos balstītu metožu palīdzību. Salīdzinot ar gēnu regulatorajiem tīkliem, hromatīna mijiedarbību tīkli ir lielāki, blīvāki un tiešākā veidā saistīti ar biomolekulu fizikālo stāvokli šūnas kodolā, tāpēc tīklu bioloģijas piedāvātās abstrakcijas metodes šajā gadījumā ir noderīgas, lai apkopotu un sistematizētu grafu informāciju, ko iespējams veidot 1-pret-1 no Hi-C kontaktu matricām, pārvēršot tās par grafu tuvības matricām.

Darbs pie hromatīna mijiedarbību tīkliem tika uzsākts ar 17 dažādu asins šūnu tipu promoteru notveres Hi-C (pcHi-C) datu kopu (Javierre et al. 2016). Šī datu kopa apstrādei bija labi piemērota, jo tā ir gan labi anotēta ar lielu papilddatu daudzumu (piemēram, hromatīna aktivitātes iezīmēm), gan arī salīdzinoši pieticīga izmēra, kas ļāva to apstrādāt ar plašu metožu klāstu bez lieliem algoritmiskiem sarežģījumiem, tajā pašā laikā nodrošinot pietiekamu mijiedarbību klāstu statistiski ticamām analīzēm. Līdzīgā kārtā mijiedarbības šajā datu kopā ir grupētas pēc to CHiCAGO (Cairns et al. 2016) indeksiem, kas nāk no publikācijas oriģinālās apstrādes metodēm, vienkāršojot datu atlasi un filtrēšanu. Visi audu tipi datu kopā ir iegūti no veseliem cilvēkiem, un ir savstarpēji bioloģiski saistīti, jo tiem ir kopīgs sencis asinsrades kokā (visu cilvēka ķermenī atrodamo asins šūnu izcelsmes koks, sākot ar hematopoiētiskajām cilmes šūnām, kas tālāk diferencējas visos asins šūnu tipos), tāpēc šai datu kopai bija arī salīdzinoši vienkārši izzināt audu tipu precīzo radniecību, ar kuru varētu salīdzināt mūsu topoloģiskās analīzes rezultātus.

Šīs publikācijas galvenais uzdevums bija atrast asinsrades koka apakškategorijām specifisku funkcionālu moduļu atrašana. Šiem moduļiem, kurus var aprakstīt kā saistības komponentes no mijiedarbību grafa, vajadzēja būt nošķiramiem gan no komponentēm, kas ir atrodamas tikai individuālos audu tipos, gan arī komponentēm, kas atrodamas visos audu tipos, lai atrastu tiešu tādas komponentes, kuras būtu iespējams sasaistīt ar šūnu diferenciāciju. Lielākajā daļā audu tipu kontaktu kartējums pēc filtrēšanas sastāvēja no vairāk nekā 150 000 mijiedarbību, tāpēc komponenšu atrašanai bija nepieciešams izstrādāt algoritmu.

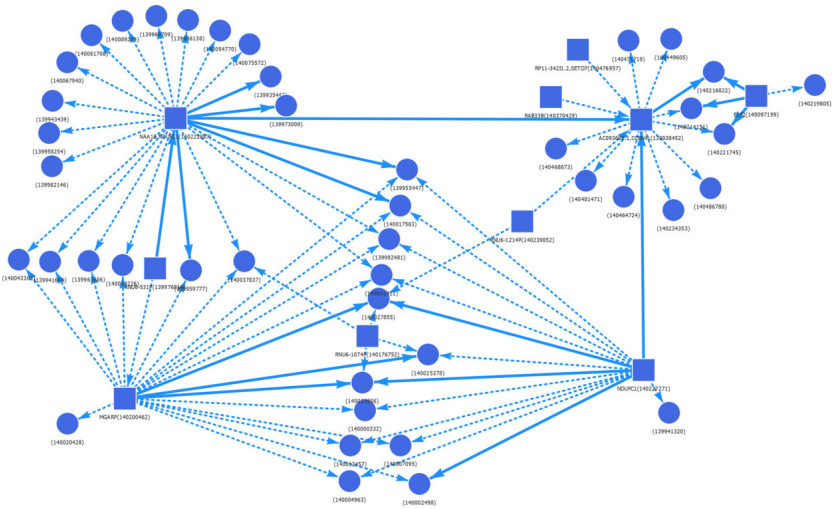
Darbā izmantotās pieejas pamatā ir pārklājums starp dažādu asins šūnu tipu hromatīna kontaktiem. Izmantotajā datu kopā ir aptuveni 700 000 unikālu mijiedarbību 17 audu tipos, no kurām ir iespējams konstruēt grafu katrai hromosomai (ir arī mazs mijiedarbību skaits starp hromosomām, bet tās šeit nav ņemtas vērā, sekojot rekomendācijām (Lajoie, Dekker, and Kaplan 2015) no tajā brīdī pieejamās analītiskās literatūras). Šajos grafos tika atrastas visas mijiedarbības visos audu tipos, kas atbilda datu filtrēšanas kritērijiem, un tās attiecīgi tika marķētas ar iezīmēm, kuros audu tipos tās ir sastopamas. Izmantojot šo informāciju, tika pielietots speciāli veidots saistīto komponenšu atrašanas algoritms (FindNetworkComponents), veicot plašummeklēšanu binomiālā kokā, kas sastāvēja no visām šūnu tipu kombinācijām, kas šajos datos bija gana vienkārši aprēķināms. Lai no atrastajām komponentēm iegūtu pēc iespējas labākus rezultātus, bija jāveic papildus atlase, izvēloties viegli analizējama izmēra (10...100 virsotņu) komponentes, kurās vairāk nekā 75% šķautņu saglabātos nelielā audu tipu komplektā, bet nesaglabātos visā datu kopā. Šo atlasi darbā izdevās veikt ar SignificanceScore funkciju, kas piešķīra skaitlisku vērtību katras komponentes atbilstībai prasītajiem kritērijiem (sk. 2. attēlu, lai redzētu šādas komponentes piemēru). Sašaurinot meklējumu loku šādā veidā, tika atlasītas komponentes, kas būtu noderīgas tālākai analīzei. Esošajos datos izvēlētais algoritms sniedza pietiekami labus rezultātus, bet tā darbības laiks visticamāk ierobežo tā lietojamību lielām datu kopām.

