Proteiinien rakenteellisten motiivien selvittäminen

Pia Laine
Pia.Laine@iki.fi

Tiedon louhinta biomolekyyliaineistosta
Helsingin yliopisto, tietojenkäsittelytieteen laitos
Raportti C-2003-52, s.51-61, marraskuu 2003

Tiivistelmä

Proteiinien toiminnan määrää niiden kolmiulotteinen rakenne eli konformaatio. PDB:n (Protein Data Bank) tietokannassa on kokeellisesti röntgenkristallografian ja NMR-menetelmän avulla selvitettyjä proteiinien kolmiulotteisia rakennemalleja yli 21 000 ja SwissProt-tietokannassa on yhteensä 136 000 proteiinisekvenssiä. Rakennemallien selvittäminen on hidasta käsityötä automaattisen proteiinin sekvenssitiedon tuottamiseen verrattuna ja tämän vuoksi on kehitetty menetelmiä, joissa käytetään hyväksi rakennemallien tietoja muiden proteiinien rakenteita ja toimintaa ennustettaessa.

Tässä raportissa kuvataan menetelmä, jolla voidaan selvittää proteiinien rakenteellisia samanlaisuuksia eli motiiveja, jotka ovat jollekin piirteelle ominaisia rakennekomponentteja ja jotka yleensä ovat myös proteiinin toiminnan kannalta keskeisiä kohtia. Menetelmässä käytetään kahta eri algoritmia. Ensin seitsemän aminohapon mittaisten proteiinisegmenttien koordinaattitietojen avulla selvitetyt konformaatiot eli rakenteet kuvataan merkkeinä, jolloin koko proteiini voidaan esittää merkkijonona (Matsuo-Kanehisa-algoritmi). Tämän jälkeen useita proteiinien rakenteista saatuja merkkijonoja rinnastamalla (Needleman-Wunsch-algoritmi) voidaan havaita proteiinien rakenteellisia motiiveja, jotka voidaan havannoida rinnastuksista konservoituneina eli samanlaisina alueina.

1. Johdanto

DNA:n sekvensoinnin ja kloonauksen kehittymisen myötä proteiinisekvenssitiedon määrä on kasvanut eksponentiaalisesti ja lukuisten genomiprojektien myötä sekvenssidatan määrä vain kasvaa. Tunnetuimpana on ihmisen genomiprojekti, jossa selvitettiin ihmisen perimä. Sekvenssimateriaalin tehokas hyödyntäminen edellyttää tärkeätä biologista informaatiota tuottavien menetelmien kehittämistä sekvenssitietoa analysoimalla. Perinteinen menetelmä proteiinien rakenteiden ennustamiseen perustuu aminohappojen ja sekundäärirakenteiden riippuvuussuhteisiin. Uudempi tietämykseen pohjautuva menetelmä perustuu empiiristen sääntöjen joukkoon, missä on sekvenssi ja rakenteellisten hahmojen suhteita. Esimerkki tällaisesta säännöstä on konservoidun aminohappohahmon GxGxxG, missä x kuvaa mitä tahansa

aminohappoa, ja nukleotidia sitovan rakenteellisen motiivin, $\beta\alpha\beta$-yksikön, välinen suhde. [MK93]

Tietämysperusteisen menetelmän onnistuminen riippuu siitä, kuinka monta sääntöä voidaan kerätä. Tällöin tulisi havaita ja karakterisoida niin monta proteiinien rakenteellista motiivia tietokantoja tutkimalla kuin on mahdollista. Rakenteiden vertailuun on kehitetty useita laskennallisia menetelmiä, mutta niiden aikavaativuus on liian suuri tietokantojen perusteelliseen tutkimiseen, sillä rakennetiedot on talletettuina tietokantoihin atomikoordinaatteina. Suurin osa näistä menetelmistä perustuu samanlaisten rakenteiden kokonaisrinnastukseen, jolloin paikalliset samanlaisuudet voivat jäädä huomioimatta. [WSD99,MK93]

Matsuon ja Kanehisan kehittämän algoritmin avulla proteiinin kolmiulotteisen rakennemallin alfa-hiilien koordinaattitiedot muutetaan symbolimerkkijonoksi. Algoritmin kehityksessä tekijät olivat kiinnostuneita sääntöpohjaisesta menetelmästä rakenteellisten motiivien, kuten $\beta\alpha\beta$-yksikön, EF-käden ja β-hiuspinnirakenteiden havaitsemisesta supersekundääritasolla. Proteiinikoordinaattitietojen symbolimerkkijonoiksi muuttamisen jälkeen voidaan saatua merkkijonoa vertailla muiden proteiinien symbolimerkkijonoihin sekvenssien rinnastusmenetelmillä. Rinnastuksen tarkoitus on löytää sekvensseistä samanlaisia domeeneja[1] ja motiiveja, ja päätellä niiden biologisia funktioita tai evolutiivista sukulaisuutta. Yleisimmin käytetyt rinnastusmenetelmät perustuvat niin sanottuihin dynaamisiin algoritmeihin, jotka voidaan jakaa kahteen kategoriaan: kokonais[2]- ja paikallisrinnastukseen[3]. Tavallisimmat sekvenssivertailumenetelmät perustuvat Needleman-Wunsch–algoritmiin (kokonaisrinnastus) ja Smith-Waterman-algoritmiin (paikallinen rinnastus). [APS99, NW70, SW81] Näistä ensimmäisen algoritmin kehittyneempää versiota käytetään merkkijonojen rinnastamiseen, jolloin saadaan sekä sekvenssi että rakenteelliset motiivit eroteltua automaattisesti. [WSD99, MK93]

