Proteiinirakenteiden kolmiulotteinen päällekkäin asetus ja niiden kolmiulotteisen samankaltaisuuden vertailu

TkK-tutkielma
Turun yliopisto
Tietotekniikan laitos
Tietotekniikka
2025
Sanni Knuutila

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck-järjestelmällä.

TURUN YLIOPISTO
Tietotekniikan laitos

Sanni Knuutila: Proteiinirakenteiden kolmiulotteinen päällekkäin asetus ja niiden kolmiulotteisen samankaltaisuuden vertailu

TkK-tutkielma, 21 s.
Tietotekniikka
Heinäkuu 2025

Tutkielmassa tarkastellaan proteiinirakenteiden kolmiulotteisen päällekkäinasetuksen menetelmiä ja niiden soveltuvuutta proteiinien rakenteellisen samankaltaisuuden arviointiin. Tavoitteena on vertailla olemassa olevia menetelmiä ja selvittää, mitkä niistä tarjoavat parhaan yhdistelmän tarkkuutta, laskennallista tehokkuutta ja käytännön sovellettavuutta eri bioinformatiikan tarpeissa. Tutkimus on teoreettinen ja perustuu pääasiassa tieteellisiin artikkeleihin sekä alan keskeiseen kirjallisuuteen ja menetelmäkuvauksiin.

Analyysin kohteena ovat menetelmät VAST, TM-align, DALI, GTalign ja Foldseek, jotka hyödyntävät erilaisia algoritmeja proteiinien kolmiulotteiseen kohdistamiseen. Näiden menetelmien toimintaperiaatteita ja eroja arvioidaan keskeisten mittareiden, kuten RMSD:n, TM-scoren ja Z-pisteytyksen, avulla. Lisäksi tutkielmassa pohditaan menetelmien soveltuvuutta erityyppisiin tietokantoihin ja käyttötarkoituksiin, kuten evoluutiotutkimuksiin, toiminnallisten alueiden tunnistamiseen ja suurten aineistojen seulontaan.

Tulosten perusteella GTalign osoittautuu tehokkaimmaksi työkaluksi laajoissa tietokanta-analyyseissä. DALI puolestaan säilyttää asemansa tarkkana menetelmänä biologisten homologioiden tunnistamisessa, ja TM-align tarjoaa tasapainon tarkkuuden ja suorituskyvyn välillä. Foldseek on hyödyllinen erityisesti alustavissa seulonnoissa nopeutensa ansiosta. Tutkielma korostaa kolmiulotteisen rakennevertailun keskeistä roolia bioinformatiikassa ja sen kasvavaa merkitystä nykyaikaisen rakennetiedon hallinnassa.

Asiasanat: proteiinirakenne, kolmiulotteinen rakenne, rakenteiden vertailu, päällekkäinasetus

Sisällys

1 Johdanto **1**

2 Proteiinien päällekkäin asetus **5**

3 Proteiinien vertailumenetelmiä **10**

3.1 VAST 11

3.2 DALI 12

3.3 TM-align 13

3.4 GTalign 14

3.5 FoldSeek 16

4 Pohdinta **18**

5 Yhteenveto **20**

Lähdeluettelo **22**

1 Johdanto

Proteiinien ja niiden rakenteiden määrä kasvaa jatkuvasti biologisen tutkimuksen ja sekvensointitekniikoiden kehityksen ansiosta. Protein Data Bank (PDB) on yksi esimerkki maailmanlaajuisesta tietokannasta, johon on tallennettu valtavia määriä proteiinien kolmiulotteisia rakennetietoja [1]. Tietokanta sisältää noin 232 000 tietuetta ja sinne lisätään viikoittain noin 500 uutta rakennetta [2]. Niiden analysointia varten tarvitaan tehokkaita tapoja vertailla proteiinirakenteita ja se onkin välttämätön osa bioinformatiikkaa sekä rakenteellista biologiaa. Tämän takia tiedon vertailu ja analysointi on tehokkainta suorittaa automatisoiduilla menetelmillä.

Proteiinirakenteiden vertailulla pyritään analysoimaan proteiinien kolmiulotteisia rakenteita tunnistamalla niiden samankaltaisuuksia ja eroja. Tämä mahdollistaa homologisten proteiinien tunnistamisen ja tarjoaa tärkeää tietoa niiden evolutiivisista ja toiminnallisista suhteista. Lisäksi rakenteiden tarkka kohdistaminen mahdollistaa proteiinien aktiivisten alueiden, kuten entsyymien aktiivisten keskusten ja sitoutumiskohtien, tutkimisen. [3]

Proteiinirakenteiden vertailuun ja tutkimiseen on monia syitä. Kolmiulotteinen vertailu mahdollistaa esimerkiksi proteiinien evolutiivisten suhteiden selvittämisen ja auttaa tunnistamaan homologioita, jotka voivat jäädä piiloon pelkässä sekvenssivertailussa. Esimerkiksi proteiinit, joiden sekvenssit eroavat huomattavasti, voivat jakaa samanlaisia laskostumismalleja ja toimintoja. Vertailulla voidaan tunnistaa proteiinien funktionaalisia alueita.Tarkka kohdistaminen voi paljastaa yhteisiä toi-

minnallisia ominaisuuksia, kuten aktiivisia keskuksia tai ligandin sitoutumiskohtia, mikä tukee esimerkiksi lääkeaineiden kehitystä. Rakenteiden vertailussa olennaista on nopeus ja tämän takia viime vuosina on pyritty kehittämään nopeita hakualgoritmeja, joiden avulla tunnettuja proteiinirakenteita voidaan vertailla mahdollisimman nopeasti [4]. Proteiinirakenteiden samankaltaisuuden havaitseminen kolmiulotteisen päällekkäisyyden avulla tarjoaa suuremman herkkyyden homologisten proteiinien tunnistamiseen [5].