Nākamais solis komponenšu raksturojumā bija to sasaistīšana ar bioloģisku informāciju no citiem avotiem, kas bija iespējams, lietojot BLUEPRINT projekta RNS sekvenēšanas datus (Adams et al. 2012) un FANTOM5 pozicionālās gēnu ekspresijas datus (Noguchi et al. 2017), ar kuru palīdzību būtu teorētiski iespējams noskaidrot gēnu ekspresijas izmaiņas komponenšu ietvaros, bet abām datu kopām bija pārāk mazs pārklājums ar izvēlētajiem audu tipiem, lai to veiksmīgi pielietotu šajā publikācijā (tālākās publikācijās šī problēma tiks risināta). Tā vietā analīzei tika izmantots Enrichr tiešsaistes rīks (Kuleshov et al. 2016), ar kuru tika pārbaudīta kopīgu transkripcijas faktoru klātbūtne algoritmiski visatbilstošākajās komponentes. Papildus tam analīzē tika iesaistītas ChromHMM anotācijas (Ernst and Kellis 2017), kuras norādīja paaugstinātas, pazeminātas vai noklusinātas aktivitātes reģionus komponentās, kas varētu potenciāli sakrist ar koordinētas gēnu ekspresijas reģioniem.

Rezultāti

Izvēlētajā Hi-C datu kopā grafu analīzes bija lielā mērā produktīvas ar nosacītām problēmām. Pirmkārt, lielu daļu no hromatīna mijiedarbību grafa veidoja viena liela komponente, kas saturēja pusi no visām grafa virsotnēm, bet vidēja izmēra komponentes, kuras varēja algoritmiski atlasīt, veidoja salīdzinoši mazu grafa daļu. Tomēr arī šī grafa daļa bija liela izmēra, saturot vairākus tūkstošus

komponenšu 10...100 virsotņu lielumā visās pētītajās hromosomās, paverot turpmākas analīzes iespējas (Y hromosoma tās mazā izmēra un minimālā grafa dēļ nebija iekļauta analīzē).



2. attēls. Algoritmiski atlasītas komponentes piemērs. Ar nepārtrauktu līniju ir norādītas šķautnes, kas saglabājas līdz ar audu tipu pievienošanu, bet šķautnes ar pārtrauktu līniju nesaglabājas.

Komponenšu bioloģiskās validācijas gaitā tika parādīts, ka izvēlētajās komponentēs bija novērojams lielāks anotāciju īpatsvars kopumā, kas nozīmē, ka gan palielinātas ekspresijas anotācijas, gan represēta hromatīna anotācijas, gan arī miera stāvokļa anotācijas bija vairāk sastopamas izvēlētajās komponentēs salīdzinot ar visu datu kopu. Transkripcijas faktoru analīzes līdzīgi norādīja uz pieticīgām asociācijām ar dažādām transkripcijas faktoru kopām. Kopumā šis rezultāts norādīja uz to, ka lai gan saistības komponentes pašas par sevi nosacīti grupē regulētus genoma reģionus, tās nav sevišķi efektīvas konkrētas bioloģiskas iezīmes izolēšanai. Tātad, lai šo analīzes virzienu tālāk attīstītu, pirmkārt bija nepieciešams atrast šaurākus topoloģiskus elementus, kuri būtu tālāk izmantojams analīžu papildināšanai. Šie elementi varētu būt noteikta veida apakšgrafi, piemēram, zvaigžņveida sakopojumi ap augstas pakāpes virsotnēm noteiktos reģionos, un šādu elementu izolēšanai ir nepieciešamas tālākas topoloģiskas analīzes. Otrkārt, izvēlētās komponentes bija nepieciešams izsmeļošāk salīdzināt ar pārējo datu kopu, lai izprastu, vai nav kāda labāka kritērija, pēc kura atrast precīzākus hromatīna sakopojumus ar skaidrākām gēnu aktivitātes iezīmēm. Šīs nepieciešamības tika tālāk apskatītas nākamajās publikācijās.

3. PUBLIKĀCIJAS III UN IV – *CHARACTERISTIC TOPOLOGICAL FEATURES OF PROMOTER CAPTURE HI-C NETWORKS, GRAPH-BASED CHARACTERIZATIONS OF CELL TYPES AND FUNCTIONALLY RELATED MODULES IN PROMOTER CAPTURE HI-C DATA*

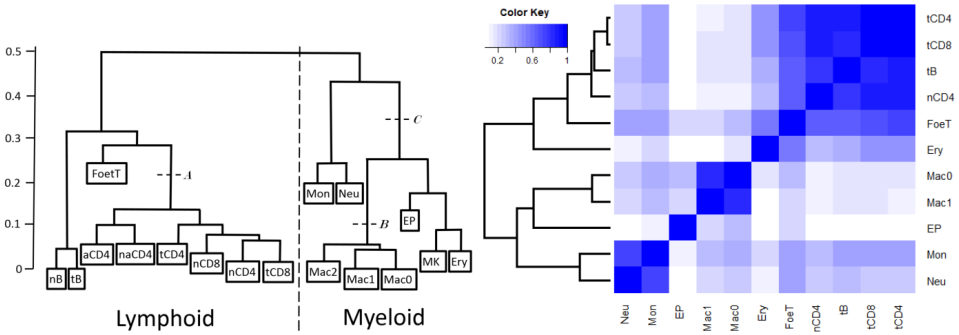
Turpinot darbu pie hromatīna mijiedarbību datu topoloģiskās analīzes, šajās publikācijās ir apskatītas papildus pieejas interesantu topoloģisku īpašību atrašanai pcHi-C datos. Galvenā no šīm pieejām ir topoloģiskas metrikas, ar kurām tika skaitīti noteikta veida apakšgrafi lielajos audu tipu grafos, lai topoloģiski novērtētu atšķirības starp pētītajiem audu tipiem. Publikācijas III un IV lielā mērā pārklājas gan izmantotajā metodoloģijā, gan galvenajās atziņās, tāpēc tās šajā gadījumā ir savienotas vienā kopsavilkuma sadaļā, lai gan tās ir norādītas atsevišķi promocijas darbā.