2. Proteiinit

Kappaleessa 2.1 kuvataan proteiineja ja niiden rakennetta. Kolmiulotteista proteiinin rakennemallia ja motiiveja käsitellään kappaleessa 2.2 ja tunnetuimpia proteiinitietokantoja kappaleessa 2.3.

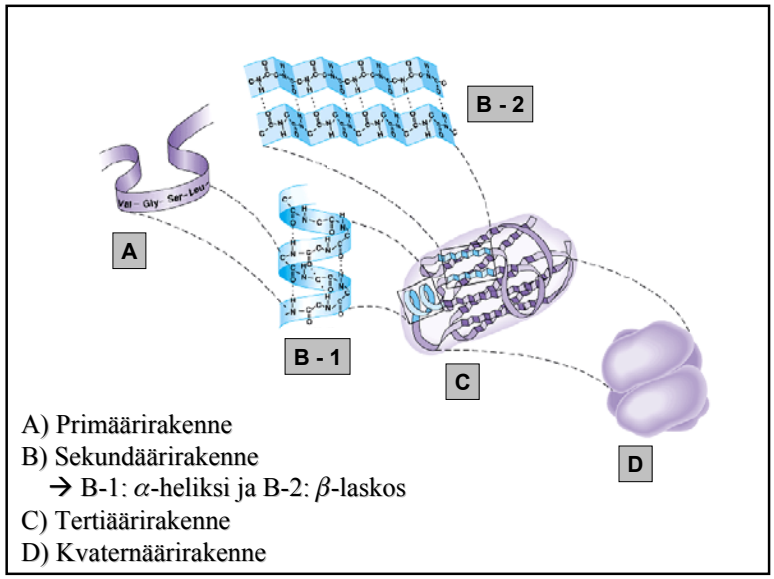
2.1. Rakenne

Proteiinit ovat monimutkaisia makromolekyylejä, jotka toimivat esimerkiksi kudosten rakennusaineina, osana metaboliaverkostoa, entsyymeinä, vasta-aineina, hormoneina ja kuljettajamolekyyleinä. Rakenteellisesti ja toiminnallisesti mitä erilaisimmat proteiinit ovat osa lähes kaikkia biologisia prosesseja. [ZPV94]

Jokainen proteiini voidaan kuvata aminohappoina, joiden liittyessä yhteen muodostuu polypeptidiketju. Aminohappoja koodaavassa DNA:ssa on koodi jokaiselle 22:lle eri aminohapolle, joista ihmiselle välttämättömiä aminohappoja on noin kymmenen. Proteiinit voidaan jaotella rakenteen mukaan primääri-, sekundääri-, tertiääri- ja kvaternäärirakenteisiin (Kuva 1). Primäärirakennetta (esim. HKIWARPRPPARLA) voidaan kuvata siten, että proteiini muodostuu jonosta toisissaan kiinniolevia aminohappoja. Sekundäärirakenteella tarkoitetaan

[1] Domeeni on proteiinirakenteen itsenäisesti poimuttuva yksikkö.
[2] Kokonaisrinnastuksessa rinnastetaan sekvenssejä niin, että sekvenssien samankaltaisuuksia pyritään löytämään koko vertailtavien sekvenssien pituudelta.
[3] Paikallisella rinnastuksella vertailtavista sekvensseistä etsitään huomattavan samankaltaisia alueita eli paikallisia samanlaisuuksia.

aminohappoketjun järjestäytymistä eli laskostumista erilaisiksi rakenteiksi, joista tärkeimmät yksiköt ovat alfa-heliksi ja beta-laskos. Proteiinin laskostumiseen vaikuttaa aminohapposekvenssin koostumus. Tertiäärirakenteella tarkoitetaan alfa-heliksien ja beta-tasojen muodostaman yhden polypeptidiketjun kokonaismuotoa, kun taas kvaternäärirakenteen muodostaa usea polypeptidiketju. Jokaisella proteiinilla on monimutkainen kolmiulotteinen rakenne, konformaatio, (Kuva 1) johon vaikuttaa sen kemiallinen ja fysikaalinen ympäristö, kuten pH-arvo ja lämpötila. [ZPV94, VJ03]



Kuva 1 Proteiinin konformaatio.