Proteiinirakenteiden vertailuun on olemassa erilaisia mittareita ja algoritmeja, joiden pohjalta itse vertailumenetelmät toimivat. Mittareita ovat esimerkiksi RMSD eli Root Mean Square Deviation, joka mittaa kohdistettujen atomien eroavaisuuksia geometrisesti. Z-pisteytys arvioi samankaltaisuuden merkitystä tilastollisesti. Menetelmistä DALI käyttää etäisyysmatriiseja proteiinirakenteiden samankaltaisuuksien tunnistamiseen hyödyntäen heuristisia algoritmeja nopeuttaakseen laskentaa [6] [7]. TM-align optimoi Cα-atomien jakautumisen samankaltaisuuden maksimoimiseksi, yhdistäen nopeuden ja tarkkuuden [8]. VAST hyödyntää vektoreiden suhteita rakenteiden samankaltaisuuden tunnistamiseksi ja evolutiivisten suhteiden analysoimiseksi [9]. FoldSeek muuntaa proteiinirakenteet rakenteellisiksi aakkosiksi ja suorittaa vertailun sekvenssipohjaisesti, mikä mahdollistaa huomattavasti nopeamman analyysin [5]. GTalign on uusi ja tehokas menetelmä, joka käyttää spatiaalista indeksointia ja k-d-puu-rakenteita proteiinirakenteiden nopeaan ja tarkkaan kohdistamiseen suurissa tietokannoissa, kuten AlphaFoldissa. Se sopii erityisesti laajamittaiseen rakenteelliseen analyysiin ja lääkekehitykseen [10]. Muita menetelmiä ovat esimerkiksi DaliLite, joka tarjoaa kevyemmän version DALI-algoritmista.

Tässä työssä perehdytään erilaisiin proteiinien vertailumenetelmiin, mihin algoritmeihin ja tekniikoihin ne perustuvat, sekä selvitetään niiden tehokkuutta, nopeutta ja soveltuvuutta tiettyihin sovelusalueihin. Tutkimuksen tavoitteena on vertailla käytettyjä menetelmiä, arvioida niiden suorituskykyä ja tunnistaa tilanteita, jois-

sa tietyt lähestymistavat ovat erityisen hyödyllisiä. Lisäksi työssä selvitetään mitkä näistä menetelmistä tarjoavat parhaimman yleiskatsauksen vertailumenetelmistä ja niiden soveltuvuudesta erilaisiin tarpeisiin. Proteiinirakenteiden kolmiulotteinen vertailu on elintärkeä työkalu molekyylibiologiassa. Se ei ainoastaan paranna ymmärrystä proteiinien toiminnasta ja evoluutiosta, vaan myös tukee lääketieteellisiä sovelluksia, kuten lääkkeiden kohdespesifisyyden suunnittelua.

1. Tutkimuskysymys: Mitä eri mittareita on käytetty kolmiulotteisen samankaltaisuuden kvantitointiin ja 3D-samankaltaisuuteen perustuviin hakuihin? Onko niiden joukossa toisista poikkeavia lähestymistapoja, vai ovatko kaikki vain muunnelmia etäisyyksien käytöstä?

2. Tutkimuskysymys: Mitä hyötyä on proteiinirakenteiden vertailusta?

3. Tutkimuskysymys: Mitä kolmiulotteisen rakenteen vertailun menetelmiä pidetään parhaina?

Tämä kirjallisuuskatsaus toteutettiin hyödyntäen Turun yliopiston Volter-kirjastoa ja Google Scholaria. Lopullisten tutkimusartikkelien hakemisessa käytettiin hakulauseita, jotka muodostettiin yhdistämällä alan keskeisiä käsitteitä loogisilla operaattoreilla. Hakuprosessissa käytettiin erikseen englannin- ja suomenkielisiä käsitteitä: esimerkiksi englanniksi protein OR proteins AND superimposition OR optimization AND ”structural superimposition” OR ”structural similarity” AND ”3D compari*” sekä suomeksi proteiini OR proteiinien AND asettelu OR päällekkäin asettelu OR ”3D comparison”. Hakulauseet muodostettiin yhdistämällä kolmen ja kahden käsitteen yhdistelmiä ja suomen- ja englanninkieliset haut tehtiin erillisinä.

Lähdeaineistoa haettiin erityisesti kansainvälisistä tieteellisistä lehdistä, joihin on viitattu usein alan tutkimuksessa. Mukaan valittiin julkaisuja, jotka esittelivät tunnettuja ja ajankohtaisia proteiinirakenteiden vertailumenetelmiä sekä artikkeleita, jotka kuvasivat näiden menetelmien tehokkuutta, sovellusalueita ja rajoituk-

sia. Karsintakriteereinä käytettiin toistuvuutta muissa lähteissä, julkaisuvuotta, sekä menetelmän relevanssia bioinformatiikan nykytarpeisiin, kuten suurten tietokantojen analysointiin. Vanhemmat tai menetelmällisesti vanhentuneet julkaisut rajattiin pois, ellei niillä ollut historiallista merkitystä vertailun kannalta.

Proteiinirakenteiden vertailumenetelmät ovat säilyttäneet yllättävän hyvin asemansa siitä lähtien, kun tunnetuimmat algoritmit kuten TM-align ja DALI kehitettiin 2000-luvun alkuvuosina [8] [2]. Näiden menetelmien kyky tunnistaa evolutiivisia yhteyksiä ja rakenteellista samankaltaisuutta on osoittautunut kestäväksi. Vaikka uusia ratkaisuja, kuten FoldSeek [5] ja GTalign [10], on kehitetty viime vuosina erityisesti nopeuden ja skaalautuvuuden parantamiseksi, klassisille menetelmille ei ole ollut helppoa löytää korvaajia niiden biologisen tulkittavuuden, herkkyyden ja tarkkuuden vuoksi [4]. Tämän vuoksi tutkimuksessa joudutaan yhä usein turvautumaan yhdistelmästrategioihin, joissa perinteisiä ja uusia menetelmiä hyödynnetään rinnakkain.

2 Proteiinien päällekkäin asetus

Proteiinit ovat elintärkeitä suuria molekyylejä, jotka koostuvat aminohapoista, joita on luonnollisesti enintään 20 erilaista. Aminohapot kiinnittyvät toisiinsa peptidisidoksin muodostaen polypeptidiketjuja. Ketjujen sidokset saavat aikaan ketjun järjestäytymisen kolmiulotteiseksi rakenteeksi. Proteiinirakenne voidaan jakaa neljään eri tasoon.

Ensimmäinen taso proteiinirakenteessa on primaarirakenne, joka määrittää aminohappojen lineaarisen järjestyksen polypeptidiketjussa. Tämä järjestys perustuu geneettisen DNA:n emäsjärjestykseen, joka ohjaa proteiinin synteesiä soluissa. Rakenne syntyy aminohappojen kiinnittyessä kovalenttisilla sidoksilla tosiinsa. Primaarirakenteen merkitys korostuu, sillä se muodostaa perustan proteiinien rakenteelle, sen toiminnalle ja pohjan seuraaville rakennetasoille. Sekundaarirakenne keskittyy proteiiniketjun paikallisiin kolmiulotteisiin muotoihin. Kaksi yleistä sekundaarirakennetta ovat α-kierre ja β-levy. α-kierteessä aminohapot muodostavat spiraalimaisen rakenteen, kun taas β-levyt ovat litteitä rakenteita, jotka muodostuvat β-juosteista. Nämä rakenteet syntyvät vety- ja ionisidoksista sekä hydrofobisista vuorovaikutuksista, ja ne muodostavat proteiinille sen ensimmäisen varsinaisen kolmiulotteisen rakenteen.