Pirms ķerties pie pārējām metodēm, ir vērts izcelt, ka šeit izmantotā topoloģiskā pieeja ir domāta orientētiem grafiem. Šķautņu orientācija nav dabiski piemērojama hromatīna mijiedarbību datiem, jo jebkurš hromatīna kontakts ir divpusējs, jo attālums starp diviem genoma reģioniem ir viens un tas pats abos virzienos un tādējādi šeit nav raksturīga tāda asimetrija, kāda novērojama, piemēram, gēnu regulācijā. Šajā gadījumā virzieni šķautnēm ir piemēroti nevis bioloģiskas atšķirības dēļ, bet gan tāpēc, ka lietotā asins audu tipu datu kopa ir promoteru notveres jeb pcHi-C datu kopa, kurā ir konkrēti atlasīti tikai tās hromatīna mijiedarbības, kurās vismaz viens no kontakta reģioniem ir iepriekš izvēlēts gēnu promotera reģions (Mifsud et al. 2015). Šos promoteru reģionus eksperimenta dizainā dēvē par "ēsmām" (*baits*), un to kontaktējošos reģionus (kas var nebūt promoteri) par "otriem galiem" (*other ends*). Viena šāda ēsma var atrasties kontaktā ar citām ēsmām, kas tādējādi veido lielākas komponentes. Visās mijiedarbībās ir vismaz viena ēsma, kas vienmēr ir skaidri izcelta eksperimentālajos datos, tāpēc nosacītā mērā pcHi-C datos mijiedarbībām ir sava veida virziens, kas tādējādi paver būtiskas topoloģiskas variācijas, kas tālāk apskatītas šajā pētījumā.

Izmantojot dažādas topoloģisko elementu variācijas, šajās publikācijās tika pārbaudītas 57 topoloģiskas metrikas, ar kurām tika mērīti dažādu topoloģisko elementu skaiti (galvenokārt dažāda veida saistības komponentes), šķautņu un virsotņu vidējie un maksimālie skaiti šajos elementos, kā arī papildus topoloģiskā informācija (pilns saraksts atrodams publikācijās). Šīs *Base57* metrikām bija paredzētas visu vienkāršo topoloģisko variāciju individuālai mērīšanai,

tāpēc daudzas no tām spēcīgi korelēja savā starpā (piemēram, maksimālais virsotņu skaits spēcīgi sasaistītās saistības komponentēs un maksimālais šķautņu skaits šajās komponentēs), kas nozīmēja, ka šo 57 metriku komplektu tad varēja samazināt līdz 34 metrikām, izslēdzot metrikas ar korelācijas koeficientiem, kas pārsniedza 0,93. Šo *Base34* varēja tālāk samazināt ar pakāpenisku regresijas modeli (izmantojot Akaike informācijas kritēriju jeb AIC), lai iegūtu galīgo 11 metriku komplektu *Base11* publikācijā III, un vēl tālāk sašaurināto *Base6* komplektu publikācijā IV. Šīs metrikas tad varēja salīdzināt ar šūnu radniecības attāluma metrikām.

Izmērīt precīzu attālumu starp audu tipiem nav vienkārši, jo nav absolūtas metodes, lai noskaidrotu, cik atšķirīgi tie varētu būt, kas nozīmē, ka metodoloģija ir potenciāli pielāgojama dažādos veidos. Lai pārbaudītu dažādus attāluma variantus, tika ieviestas sešas attāluma metrikas, no kurām divas (D_{cont} un D_b) bija balstītas hierarhiskas šūnu tipa klāsterēšanas rezultātiem no oriģinālās datu kopas, kā redzams 3. attēlā (Javierre et al. 2016), kamēr pārējās četras (D_A, D_A, D_B, D_C) pamatojās asinsrades kokā, izmantojot tā sazarojumus kā attāluma mērīšanas metodi. No šīm attāluma metrikām D_A, D_B, D_C and D_b ir bināras (pārbaudot to, vai divi šūnu tipi atrodas vienā klāsterētā vai asinsrades koka zarā), kamēr D_4 nedaudz sarežģītākā veidā modelē attālumus starp šūnām pēc asinsrades koka, sadalot audu tipus 5 apakšgrupās, bet D_{cont} izmanto hierarhiskās klāsterēšanas koka attālumus no oriģināldatiem tiešā veidā. Ar šiem sešiem attāluma variantiem tika pārbaudīta metriku spēja izšķirt starp dažāda attāluma audu tipiem.



3. attēls D_{cont} attāluma metrika, vizualizēta kā divdimensionāls kartējums ar iekrāsotu līdzību starp audu tipiem, kas iegūta no oriģināldatiem (Javierre et al. 2016).

Šajās publikācijās arī tika turpināta komponenšu struktūras analīze un validācija, izmantojot gan Enrichr rīku (Kuleshov et al. 2016), gan arī pielietojot

FANTOM5 ekspresijas datus (Noguchi et al. 2017). FANTOM5 dati šajā gadījumā bija obligāti nepieciešami, lai atrastu saikni starp hromatīna topoloģiju un gēnu ekspresiju, ja tāda patiešām pastāv datu kopā. Līdzīgi kā publikācijā II, FANTOM5 pārklājums ar datu kopu bija pieticīgs (18% no šķautnēm 11 no 17 audu tipos), bet šajā gadījumā pārklājums bija pietiekams, lai analizētu gēnu ekspresijas korelācijas komponenšu iekšienē salīdzinot ar gēnu ekspresijas korelācijām starp komponentēm, kā arī korelācijas starp tuvāk un tālāk radniecīgiem audu tipiem komponenšu gēnu ekspresijas ziņā.

Rezultāti

Ar darbā izvirzīto topoloģisko metriku komplektu bija iespējams aprakstīt un izšķirt vairākus audu tipus no izvēlētās datu kopas, lai gan šīm metrikām ne vienmēr bija novērojami vienādi rezultāti dažādās hromosomās. Hromosomām raksturīgas topoloģiskas īpatnības šķietami ietekmēja metriku rezultātus, un lai gan bija novērojams, piemēram, divsakarības komponenšu būtiskums audu tipu atšķiršanā, šeit izvirzītā pieeja nesniedza sevišķi izsmeļošu skaidrojumu, kāpēc tieši šādas topoloģiskas sakarības varētu būt zīmīgas, kas padarīja validāciju vēl nozīmīgāku nākamajos soļos. Papildus sarežģījumi veidojās no izvēlētā attālumu metriku komplekta – lai gan šajā pētījumā bija iespējams iegūt kvantitatīvus un lielā mērā atšķirīgus attālumus starp audu tipiem ar D_{cont} metriku, šī metrika tajā pašā laikā nāca no šīs pašas datu kopas, kas nozīmē, ka lai gan topoloģiskās un attāluma metrikas teorētiski bija attiecināmas uz dažādiem objektiem un mērījumiem, D_{cont} metrikā pamatotie rezultāti nebija pilnībā objektīvi salīdzināmi ar mazāk precīzajām asinsrades koka metrikām, kas kopumā sarežģīja analīzi un laika gaitā noveda pie metriku pieejas atmešanas publikācijā V.