α-kierteen muodostumista ohjaavat ϕ- ja ψ-kiertokulmat, jotka mahdollistavat selkärangan atomien tiiviin sijoittumisen. Tämä rakenne mahdollistaa vetysidosten muodostumisen tiettyjen runkoatomien välillä. Vetysidokset syntyvät, kun amino-

happojäännöksen karbonyyliryhmän happiatomi muodostaa sidoksen polypeptidiketjussa neljän jäännöksen päässä olevan aminoryhmän vetymolekyylin kanssa.

α-kierteessä vetysidokset muodostuvat proteiinin selkärangan karbonyyliryhmän (C=O) ja neljän jäännöksen päässä olevan aminoryhmän (N–H) välillä. Näiden vetysidosten suuntautuminen johtuu happiatomin suuresta elektronegatiivisuudesta, mikä aiheuttaa sidoksille sisäisen varausjakauman eli polarisaation. Tämän seurauksena α-kierteeseen syntyy dipolimomentti, jossa aminopää on keskimäärin negatiivinen ja karboksyylipää positiivinen. α-kierteen alku- ja loppupäässä vetysidoksia ei yleensä pääse muodostumaan. Erityisesti ensimmäiset neljä NH-ryhmää ja viimeiset neljä CO-ryhmää jäävät ilman selkärangan vetysidoksia, jotka normaalisti vakauttavat rakennetta. Tämän seurauksena lyhyet α-kierteet voivat olla alttiita rakenteellisille vääristymille tai muodostaa vaihtoehtoisia vetysidoksia muiden ryhmien kanssa.

Tertiaarirakenne kuvaa koko proteiinin lopullista kolmiulotteista muotoa, joka syntyy aminohapposivuketjujen välisistä vuorovaikutuksista. Näitä ovat muun muassa vetysidokset, ionisidokset, hydrofobiset vuorovaikutukset ja disulfidisillat. Näiden yhteisvaikutus saa polypeptidiketjun taittumaan ja kiertymään, muodostaen monimutkaisen rakenteen, joka on ratkaiseva proteiinin toiminnalle. Tertiaarirakenteen muutokset voivat vaikuttaa merkittävästi proteiinin biologiseen aktiivisuuteen.

Kvaternaarirakenne puolestaan viittaa useamman aminohappoketjun eli monomeerin järjestäytyneeseen vuorovaikutukseen. Monomeerit voivat olla identtisiä tai toisistaan poikkeavia, ja niiden välinen sitoutuminen perustuu samoihin eikovalenttisiin ja kovalenttisiin vuorovaikutuksiin kuin tertiaarirakenteessa. Näiden avulla muodostuu toiminnallisia proteiinikomplekseja, joiden rakenne ja yhteistyö ovat välttämättömiä monimutkaisten biologisten prosessien kannalta. [3]

Proteiinien rakenteita voidaan sovittaa yhteen 3D-mallinnuksen avulla. Yksi yleisimmin käytetyistä tavoista on asettaa kahden proteiinin Cα-atomit mahdollisim-

man lähelle toisiaan. Rakenteiden samankaltaisuutta voidaan arvioida mittaamalla atomien etäisyyksiä. Pienet etäisyydet viittaavat rakenteelliseen samankaltaisuuteen, kun taas suuremmat etäisyydet osoittavat proteiinien välillä olevan rakenteellisi eroja. Vaikka Cα-atomien käyttö on yleinen lähestymistapa, myös muita sovitusmenetelmiä on olemassa.

Proteiinirakenteiden vertailulla on mahdollista selvittää proteiinien laskostumismalli, joka antaa proteiineille niiden kolmiulotteisen rakenteen. Vertailulla voidaan etsiä myös proteiineista erilaisia ja samankaltaisia alueita, joihin entsyymien ”liikkuvat osat” eli substraatit ja inhibiittorit voivat kiinnittyä. Lisäksi proteiinien päällekkäisyydellä voidaan hyödyntää erilaisten, mutta samankaltaisten proteiinien visualisointia. Tämä mahdollistaa proteiinikompleksien muokkaamisen uusilla komponenteilla sekä uusien komponenttien asettamisen olemassa olevien kompleksien päälle.

Kolmiulotteisella proteiinien rakennevertailulla voidaan myös mahdollistaa proteiinirakenteiden evoluution tarkastelu kauas taaksepäin ja näin on mahdollista löytää yhteneväisyyksiä eri proteiiniperheille. Yhteneväisyyksiä voi löytyä myös proteiineille, jotka näyttävät erilaiselta esimerkiksi sekvenssivertailua hyödyntäen. Esimerkkinä kierteisten sytokiinien sekvenssit eivät ole havaittavasti samakaltaisia, mutta rakennevertailulla voidaan huomata niiden samakaltaisia reseptoreita ja signaalireittejä. Näin voidaan todeta niiden kuuluvan samaan sukulinjaan. [8]

Proteiinien tutkimiseen käytetään usein rakenteellisia menetelmiä, jotka mahdollistavat proteiinien päällekkäin asettelun ja niiden kolmiulotteisen rakenteen yksityiskohtaisemman vertailun. Tavoitteena on määritellä kahden proteiinin geometrinen rakenne mahdollisimman tarkasti usean pisteen avulla ja hyvin pienin poikkeamin [2]. Seuraavassa tutkielman osassa tuodaan esille käytetyimpiä kolmiulotteisia proteiinien vertailumenetelmiä, niiden tärkeimpiä käyttötarkoituksia ja hyötyjä bioinformatiikassa.

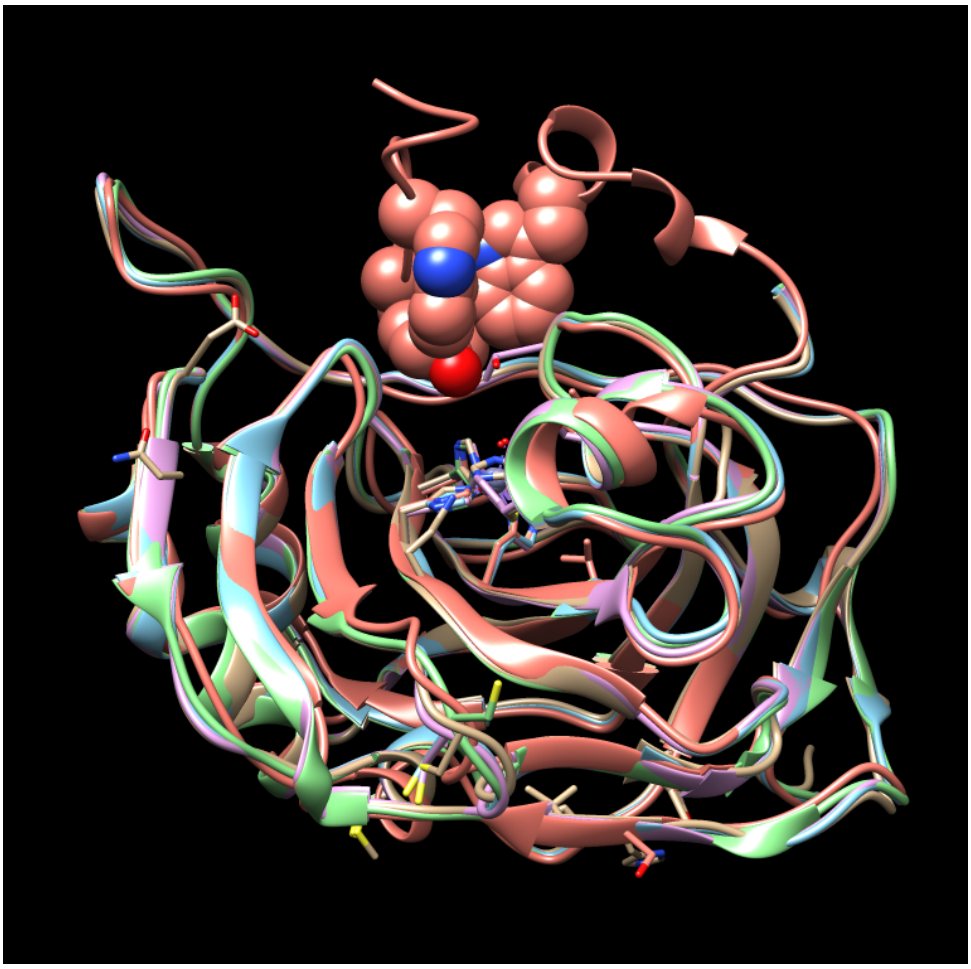
Kun kahden domeenin tertiaarirakenne on samankaltainen, se tarkoittaa, että niiden sekundaarirakenteen elementit (SSE engl. Secondary Structure Element) asettuvat avaruudessa samanlaiseen järjestykseen. Tämä edellyttää rakenteiden päällekkäisyyden sekä ketjujen liitettävyyden yhteneväisyyden. Artikkelissa ”Analysis of protein homology by assessing the (dis)similarity in protein loop regions” tarkastellaan rakenteellista samankaltaisuutta kuuden mittarin avulla. [11]

Prosentuaalinen identiteetti kuvaa aminohappojen vastaavuusprosenttia kohdistettujen alueiden välillä. RMSD mittaa kohdistettujen atomien keskimääräisen geometrisen eroavuuden. Kohdistettujen alueiden osuus määritellään suhteuttamalla kohdistettujen alueiden aminohappojen määrä pienemmän alueen aminohappojen kokonaismäärään.

Silmukoiden pituudet mittaavat, kuinka pitkiksi silmukka-alueet muodostuvat rakenteessa. Kohdistettujen silmukoiden osuus ilmaisee kohdistettujen silmukoiden suhteellisen osuuden aminohappojen määrästä. LHM (Loop Homology Measure) arvioi silmukka-alueiden yhdenmukaisuutta ja samankaltaisuutta.

RMSD lasketaan VAST-kohdistuksille käyttäen McLachlanin algoritmia, ja normalisoitu RMSD saadaan jakamalla alkuperäinen RMSD kohdistettujen aminohappojen osuudella. Kohdistettujen aminohappojen osuus puolestaan määritetään suhteuttamalla kohdistettujen ja pienemmän alueen aminohappojen lukumäärä. [11]

Kuvassa 2.1 havainnollistetaan CA5-proteiinin lyhyemmän sekvenssin vaikutusta rakenteeseen verrattuna täysipitkään CA2:een. Päällekkäin asetetuista rakenteista nähdään, että puuttuva N-terminaalialue ei vaikuta jäljelle jäävän rakenteen laskostumiseen, mikä viittaa rakenteelliseen vakauteen ja funktionaalisesti säilyneisiin osiin.



Kuva 2.1: CA5-proteiinin neljä mallinnettua rakennetta (sininen, vihreä, beige ja lila) päällekkäin asetettuina CA2-proteiinin kanssa (punainen). Kuvassa näkyy, että CA5 on sekvenssiltään hieman lyhyempi kuin muut ihmisen hiilihappoanhydraasit, erityisesti puuttuvan N-terminaalialueen osalta. CA2:ssa tämä alkupään osa muodostaa selkeästi erillisen pallomaisen laskostuman (kuvan yläoikealla), jota ei esiinny CA5-malleissa. Tästä huolimatta CA5-proteiinit säilyttävät hyvin samankaltaisen rakenteen vertailukelpoisilla alueilla, mikä viittaa siihen, ettei N-terminaalin puuttuminen haittaa muun rakenteen laskostumista.

3 Proteiinien vertailumenetelmiä

Proteiinien kohdistaminen tai päällekkäin asettaminen on keskeinen menetelmä, jolla vertaillaan proteiinien aminohapposekvenssejä tai kolmiulotteisia rakenteita niiden samankaltaisuuksien ja erojen tunnistamiseksi. Tämä prosessi on olennainen osa proteiinien evolutiivisten suhteiden, rakenteellisten samankaltaisuuksien ja toiminnallisten merkitysten selvittämistä. Kohdistusmenetelmien avulla voidaan esimerkiksi tutkia, missä määrin tietyt toiminnalliset motiivit ovat säilyneet eri lajien välillä tai saman proteiiniperheen sisällä. Näin ollen proteiinien kohdistaminen on perusvaihe bioinformatiikassa ja rakenteellisessa biologiassa, kun pyritään ymmärtämään proteiinien ominaisuuksia ja suhteita toisiinsa.