Vēl viens metodoloģisks apsvērums, kas jāņem vērā, ir šīs metodoloģijas izstrāde un pielietojums šajā konkrētajā datu kopā. Izmantotā asins šūnu pcHi-C datu kopa bija labi piemērota eksperimentālajām vajadzībām tās plašā audu tipu klāsta, normālu audu lietojuma un tuvās radniecības dēļ atšķirībā no, piemēram, vēža šūnu kultūru Hi-C datiem, kur audu tipi nav radniecīgi un nepilnīgi atbilst vesela organisma genomam dažādos veidos. Tajā pašā laikā pcHi-C dati ir maza apakškopa no Hi-C eksperimentiem kopumā, un citu līdzīgu pcHi-C datu kopu atrašana, kas šo publikāciju izstrādes gaitā nebija sekmīga, sarežģīja rezultātu validāciju plašākā nozīmē. Papildus tam, metodes pielietošana zināmām Hi-C datu kopām arī nebija vienkārša, jo tās ir ne tikai daudz lielākas, bet arī ar ievērojami atšķirīgām struktūrām, kas nav raksturīgas pcHi-C datu kopām – labs piemērs ir viena no zelta standarta augstas izšķirtspējas Hi-C datu kopām (Rao et al. 2014), ko praktiski nebija iespējams apstrādāt ar pētījumu grupai pieejamajiem resursiem, izmantojot šīs mazākām pcHi-C datu kopām piemērotas metodes.

Par spīti šiem metodoloģiskajiem sarežģījumiem, šīs publikācijas jebkurā gadījumā apstiprināja, ka noderīgākā pieeja tālāku topoloģisku sakarību

atrašanai ir pievēršanās konkrētam, mazākam elementam, kas tālāk tiks izvērsta pētījumos, kur kliķes, nevis saistības komponentes kopumā, ir izmantotas kā galvenais topoloģiskais elements hromatīna struktūru topoloģiskai atlasei. Šajos pētījumos arī būtiska ir ekspresijas datu piesaiste analīzēm, kas ļauj materiāli salīdzināt ekspresiju topoloģisko elementu ietvaros. Papildinot šos centienus tālākā darbā un pielāgojot metodes jauniem datiem, šie panākumi tika tālāk attīstīti kā efektīvākas metodoloģiskas pieejas gēnu regulācijas un topoloģijas saistības pārbaudīšanai, kas veidoja nākamās publikācijas pamatmetodes.

4. PUBLIKĀCIJA V – *TRANSCRIPTIONAL HUBS WITHIN CLIQUES IN ENSEMBLE HI-C CHROMATIN INTERACTION NETWORKS*

Pamatojoties uz iepriekšējo publikāciju ietvaros iegūtajām atziņām par hromatīna mijiedarbību datu un to saistības komponenšu topoloģisko struktūru, kā arī nepublicētu darbu ar audu līdzību un topoloģisko metriku savienošanu, iepriekšminētās analītiskās pieejas tika tālāk pielāgotas jaunām datu kopām un pilnveidotas, sasniedzot lielākos panākumus ar kliķēs balstītām analīzēm, kas aprakstītas publikācijā V, kas ir papildinājums iepriekš konferencē iesniegtam rakstam (Melkus et al. 2023).

Kliķēm kā galvenajam pētījumu objektam ir vairākas metodoloģiskas priekšrocības salīdzinājumā ar topoloģisko metriku pieeju. Pirmkārt, kliķes ir skaitāmas un aprēķināmas visa hromatīna mijiedarbību grafa ietvaros, kas lielā mērā samazina problēmas ar iepriekšējo saistības komponenšu atlases algoritmu un tā spēju atrast vislabāko komponenšu kombināciju. Otrkārt, kliķes ir šaurāk definēts elements nekā saistības komponente, kas nozīmē, ka tās atrast ir relatīvi vienkārši lielākā grafā, samazinot meklējumu loku līdz kliķēm izmērā 3 (C3), kas ir pilns trīs virsotņu grafs, ko ir salīdzinoši vienkārši atrast, bet kurš arī nosedz visas lielāka izmēra kliķes kā šo mazāko kliķu sakopojumus. Treškārt, kliķes ir retāk izvietotas datos un norāda uz specifiskiem blīviem hromatīna sakopojumiem, kas padara tās noderīgākas konkrētu regulācijas moduļu atrašanai, jo tās būtībā norāda uz vietējiem maksimuma punktiem grafa virsotņu savstarpējai saistībai noteiktos hromatīna reģionos.

Šī kliķu skaitīšanas pieeja tika piemērota ne tikai jau lietotajiem asins šūnu pcHi-C datiem, bet arī jaunām datu kopām, kas tika iegūtas no *3DIV* datubāzes, tostarp 27 audu tipu pcHi-C datu kopai (Jung et al. 2019) un šai kopai atbilstošu 20 audu tipu Hi-C datu komplektam (Kim et al. 2021), kas ļāva izdarīt plašāka mēroga secinājumus par Hi-C datu topoloģiju neatkarīgi no konkrēta eksperimenta parametriem. Atšķirībā no iepriekš lietotajiem datiem, šie audu tipu dati bija iegūti no dažādiem audu tipiem, tāpēc tiem bija ievērojami grūtāk izvirzīt loģisku radniecības shēmu – lai gan bija iespējams iegūt nosacītu audu līdzību, kas tika aprēķināta no ekspresijas vērtībām (Manatakis, VanDevender, and Manolakos 2021), šie dati gan to struktūras, gan atšķirīgo eksperimenta nostādņu dēļ kopumā nebija savietojami ar iepriekšējo topoloģisko metriku pieeju.

Arī grafu veidošanas process šajā gadījumā tika būtiski mainīts. Asins šūnu pcHi-C datu gadījumā tika veidots grafs ar orientētām šķautnēm, lai atspoguļotu pcHi-C datu struktūru, kas nebija nepieciešams šeit izmantotajos Hi-C

datos, un datu integrācijas un savstarpējā salīdzinājuma vienkāršošanas nolūkos gan Hi-C, gan pcHi-C dati šeit ir modelēti ar neorientētiem grafiem. Gluži kā iepriekšējās publikācijās, iegūtie grafi G_t bija konstruēti no virsotnēm V_t, kas bija genomiski reģioni Hi-C kontaktu matricā, un šķautnēm E_t, kas bija mijiedarbības starp tiem, kas atbilda mūsu datu filtrēšanas kritērijiem. Šajā gadījumā dati nebija apstrādāti ar CHiCAGO procedūru, bet gan izvērtēti ar datos iekļautu p-vērtību katram kontaktam, pēc kuras tika atfiltrēti mazāk ticamās mijiedarbības, vadoties pēc vajadzīgā grafa blīvuma. Ņemot vērā, ka Hi-C grafs bija kopumā daudz blīvāks, tā izvēlētā filtrēšanas vērtība bija arī attiecīgi augstāka. Katram datu tipam tika izveidots grafs, kurā katrai šķautnei bija piekārtots audu komplekts *t*, kurā šīs šķautnes bija atrodamas. Šajos grafos tad tika meklētas un skaitītas kliķes, reģistrējot tajos iesaistītās šķautnes un virsotnes, lai apskatītu ne tikai to skaitu, bet arī izvietojumu hromosomā.

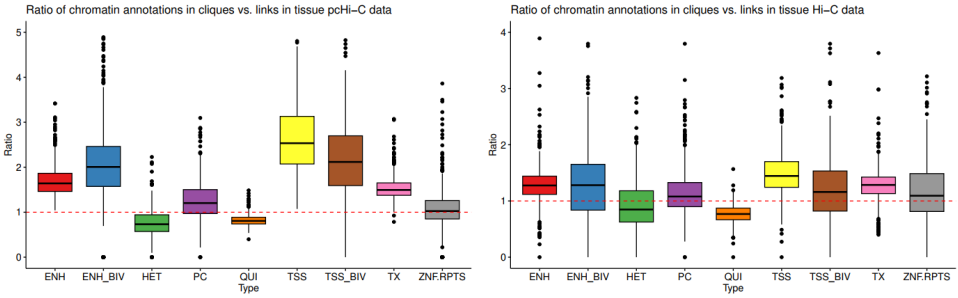
Pēc kliķu atrašanas un saskaitīšanas bija nepieciešams novērtēt to bioloģisko raksturu, un šim nolūkam tika izmantots plašs datu klāsts. Pirmkārt, ar *Ensembl* gēnu anotāciju palīdzību pētījumā lietotajiem datiem tika piesaistīti gēni, kas ļāva veikt gēnu ontoloģiju analīzi ar GOEA kliķu mijiedarbībām salīdzinot ar mijiedarbībām visā kontaktu kartējumā (Klopfenstein et al. 2018). Otrkārt, no *Roadmap Epigenomics* hromatīna stāvokļu modelēšanas datiem (Abascal et al. 2020), kas pamatojas uz epigenoma sekvenēšanas rezultātiem, tika iegūts dažādu hromatīna aktivitātes un regulācijas stāvokļu kartējums, ar kura palīdzību varēja izvērtēt, piemēram, transkripcijas sākumpunktu daudzumu kliķēs salīdzinājumā ar to kopējo sastopamību genomā un hromatīna mijiedarbību reģionos kopumā. Treškārt, izmantojot *Genotype-Tissue Expression* atlasa (Lonsdale et al. 2013) un FANTOM5 (Noguchi et al. 2017) datus, virsotnes tika sasaistītas ar pozicionālās ekspresijas datiem ar papildus nosacījumu, ka šīm ekspresijas vērtībām nebija precīzi jāsakrīt ar mijiedarbību reģionu, pamatojoties uz jaunāko literatūru, kur skaidrots, ka gēnu ekspresijas izmaiņas var notikt ar ievērojamu nobīdi no hromatīna mijiedarbības vietas (Wurmser and Basu 2022). Visbeidzot, papildus validācijas nolūkos grafa datiem arī tika piesaistīti dati no *Encyclopedia of DNA Elements* (ENCODE) projekta (Jou et al. 2019), no kuriem būtiskākā ir RNS polimerāzes II saistības vietu kartējums, kas ļāva izšķirt aktīvus gēnu ekspresijas lokusus kliķu iekšienē un pārbaudīt tos salīdzinājumā ar aktīvām transkripcijas zonām ārpus kliķēm, lai gūtu papildus izpratni par kliķu funkciju.

Rezultāti

Apstrādājot visas 3 Hi-C datu kopas un atrodot tajās kliķes izmērā 3, tika atklātas vairākas zīmīgas sakarības. Galvenā no tām bija novērojamais kliķu sakopojumu izvietojums uz hromosomām, kas sakrita ar literatūrā minētiem "hromatīna karstajiem punktiem", kas ir aktīvu gēnu sakopojumi (Liu et al.

2019). Kliķu atlase šajā gadījumā pavēra iespēju padziļinātāk pētīt šos hromatīna sakopojumus un atklāt tajos specifiskas epiģenētiskas un regulatoras sakarības.

Izmantotā daudzpakāpju validācijas stratēģija sniedza sekmīgus rezultātus šo sakarību izsekošanā. Pirmkārt, gēnu ontoloģijas analīzes parādīja, ka kliķēs ir sastopamas labi zināmas gēnu grupas, piemēram, cilvēka leikocītu antigēnu (*HLA*) grupas gēni vai ožas receptoru gēni, kuru koordinēta ekspresija ir iepriekš minēta literatūras avotos, bet arī vairākas citas gēnu grupas, kuru funkcija un koordinācija ir mazāk skaidra. Otrkārt, kliķēs bija novērojams zīmīgs hromatīna aktivitātes anotāciju palielinājums, galvenokārt transkripcijas sākumpunktu un aktīvas transkripcijas iezīmju formā, kā arī ar salīdzinoši nelielu enhānseru aktivitātes palielinājumu (sk. 4. attēlu), kas liecināja par to, ka kliķes šķietami sagrupē genoma reģionus pēc "transkripcijas rūpnīcas" principiem (Sutherland and Bickmore 2009; Mora et al. 2022), nevis promoteru-enhānseru mijiedarbību principiem. Treškārt, gēnu ekspresijas analīzēs arī bija novērojams pārliecinošs gēnu ekspresiju aktivitātes palielinājums pēc *FANTOM5* un *GTEx* datiem kliķēs salīdzinājumā ar mijiedarbību datiem kopumā. Šī ietekme kļuva vēl izteiktāka, salīdzinot RNS polimerāzes II saistības vietas kliķēs un un datos kopumā, kur kliķēs ekspresijas vērtības pat šī salīdzinājuma ietvaros bija ievērojami paaugstinātas. Tas nozīmē, ka šīs kliķes visticamāk atlasa koordinētas transkripcijas centrus šajos hromatīna mijiedarbību datos, nevis konkrētus cis-regulatoros elementus.