Proteiinien kohdistaminen voidaan jakaa kahteen päämenetelmään: sekvenssirinnastukseen (engl. *protein alignment*) ja kolmiulotteiseen päällekkäinasetukseen (engl. *protein superposition*). Sekvenssirinnastus tarkastelee aminohappojärjestysten samankaltaisuutta, kun taas päällekkäinasetus vertailee proteiinien 3D-rakenteita. Päällekkäinasetusmenetelmiin kuuluu esimerkiksi rakenteellinen superpositio, jossa pyritään minimoimaan atomien koordinaattien välinen etäisyys maksimaalisen yhteensopivuuden saavuttamiseksi. Toinen yleinen lähestymistapa on SSM (Secondary Structure Matching), jossa proteiinirakenteet kohdistetaan niiden sekundaarirakenteiden — kuten α-kierteiden ja β-levyjen — perusteella.

Erilaiset kohdistusmenetelmät lähestyvät ongelmaa eri tavoin. TM-align optimoi kahden proteiinirakenteen päällekkäisyyttä maksimoimalla niiden Cα-atomien

jakauman samankaltaisuuden. VAST (Vector Alignment Search Tool) hyödyntää vektoripohjaista representaatiota, jossa proteiinien sekundaarirakenteet kuvataan vektoreina ja näiden suhteiden avulla tunnistetaan samankaltaisia alueita. DALI (Distance-matrix Alignment) puolestaan perustuu etäisyysmatriiseihin, jotka lasketaan Cα-atomien välisten etäisyyksien perusteella. DALI käyttää iteratiivista algoritmia löytääkseen rakenteellisesti parhaiten yhteensopivat alueet.

Uusien rakenteellisen samankaltaisuuden hakumenetelmien perusta on usein tilastollinen, mutta niiden lähestymistavat merkitsevyysrajojen määrittelyyn voivat erota huomattavasti toisistaan. Esimerkiksi empiirisissä menetelmissä kohdistusten pistemäärä suhteutetaan satunnaisesti valittujen, samankokoisten proteiinien pistemäärien jakaumaan. Tämä ottaa huomioon sen, että suuremmissa proteiineissa vaihtoehtoisten sekundaarirakennekohdistusten määrä kasvaa. Toisaalta aminohappopohjaiset menetelmät säätävät pistemäärää sen perusteella, kuinka suuri osa aminohapoista sisältyy kohdistukseen. Tämä voi johtaa alempiin pistemääriin, jos kohdistus kattaa vain osan rakenteista, sillä suuri osa aminohapoista jää tällöin kohdistamatta. Tällaiset erot vaikuttavat menetelmien kykyyn arvioida biologisesti merkittäviä rakenteellisia samankaltaisuuksia [4].

3.1 VAST

VAST (Vector Alignment Search Tool) on tehokas menetelmä proteiinien rakenteellisen samankaltaisuuden vertailuun. Se perustuu sekundaarirakenteiden, kuten α-kierteiden ja β-juosteiden, esittämiseen kolmiulotteisina vektoreina, joiden suunta ja pituus kuvastavat rakenteen avaruudellista rakennetta. Kohdistus perustuu näiden vektorien välisiin geometrisiin suhteisiin, joita analysoidaan tehokkailla hakualgoritmeilla. Menetelmä soveltuu erityisesti proteiinien säilyneiden rakenteellisten alueiden tunnistamiseen, esimerkiksi evolutiivisten suhteiden tai toiminnallisten domeenien tutkimuksessa.

VAST on hierarkkinen algoritmi, joka koostuu kahdesta päävaiheesta. Ensimmäisessä vaiheessa sekundaarirakenteen vektorit kohdistetaan Bron–Kerbosch-algoritmilla, jolla etsitään vektoriparit, joilla on pieni RMSD-arvo. Toisessa vaiheessa kohdistus laajennetaan aminohappotasolle käyttäen Gibbs-sampling-tekniikkaa, jonka avulla voidaan tunnistaa merkittäviä paikallisia samankaltaisuuksia ilman täsmällistä kokonaista kohdistusta [9].

Kohdistusprosessin tuloksena syntyy joukko rakennekohdistuksia, joiden pisteet klusteroidaan. Näiden perusteella muodostetaan rotaatiomatriisi, jonka avulla proteiinien rakenteet asetetaan kolmiulotteisesti päällekkäin. Menetelmä ottaa huomioon hakuavaruuden koon ja laskee p-arvon todennäköisimmälle alarakenteen superpositiolle. Tämä tilastollinen p-arvo mahdollistaa merkityksellisten rakenneyhtäläisyyksien erottamisen satunnaisista kohdistuksista [4].

3.2 DALI

DALI (Distance-matrix Alignment) on laskennallinen työkalu, joka on suunniteltu tunnistamaan ja vertailemaan proteiinirakenteiden yhtäläisyyksiä. Sen ensisijaisena tavoitteena on paljastaa rakenteellisia samankaltaisuuksia, jotka voivat viitata proteiinien toiminnallisiin tai evolutiivisiin suhteisiin [6].

Menetelmän ydin perustuu etäisyysmatriisien laskentaan. Näissä matriiseissa Cα-atomien väliset etäisyydet lasketaan ja tallennetaan matemaattiseen muotoon, minkä jälkeen eri proteiinien matriiseja verrataan toisiinsa. Näin voidaan tehokkaasti tunnistaa rakenteellisesti samankaltaiset alueet, vaikka aminohapposekvenssit poikkeaisivat toisistaan. Tämä tekee DALI:sta erityisen käyttökelpoisen, sillä proteiinien kolmiulotteinen rakenne säilyy yleensä paremmin evoluution aikana kuin niiden primäärisekvenssi.

Samankaltaisuuksien kvantifioinnissa käytetään RMSD-arvoja, jotka mittaavat rakenteiden päällekkäisyyttä kolmiulotteisesti. Tulokset pisteytetään z-arvolla, jonka

avulla voidaan erottaa tilastollisesti merkittävät yhtäläisyydet satunnaisista. Korkea z-pisteytys viittaa siihen, että kahden proteiinin välinen samankaltaisuus on epätodennäköistä sattumaa ja siten todennäköisesti biologisesti merkittävää.