4. attēls. Hromatīna anotāciju palielinājums kliķēs salīdzinot ar grafu kopumā, mērīts kā aprēķināta attiecība visās hromosomās. Sarkanā pārtrauktā līnija norāda proporciju 1, un šeit kalpo kā vizuāla robeža starp palielinājumu un samazinājumu.

Šie rezultāti skaidri parāda kliķu lomu koordinētas transkripcijas nodrošināšanā, kas nozīmē, ka tālākā darba mērķis varētu būt precīzāk raksturot specifiskus elementus šo kliķu ietvaros, ko ir grūti paveikt ar šiem grafu topoloģijas datiem (jo kliķu sakopojums pēc definīcijas ir pilns vai gandrīz pilns grafs). Šo tālāko mērķi varētu būt iespējams realizēt ar papildus bioloģisku

informāciju līdzīgi RNS polimerāzes II iekļaušanai šajos datos, jo esošajā kliķu skaitīšanas sistēmā ir salīdzinoši vienkārši iekļaut jaunu genomisku informāciju. Lai turpinātu tālāku topoloģisku hromatīna datu pētījumu, ir nepieciešami sīkāki oriģināldati, visticamāk vienas šūnas Hi-C datu formātā, kur atšķirībā no parastās Hi-C datiem nav sniegs vidējots hromatīna konformācijas attēls, bet gan kontakti individuālās šūnās, kas labāk atspoguļo hromatīna dinamiku. Jebkurā gadījumā šajā pētījumā ir veiksmīgi izstrādāta grafos balstīta topoloģiska pieeja Hi-C datu analīzei, kuras ietvaros pēc topoloģiskiem kritērijiem ir atrasti hromatīna sakopojumi ar skaidru bioloģisku nozīmi, kas saskan ar literatūrā pieejamo informāciju par hromatīna funkcionalitāti. No šīs publikācijas ir iespējams attīstīt turpmākus hromatīna konformācijas pētījumus, un šo metožu pielietojums vienas šūnas Hi-C datu izpētē visticamāk sniegs vēl papildus rezultātus.

SECINĀJUMI

Šī darba ietvaros izstrādātie un izmantotie grafos balstītie biomolekulāro datu modeļi kopumā sekmēja sistemātiskas biomolekulāro tīklu struktūras un molekulu mijiedarbības principu izpratni. Galvenie secinājumi, kas gūti promocijas darba izstrādē, ir šādi:

- Ar topoloģiskām metodēm ir iespējams izšķirt paralogu gēnu pāru radniecību, pamatojoties *bi-fan* vienību simetrijā rauga un nematodes gēnu regulatorajā tīklā. Šis secinājums pamatā saskan ar teorētisko paredzējumu, ka gēnu dublikāti saglabā daļu no savām kopīgajām mijiedarbībām evolūcijas gaitā, kas ir novērojams palielināta motīvu pārklājuma formā.
- Saistības komponentes hromatīnu mijiedarbību tīklos ir labs pamats topoloģiskiem hromatīna pētījumiem, paverot iespēju pārbaudīt šūnu diferenciācijas, genoma funkcionālo moduļu un regulatoro lokusu topoloģisko saistību šajos datos.
- Lielizmēra saistības komponentes un līdzīgus lielizmēra topoloģiskus elementus ir grūti kategoriski aprakstīt to bioloģiskā nozīmīguma kontekstā, kā arī atkārtot citās datu kopās, tāpēc lietojamu bioloģisku secinājumu iegūšanai ir ieteicams lietot šaurāk definētus topoloģiskus elementus, piemēram, kliķes.
- Genomikas datu kopas (tostarp Hi-C datus) var ievērojami papildināt un interpretēt ar pozicionālu genomisku informāciju, piemēram, epiģenētiskām iezīmēm un gēnu ekspresijas datiem, lai veidotu izsmeļošākus integratīvus grafu modeļus topoloģisku īpašību skaidrošanai un tālākai pētniecībai.
- Kliķes hromatīna mijiedarbību datos norāda uz spēcīgi sasaistītiem hromatīna reģioniem. To vienkāršība un aprēķināmība gan orientētos, gan neorientētos grafos, kā arī vieglā saskaņojamība ar molekulārās bioloģijas atziņām par hromatīna struktūras saistību ar gēnu transkripciju, kā arī sasaiste ar palielinātām gēnu ekspresijām un funkcionālu iezīmju bagātinājumu, padara tās plaši lietojamas un noderīgas topoloģisku analīžu veikšanai.

Šajā darbā veiktie pētījumi var tikt tālāk attīstīti gan ar jaunu metožu piesaisti (piemēram, mašīnmācīšanās metožu pielietojumu tīklu konstruēšanai un pielāgošanai), gan ar jaunu datu pielietošanu (īpaši vienas šūnas Hi-C datu kopu analīzi), gan ar jaunām bioloģiskās informācijas integrācijas metodēm. Šie attīstības virzieni pamatojas daudzveidīgajā literatūrā, kas ir pieejama gan gēnu regulācijas, gan hromatīna mijiedarbību pētījumu nozarēs, kā arī joprojām jaunu metožu attīstībā līdz ar vienas šūnas, telpiskās transkripcijas un multiomikas datu pieejamības paplašināšanās rezultātā.

IZMANTOTĀ LITERATŪRA

Abascal, Federico, Reyes Acosta, Nicholas J. Addleman, Jessika Adrian, Veena Afzal, Rizi Ai, Bronwen Aken, et al. 2020. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." *Nature* 583 (7818): 699–710. https://doi.org/10.1038/s41586-020-2493-4.

Adams, David, Lucia Altucci, Stylianos E Antonarakis, Juan Ballesteros, Stephan Beck, Adrian Bird, Christoph Bock, et al. 2012. "BLUEPRINT to Decode the Epigenetic Signature Written in Blood." *Nature Biotechnology* 30 (3): 224–26. https://doi.org/10.1038/nbt.2153.

Alon, Uri. 2007. "Network Motifs: Theory and Experimental Approaches." *Nature Reviews Genetics* 8 (6): 450–61. https://doi.org/10.1038/nrg2102.

Babu, M. Madan, Nicholas M. Luscombe, L. Aravind, Mark Gerstein, and Sarah A. Teichmann. 2004. "Structure and Evolution of Transcriptional Regulatory Networks." *Current Opinion in Structural Biology* 14 (3): 283–91. https://doi.org/10.1016/j.sbi.2004.05.004.

Barabási, Albert László, and Zoltán N. Oltvai. 2004. "Network Biology: Understanding the Cell's Functional Organization." *Nature Reviews Genetics* 5 (2): 101–13. https://doi.org/10.1038/nrg1272.

Broido, Anna D., and Aaron Clauset. 2019. "Scale-Free Networks Are Rare." *Nature Communications* 10 (1): 1017. https://doi.org/10.1038/s41467-019-08746-5.

Byrne, Kevin P., and Kenneth H. Wolfe. 2005. "The Yeast Gene Order Browser: Combining Curated Homology and Syntenic Context Reveals Gene Fate in Polyploid Species." *Genome Research* 15 (10): 1456–61. https://doi.org/10.1101/gr.3672305.

Cairns, Jonathan, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, et al. 2016. "CHiCAGO: Robust Detection of DNA Looping Interactions in Capture Hi-C Data." *Genome Biology* 17 (1): 127. https://doi.org/10.1186/s13059-016-0992-2.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. "Capturing Chromosome Conformation." *Science* 295 (5558): 1306–11. https://doi.org/10.1126/science.1067799.

Ernst, Jason, and Manolis Kellis. 2017. "Chromatin-State Discovery and Genome Annotation with ChromHMM." *Nature Protocols* 12 (12): 2478–92. https://doi.org/10.1038/nprot.2017.124.

Harrison, Peter W, M Ridwan Amode, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Arne Becker, et al. 2024. "Ensembl 2024." *Nucleic Acids Research* 52 (D1): D891–99. https://doi.org/10.1093/nar/gkad1049.

Javierre, Biola M., Sven Sewitz, Jonathan Cairns, Steven W. Wingett, Csilla Várnai, Michiel J. Thiecke, Paula Freire-Pritchett, et al. 2016. "Lineage-Specific Genome Architecture Links Enhancers and Non-Coding Disease Variants to Target Gene Promoters." *Cell* 167 (5): 1369-1384.e19. https://doi.org/10.1016/j.cell.2016.09.037.

Jou, Jennifer, Idan Gabdank, Yunhai Luo, Khine Lin, Paul Sud, Zachary Myers, Jason A. Hilton, et al. 2019. "The ENCODE Portal as an Epigenomics Resource." *Current Protocols in Bioinformatics* 68 (1). https://doi.org/10.1002/cpbi.89.

Jung, Inkyung, Anthony Schmitt, Yarui Diao, Andrew J. Lee, Tristin Liu, Dongchan Yang, Catherine Tan, et al. 2019. "A Compendium of Promoter-Centered Long-Range Chromatin Interactions in the Human Genome." *Nature Genetics* 51 (10): 1442–49. https://doi.org/10.1038/s41588-019-0494-8.

Kellis, Manolis, Bruce W. Birren, and Eric S. Lander. 2004. "Proof and Evolutionary Analysis of Ancient Genome Duplication in the Yeast Saccharomyces Cerevisiae." *Nature* 428 (6983): 617–24. https://doi.org/10.1038/nature02424.

Kim, Kyukwang, Insu Jang, Mooyoung Kim, Jinhyuk Choi, Min-Seo Kim, Byungwook Lee, and Inkyung Jung. 2021. "3DIV Update for 2021: A Comprehensive Resource of 3D Genome and 3D Cancer Genome." *Nucleic Acids Research* 49 (D1): D38–46. https://doi.org/10.1093/nar/gkaa1078.

Klopfenstein, D. V., Liangsheng Zhang, Brent S. Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J. Mungall, et al. 2018. "GOATOOLS: A Python Library for Gene Ontology Analyses." *Scientific Reports* 8 (1): 10872. https://doi.org/10.1038/s41598-018-28948-z.

Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, et al. 2016. "Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update." *Nucleic Acids Research* 44 (W1): W90–97. https://doi.org/10.1093/nar/gkw377.

Lajoie, Bryan R, Job Dekker, and Noam Kaplan. 2015. "The Hitchhiker's Guide to Hi-C Analysis: Practical Guidelines." *Methods* 72 (January): 65–75. https://doi.org/10.1016/j.ymeth.2014.10.031.

Larsen, Simon J., Richard Röttger, Harald H.H.W. Schmidt, and Jan Baumbach. 2019. "E. Coli Gene Regulatory Networks Are Inconsistent with Gene Expression Data." *Nucleic Acids Research* 47 (1): 85–92. https://doi.org/10.1093/nar/gky1176.

Lieberman-Aiden, Erez, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Lima-Mendez, Gipsi, and Jacques van Helden. 2009. "The Powerful Law of the Power Law and Other Myths in Network Biology." *Molecular BioSystems* 5 (12): 1482. https://doi.org/10.1039/b908681a.

Liu, Li, Qian-Zhong Li, Wen Jin, Hao Lv, and Hao Lin. 2019. "Revealing Gene Function and Transcription Relationship by Reconstructing Gene-Level Chromatin Interaction." *Computational and Structural Biotechnology Journal* 17: 195–205. https://doi.org/10.1016/j.csbj.2019.01.011.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. "The Genotype-Tissue Expression (GTEx) Project." *Nature Genetics* 45 (6): 580–85. https://doi.org/10.1038/ng.2653.