DALI on saatavilla verkkopalvelimena, joka tarjoaa heuristisia optimointeja suurten tietokantojen läpikäyntiin. Palvelimen käyttöliittymä mahdollistaa myös rakennekohdistusten ja kolmiulotteisten superpositioiden visualisoinnin. Tämä tekee työkalusta käytännöllisen erityisesti uusien rakenteellisten ja toiminnallisten löydösten analysointiin. [6] [7]

DALIn rinnalla toimii myös DaliLite, joka on sen pohjalta kehitetty itsenäinen ja kevyempi versio. DaliLite keskittyy tehokkaaseen rakenteelliseen vertailuun erityisesti proteiinien sekundaarirakenteiden perusteella.

3.3 TM-align

TM-align (Template Modeling alignment) on laajalti käytetty proteiinirakenteiden kohdistusalgoritmi, joka yhdistää TM-pisteytyksen, kiertomatriisit ja dynaamisen ohjelmoinnin tehokkaaksi laskentamenetelmäksi. Menetelmä perustuu Cα-atomien koordinaatteihin ja hyödyntää niiden välisten etäisyyksien vaihtelua tunnistaakseen proteiinien rakenteelliset samankaltaisuudet. TM-alignin erityinen vahvuus on sen nopeus ja tarkkuus. Se on noin 20 kertaa nopeampi kuin esimerkiksi DALI, ja sen avulla saavutetut kohdistukset ovat keskimäärin tarkempia [8].

Algoritmin toiminta koostuu kolmesta päävaiheesta. Ensimmäisessä vaiheessa hyödynnetään dynaamista ohjelmointia sekundaarirakenteiden (SS) vertailuun. Kohdistus perustuu pisteytysmatriisiin, jossa kaksi SS-elementtiä saa arvon 1, jos ne ovat identtiset (esimerkiksi α-kierre tai β-levy), ja arvon 0, jos ne poikkeavat toisistaan. SS-tilat määritellään Cα-koordinaattien perusteella viiden vierekkäisen aminohapon muodostamasta segmentistä. Kohdistuksen johdonmukaisuus varmistetaan hienosäätämällä SS-elementtejä yhdistämällä samankaltaisia tiloja ja poista-

malla epäyhtenäisiä rakenteellisia poikkeamia.

Toisessa vaiheessa algoritmi etsii optimaalisen rakenteellisen vastaavuuden proteiinien välillä. Tämä tapahtuu käyttäen merkkijonohakua ja TM-pisteytystä. Tavoitteena on löytää aukoton kohdistus, jossa pienempi proteiini asemoidaan suurempaan rakenteeseen mahdollisimman tarkasti. TM-score tarjoaa tässä vaiheessa pisteytysperustan, joka huomioi tehokkaasti aminohappojen väliset etäisyyserot.

Kolmannessa vaiheessa dynaaminen ohjelmointi toistuu, mutta nyt yhdistetään kaksi eri matriisia: SS-elementtien pisteytysmatriisi ja kolmiulotteinen etäisyysmatriisi. Näiden perusteella lasketaan entistä tarkempi kohdistus. Alustava kohdistus syötetään heuristiseen iteratiiviseen algoritmiin, jossa transformaatiomatriiseja hienosäädetään usean iteraation ajan TM-scoren maksimoimiseksi. Uudet kiertomatriisit ja pisteytysmatriisit parantavat kohdistusta kerta toisensa jälkeen, kunnes rakenneyhteensopivuus vakiintuu. Käytännössä tarkka kohdistus saavutetaan usein 2–3 iteraation jälkeen.

Kohdistuksen aikana otetaan huomioon myös aukkojen käsittely. Tätä varten käytetään aukonavausvirhepisteitä, jotka minimoivat aukkojen haitallisen vaikutuksen kohdistuksen laatuun. Lisäksi on ehdotettu vaihtoehtoisia paikallisia malleja, esimerkiksi yhteistyötermin (cooperativity term) hyödyntämistä, erityisesti silloin kun kohdistus tapahtuu SS-elementtien sisällä kuten α-kierteissä ja β-juosteissa.

TM-align on vakiinnuttanut asemansa keskeisenä työkaluna proteiinien rakenteellisessa vertailussa. Sen erinomainen tasapaino tarkkuuden, suorituskyvyn ja skaalautuvuuden välillä tekee siitä sopivan menetelmän laajamittaisiin rakenteellisiin analyyseihin sekä proteiinien evolutiivisten yhteyksien selvittämiseen. [8]

3.4 GTalign

GTalign (Giga-scale Targeted alignment) on moderni proteiinirakenteiden kohdistusmenetelmä, joka on suunniteltu erityisesti suurten tietokantojen, kuten Alpha-

Foldin, tehokkaaseen ja tarkkaan analysointiin. Menetelmä hyödyntää spatiaalista indeksointia ja iteratiivista TM-score-optimointia saavuttaakseen sekä korkean tarkkuuden että erittäin nopean laskennan [10].

GTalignin ytimessä on k-d-puu-rakenne (engl. *k-dimensional tree*), joka järjestää Cα-atomien koordinaatit siten, että lähimmän naapurin haku voidaan suorittaa laskenta-ajassa $O(1)$. Tämä mahdollistaa superpositioiden laskennan huomattavasti perinteisiä menetelmiä nopeammin ja skaalaa hyvin massiivisten tietoaineistojen käsittelyyn.

Kohdistusprosessi alkaa karkealla superpositiolla, jota seuraa iteratiivinen hienosäätö transformaatiomatriiseja optimoimalla. Näissä vaiheissa TM-score toimii kohdistuksen laadun mittarina, ja sen maksimoiminen mahdollistaa proteiinien rakenteiden yhteensopivuuden jatkuvan parantamisen. Kohdistuksen pituus ja sijainti mukautuvat dynaamisesti jokaisessa vaiheessa, mikä lisää joustavuutta ja tarkkuutta erityisesti rakenteellisesti monimuotoisissa proteiineissa.

GTalign sisältää kaksivaiheisen esisuodatuksen vähentääkseen laskennallista kuormitusta. Ensimmäisessä vaiheessa sekvenssipohjainen suodatus arvioi samankaltaisuudet proteiinisekvenssien perusteella ja sulkee pois ilmeisen epärelevantit parit. Toisessa vaiheessa käytetään alustavaa TM-score-arviointia rakenteellisen samankaltaisuuden seulontaan ennen tarkempaa kohdistusta.

Lisäksi GTalign tukee rinnakkaislaskentaa ja hyödyntää GPU-kiihdytystä, mikä mahdollistaa nopean analyysin suurista tietomääristä, kuten virusten kapsidirakenteista tai koko proteomiin perustuvista tietokannoista. Tämä tekee siitä erityisen hyödyllisen työkalun bioinformatiikan skaalautuviin haasteisiin ja lääketieteellisiin sovelluksiin, kuten lääkekehitykseen.