Manatakis, Dimitris V, Aaron VanDevender, and Elias S Manolakos. 2021. "An Information-Theoretic Approach for Measuring the Distance of Organ Tissue Samples Using Their Transcriptomic Signatures." Edited by Pier Luigi Martelli. *Bioinformatics* 36 (21): 5194–5204. https://doi.org/10.1093/bioinformatics/btaa654.

Melkus, Gatis, Sandra Silina, Andrejs Sizovs, Peteris Rucevskis, Lelde Lace, Edgars Celms, and Juris Viksna. 2023. "Clique-Based Topological Characterization of Chromatin Interaction Hubs." In , 476–86. https://doi.org/10.1007/978-981-99-7074-2_38.

Mifsud, Borbala, Filipe Tavares-Cadete, Alice N. Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W. Wingett, et al. 2015. "Mapping Long-Range Promoter Contacts in Human Cells with High-Resolution Capture Hi-C." *Nature Genetics* 47 (6): 598–606. https://doi.org/10.1038/ng.3286.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. "Network Motifs: Simple Building Blocks of Complex Networks." *Science* 298 (5594): 824–27. https://doi.org/10.1126/science.298.5594.824.

Mora, Antonio, Xiaowei Huang, Shaurya Jauhari, Qin Jiang, and Xuri Li. 2022. "Chromatin Hubs: A Biological and Computational Outlook." *Computational and Structural Biotechnology Journal* 20: 3796–3813. https://doi.org/10.1016/j.csbj.2022.07.002.

Noguchi, Shuhei, Takahiro Arakawa, Shiro Fukuda, Masaaki Furuno, Akira Hasegawa, Fumi Hori, Sachi Ishikawa-Kato, et al. 2017. "FANTOM5 CAGE Profiles of Human and Mouse Samples." *Scientific Data* 4 (1): 170112. https://doi.org/10.1038/sdata.2017.112.

Pancaldi, Vera. 2021. "Chromatin Network Analyses: Towards Structure-Function Relationships in Epigenomics." *Frontiers in Bioinformatics* 1 (October). https://doi.org/10.3389/fbinf.2021.742216.

———. 2023. "Network Models of Chromatin Structure." *Current Opinion in Genetics & Development* 80 (June): 102051. https://doi.org/10.1016/j.gde.2023.102051.

Pržulj, Nataša. 2007. "Biological Network Comparison Using Graphlet Degree Distribution." *Bioinformatics* 23 (2): e177–83. https://doi.org/10.1093/bioinformatics/btl301.

Rao, Suhas S.P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159 (7): 1665–80. https://doi.org/10.1016/j.cell.2014.11.021.

Reece-Hoyes, John S., Carles Pons, Alos Diallo, Akihiro Mori, Shaleen Shrestha, Sreenath Kadreppa, Justin Nelson, et al. 2013. "Extensive Rewiring and Complex Evolutionary Dynamics in a C.Elegans Multiparameter Transcription Factor Network." *Molecular Cell* 51 (1): 116–27. https://doi.org/10.1016/j.molcel.2013.05.018.

Sarajlić, Anida, Noël Malod-Dognin, Ömer Nebil Yaveroğlu, and Nataša Pržulj. 2016. "Graphlet-Based Characterization of Directed Networks." *Scientific Reports* 6 (1): 35098. https://doi.org/10.1038/srep35098.

Sizovs, Andrejs, Gatis Melkus, Peteris Rucevskis, Sandra Silina, Lelde Lace, Edgars Celms, and Juris Viksna. 2024. "A Technique for Preserving Network Structure in Randomized Hi-C Data." *Journal of Bioinformatics and Computational Biology* 22 (05). https://doi.org/10.1142/S0219720024400018.

Sizovs, Andrejs, Sandra Silina, Gatis Melkus, Peteris Rucevskis, Lelde Lace, Edgars Celms, and Juris Viksna. 2024. "Exploration and Visualization Methods for Chromatin Interaction Data." In , 101–13. https://doi.org/10.1007/978-981-97-5128-0_9.

Sorrells, Trevor R., and Alexander D. Johnson. 2015. "Making Sense of Transcription Networks." *Cell* 161 (4): 714–23. https://doi.org/10.1016/j.cell.2015.04.014.

Stone, Lewi, Daniel Simberloff, and Yael Artzy-Randrup. 2019. "Network Motifs and Their Origins." Edited by Ruth Nussinov. *PLOS Computational Biology* 15 (4): e1006749. https://doi.org/10.1371/journal.pcbi.1006749.

Sutherland, Heidi, and Wendy A. Bickmore. 2009. "Transcription Factories: Gene Expression in Unions?" *Nature Reviews Genetics* 10 (7): 457–66. https://doi.org/10.1038/nrg2592.

Teixeira, Miguel Cacho, Romeu Viana, Margarida Palma, Jorge Oliveira, Mónica Galocha, Marta Neves Mota, Diogo Couceiro, et al. 2023. "YEASTRACT+: A Portal for the Exploitation of Global Transcription Regulation and Metabolic Model Data in Yeast Biotechnology and Pathogenesis." *Nucleic Acids Research* 51 (D1): D785–91. https://doi.org/10.1093/nar/gkac1041.

Tierrafría, Víctor H., Claire Rioualen, Heladia Salgado, Paloma Lara, Socorro Gama-Castro, Patrick Lally, Laura Gómez-Romero, et al. 2022. "RegulonDB 11.0: Comprehensive High-Throughput Datasets on Transcriptional Regulation in Escherichia Coli K-12." *Microbial Genomics* 8 (5). https://doi.org/10.1099/mgen.0.000833.

Vandereyken, Katy, Alejandro Sifrim, Bernard Thienpont, and Thierry Voet. 2023. "Methods and Applications for Single-Cell and Spatial Multi-Omics." *Nature Reviews Genetics* 24 (8): 494–515. https://doi.org/10.1038/s41576-023-00580-2.

Ward, Jonathan J., and Janet M. Thornton. 2007. "Evolutionary Models for Formation of Network Motifs and Modularity in the Saccharomyces Transcription Factor Network." *PLoS Computational Biology* 3 (10): 1993–2002. https://doi.org/10.1371/journal.pcbi.0030198.

Wurmser, Annabelle, and Srinjan Basu. 2022. "Enhancer-Promoter Communication: It's Not Just About Contact." *Frontiers in Molecular Biosciences* 9 (April). https://doi.org/10.3389/fmolb.2022.867303.