3.5 FoldSeek

Proteiinien rakenteen ennustusmenetelmät ovat tuottaneet valtavia määriä julkisesti saatavia proteiinirakenteita, mikä on lisännyt tarvetta tehokkaille rakenteellisille hakumenetelmille. Näiden tietokantojen kattava analysointi on erittäin aikaa vievää, ja perinteiset työkalut eivät skaalaudu riittävän nopeasti vastaamaan nykyaikaisten tietokantojen vaatimuksia.

FoldSeek on kehitetty ratkaisemaan tämä ongelma, ja sen on todettu lyhentävän laskenta-aikaa jopa neljästä viiteen kertaluokkaa verrattuna perinteisiin rakennekohdistusmenetelmiin. Lisäksi sen herkkyys on raportoitu olevan 86 %, 88 % ja 133 % verrattuna DALI:n, TM-alignin ja CE:n vastaaviin herkkyyksiin.

Klassiset rakennekohdistimet, kuten TM-align ja DALI, perustuvat iteratiivisiin tai stokastisiin optimointimenetelmiin, jotka ovat laskennallisesti raskaita. Esimerkiksi TM-alignin avulla yhden rakenteen etsiminen 100 miljoonan proteiinirakenteen tietokannasta vaatisi kuukauden yhdellä CPU-ytimellä, ja kaikkien proteiinien välinen vertailu kestäisi noin 10 000 vuotta 1 000 ytimen klusterilla [5]. Vastaavasti sekvenssipohjainen haku, kuten MMseqs2:lla, suoriutuu samasta tehtävästä noin viikossa samalla laskentakapasiteetilla.

Rakennehakutyökalujen hitaus johtuu kahdesta keskeisestä syystä. Ensinnäkin sekvenssihakutyökalut hyödyntävät nopeita ja herkkiä esisuodatusalgoritmeja, joita ei ole yhtä laajalti kehitetty rakennevertailuun. Toiseksi rakenteellinen samankaltaisuuspisteytys on epäpaikallista: yhden osan kohdistuksen muuttaminen vaikuttaa pisteytykseen koko rakenteessa, mikä lisää optimoinnin monimutkaisuutta.

FoldSeek ratkaisee tämän pullonkaulan muuttamalla proteiinien kolmiulotteiset selkärangat lineaarisiksi sekvensseiksi niin kutsutussa rakenteellisessa aakkostossa. Näin rakennevertailu voidaan muuntaa huomattavasti nopeammiksi sekvenssikohdistuksiksi. Tämä lähestymistapa mahdollistaa jopa satojen miljoonien proteiinien vertailun realistisessa ajassa. Rakenteelliset aakkoset muodostetaan diskretoimalla

paikallisia aminohapposelkärangan konformaatioita, usein 3–5 Cα-atomin jaksoissa. Useita diskretisointimenetelmiä on kehitetty, mukaan lukien CLE, 3D-BLAST ja Protein Blocks [5].

4 Pohdinta

GTalign osoittautui parhaaksi menetelmäksi suurten proteiinitietokantojen analysoinnissa, tarjoten samanaikaisesti korkeaa tarkkuutta ja huomattavaa nopeutta. Se saavutti selvästi paremmat TM-scoret ja kattavuuden verrattuna TM-aligniin ja DALIin, erityisesti massiivisten tietokantojen, kuten AlphaFoldin rakenteiden, kohdalla. GTalignin suorituskyky nopeudessa oli ylivoimainen: se oli jopa 104–1424 kertaa nopeampi kuin TM-align, mikä tekee siitä erityisen sopivan laajamittaisiin rakenteellisiin analyyseihin ja lääkekehityksen tarpeisiin. Foldseek puolestaan suoriutui parhaiten silloin, kun analyysin ensisijaisena tavoitteena oli nopeus, mutta sen tarkkuus ei riittänyt monimutkaisissa vertailuissa. TM-align säilytti edelleen käyttökelpoisuutensa kompromissina tarkkuuden ja suorituskyvyn välillä, mutta jäi selvästi GTalignin varjoon molemmilla osa-alueilla.

Menetelmien soveltuvuus vaihtelee käyttötarkoituksen mukaan. GTalignin ja DALIn vahvuudet korostuivat erityisesti rakenteellisten homologioiden ja analogioiden tunnistamisessa tilanteissa, joissa sekvenssipohjaiset menetelmät eivät tuota luotettavia tuloksia. DALI:n Z-pisteytys antaa biologisesti merkityksellisen arvion rakenteiden samankaltaisuudesta, mikä tekee siitä edelleen arvokkaan työkalun erityisesti evoluutiotutkimuksessa. Foldseekin etuna on sen käyttö alustavissa seulonnoissa, mutta tarkempi analyysi vaatii edelleen menetelmiä kuten GTalign tai DALI.

Käytetyt mittarit vaikuttivat huomattavasti analyysin luotettavuuteen. TM-score osoittautui tehokkaaksi globaalin samankaltaisuuden mittaamisessa, mutta sen

herkkyys lyhyille proteiineille jäi heikommaksi. RMSD-mittari puolestaan kärsi kohdistuspituuteen liittyvästä vääristymästä, mikä rajoittaa sen sovellettavuutta tarkassa analyysissä. GTalignin hyödyntämä iteratiivinen TM-score-optimointi tuotti johdonmukaisesti tarkempia ja luotettavampia tuloksia verrattuna perinteisiin, staattisiin lähestymistapoihin.

Tulevaisuuden kehityksessä GTalignin algoritmin parantaminen lyhyiden proteiinien kohdalla muodostuu keskeiseksi tavoitteeksi. Samoin joustavien rakenteiden vertailu edellyttää uusia lähestymistapoja, sillä nykyiset menetelmät eivät kykene tehokkaasti käsittelemään konformaatioltaan vaihtelevia alueita. Matalan tarkkuuden laskennan hyödyntäminen esikarsintaan ja tekoälypohjaisten menetelmien integrointi tarjoavat lupaavia mahdollisuuksia sekä nopeuden että tarkkuuden lisäämiseksi. Näiden kehityssuuntien edistäminen voisi merkittävästi laajentaa rakenteellisen vertailun sovellettavuutta ja syventää biologista ymmärrystä.

5 Yhteenveto

Proteiinirakenteiden vertailu on yksi bioinformatiikan keskeisimmistä menetelmistä, joka mahdollistaa proteiinien evolutiivisten suhteiden, toiminnallisten ominaisuuksien ja rakenteellisten yhtäläisyyksien ymmärtämisen. Tämä prosessi hyödyttää laajasti biologian ja lääketieteen sovelluksia, kuten lääkekehitystä, evoluutiotutkimuksia ja uusien biologisten löydösten tekemistä. Rakenteellisen vertailun merkitys korostuu entisestään, kun otetaan huomioon proteiinirakennetietokantojen, kuten AlphaFold-ennusteiden, jatkuvasti kasvava laajuus. Tämä vaatii yhä tehokkaampia ja tarkempia työkaluja, jotka pystyvät käsittelemään suuria tietomääriä nopeasti ja luotettavasti.

GTalign erottuu edukseen nopeutensa, tarkkuutensa ja skaalautuvuutensa ansiosta. Se hyödyntää k-d-puu-pohjaista tilanhakua ja iteratiivista TM-scoren optimointia, mikä tekee siitä huomattavasti nopeamman ja tarkemman kuin perinteiset menetelmät, kuten TM-align ja DALI. GTalign on erityisen hyödyllinen laajojen proteiinitietokantojen, kuten AlphaFoldin, analysoinnissa, missä sen nopeus on kriittinen tekijä. Vaikka GTalign ylittää useimmat menetelmät suurissa tietokannoissa, se kohtaa haasteita lyhyiden proteiinien kohdalla, joissa joitakin tarkkuustekijöitä on vielä parannettava.

DALIn vahvuus on sen kyky tunnistaa homologioita tilastollisesti merkittävällä tavalla. Sen etäisyysmatriiseihin perustuva analyysi ja Z-pisteytys tekevät siitä erityisen tehokkaan evolutiivisten suhteiden tutkimisessa, vaikka sen hitaus suurissa

tietokannoissa rajoittaa sen soveltuvuutta nykyajan massiivisiin tietomääriin.

TM-align tarjoaa erinomaisen tasapainon nopeuden ja tarkkuuden välillä. Sen TM-score-pohjainen lähestymistapa on vakiinnuttanut asemansa rakenteellisen vertailun perusmenetelmänä. Kuitenkin GTalign on suurelta osin syrjäyttänyt TM-alignin erityisesti suurten tietokantojen käsittelyssä.

Foldseek erottuu nopeudellaan ja soveltuu hyvin alustavaan seulontaan suurista tietokannoista. Sen sekvenssipohjainen lähestymistapa tekee siitä vähemmän tarkkaa rakenteellisesti monimutkaisissa proteiineissa, mutta sen laskennallinen tehokkuus tarjoaa merkittäviä hyötyjä suurissa analyyseissä.

Rakenteellinen vertailu osoittautui erityisen tehokkaaksi proteiinien homologioiden ja analogioiden tunnistamisessa. GTalign ja DALI kykenivät paljastamaan sellaisia rakenteellisia samankaltaisuuksia, joita pelkät sekvenssipohjaiset menetelmät eivät pystyneet havaitsemaan. Tämä korostaa rakenteellisen analyysin kriittistä roolia proteiinien toiminnan ja evoluution ymmärtämisessä.

Tulokset osoittavat, että GTalign on tällä hetkellä erinomainen valinta suurten tietokantojen nopeaan ja tarkkaan analyysiin. DALI ja TM-align säilyttävät vahvuutensa erityisesti biologisesti syvällisissä ja evolutiivisesti merkittävissä kohdistuksissa, kun taas Foldseek täydentää kokonaisuutta alustavissa seulonnoissa. Proteiinirakenteiden vertailu on ja tulee olemaan keskeinen työkalu biologian, lääketieteen ja bioteknologian innovaatioissa. [12]

Lähdeluettelo

[1] S. K. Burley, H. M. Berman ja muut, ”Protein Data Bank: the single global archive for 3D macromolecular structure data”, *Nucleic Acids Research*, vol. 47, nro D1, 2018. DOI: 10.1093/nar/gky949.

[2] H. Hasegawa ja L. Holm, ”Advances and pitfalls of protein structural alignment”, *Current Opinion in Structural Biology*, vol. 19, nro 3, s. 341–348, 2009. DOI: 10.1016/j.sbi.2009.04.003.

[3] D. Whitford, *Proteins: structure and function*. John Wiley & Sons, 2013.

[4] J.-F. Gibrat, T. Madej ja S. H. Bryant, ”Surprising similarities in structure comparison”, *Current Opinion in Structural Biology*, vol. 6, nro 3, s. 377–385, 1996. DOI: 10.1016/S0959-440X(96)80058-3.

[5] M. Van Kempen, S. S. Kim ja muut, ”Fast and accurate protein structure search with Foldseek”, *Nature biotechnology*, vol. 42, nro 2, s. 243–246, 2024. DOI: 10.1038/s41587-023-01773-0.

[6] L. Holm, A. Laiho, P. Törönen ja M. Salgado, ”DALI shines a light on remote homologs: One hundred discoveries”, *Protein science*, vol. 32, nro 1, e4519, 2023. DOI: 10.1002/pro.4519.

[7] L. Holm, ”DALI and the persistence of protein shape”, *Protein science*, vol. 29, nro 1, s. 128–140, 2020. DOI: 10.1002/pro.3749.

[8] Y. Zhang ja J. Skolnick, ”TM-align: a protein structure alignment algorithm based on the TM-score”, *Nucleic acids research*, vol. 33, nro 7, s. 2302–2309, 2005. DOI: 10.1093/nar/gki524.

[9] T. Madej, A. Marchler-Bauer, C. Lanczycki, D. Zhang ja S. H. Bryant, ”Biological Assembly Comparison with VAST+”, *Structural Bioinformatics: Methods and Protocols*, s. 175–186, 2020. DOI: 10.1007/978-1-0716-0270-6_13.

[10] M. Margelevičius, ”GTalign: spatial index-driven protein structure alignment, superposition, and search”, *Nature Communications*, vol. 15, nro 1, s. 7305, 2024. DOI: 10.1038/s41467-024-51669-z.

[11] A. R. Panchenko ja T. Madej, ”Analysis of protein homology by assessing the (dis) similarity in protein loop regions”, *PROTEINS: Structure, Function, and Bioinformatics*, vol. 57, nro 3, s. 539–547, 2004. DOI: 10.1002/prot.20237.

[12] M. L. Sierk ja W. R. Pearson, ”Sensitivity and selectivity in protein structure comparison”, *Protein Science*, vol. 13, nro 3, s. 773–785, 2004. DOI: 10.1110/ps.03328504.