2.2. Kolmiulotteinen (3D) rakennemalli ja motiivi

Proteiinien toiminnan määrää niiden konformaatio. Kaikkia proteiinien rakenteeseen vaikuttavia tekijöitä ei tunneta, joten pelkän aminohapposekvenssin avulla ei pystytä selvittämään kolmiulotteista proteiinirakennetta. Rakennemallien tunteminen helpottaa ymmärtämään proteiinien toimintaa ja tunnettujen proteiinirakennemallien avulla voidaan ennustaa muiden proteiinien toimintoja. Ennustamisessa täytyy ottaa kuitenkin huomioon, että kaksi samankaltaisesti laskostuvaa (kolmiulotteisen rakenteen syntyminen) proteiinia voi sekvensseiltään olla täysin erilaisia tai kahdella sekvenssiltään samankaltaisella proteiinilla voi olla eri funktiot. [ZPV94]

Motiivit ovat lyhyehköjä jollekin piirteelle ominaisia sekvenssijaksoja tai rakennekomponenteja, joita esiintyy samansukuisissa proteiineissa. Motiivit ovat proteiinin toiminnalle keskeisiä alueita, esimerkiksi sitoutumiskohtia, ja tämän vuoksi ne ovat säilyneet evoluution kuluessa [HKM+02]. Rakenteellisten motiivien tai paikallisten rakenteiden ymmärtäminen ja tulkitseminen on tärkeää proteiinien toiminnan kannalta, sillä on osoitettu, että rakenteeltaan erilaisilla proteiineilla on samanlaisia paikallisia laskostumishahmoja, ja tällöin kyseessä olevilla proteiineilla on konservoituja toimintoja. Proteiinien tertiäärirakennetasolla tunnetuimpia rakenteellisia motiiveja ovat $\beta\alpha\beta$-yksikkö, EF-käsi ja heliksi-käännös-heliksi. Esimerkkinä voidaan mainita dehydrogenaasit, joilla on rakenteellinen motiivi nimeltään

Rossman-laskos (fold), joka muodostuu kahdesta peräkkäisestä $\beta\alpha\beta$-yksiköstä. Useimmilta dehydrogenaaseilta on löydetty paikallinen sekvenssimotiivi GxGxxG nukleotidia sitovan alueen läheisyydessä. [ASP99, XSD99]

2.3. Proteiinitietokannat PROSITE ja PDB

Prosite on biologisesti merkittävien proteiiniperheiden ja motiivien tietokanta, joka kokoaa yhteen sekvenssihahmoja ja niihin liittyviä toimintoja. Tietokanta sisältää paljon erilaisia proteiineja, joista suurin osa on ryhmitelty sekvenssihomologioiden perusteella perheisiin. Proteiinit tai proteiinimotiivit kuuluvat tiettyyn perheeseen, joilla on toiminnallisia samanlaisuuksia ja näin ollen ne periytyvät yhteisestä kantaisästä. Tutkittaessa proteiinien sekvenssiperheitä jotkut alueet säilyvät paremmin kuin toiset evoluution edetessä. Tällaiset alueet ovat yleensä tärkeitä proteiineille ja kolmiulotteisen rakenteen ylläpitämiselle. Tietokantaa käyttämällä on mahdollista ennustaa aminohapposekvenssien perusteella proteiinin paikalliset rakenteet ja toiminta. [TU03, P03, WSD99]

Rakennebiologian keskeisin tietokanta Protein Data Bank (PDB) sisältää biologisten makromolekyylien kolmiulotteista rakennetietoa. PDB:ssa olevan tiedon määrä kasvaa nopealla vauhdilla ja tällä hetkellä tietokannassa on noin 22810 rakennetta, joista valtaosa on proteiineja, mutta lisäksi tietokannasta löytyy myös nukleiinihappojen (DNA, RNA) ja hiilihydraattien rakenteita. Rakenteet on pääosin selvitetty kokeellisesti röntgenkristallografian ja NMR:n (ydinmagneettiresonanssispektroskopia) avulla. Kolmiulotteiset rakennemallit talletetaan koordinaattitiedostoina, joissa proteiinimolekyylin jokainen tunnistettu atomi on määritetty koordinaatteineen. Koordinaattitiedoston standardimuoto on PDB-formaatti. Keskeisin asema proteiinien kolmiulotteisilla rakennemalleilla on lääketeollisuudessa lääkkeiden suunnittelussa, bioteknologiassa ja proteiinien suunnittelussa. [PDB03, KS00, WSD99]

3. Rakenteellisten motiivien selvittäminen

Proteiinien rakenteellisten motiivien selvittäminen perustuu kahteen menetelmään, joista ensimmäistä Matsuo-Kanehisa-algoritmia kuvataan yksityiskohtaisemmin kappaleessa 3.1 ja rinnastusalgoritmeista kappaleessa 3.2. Kappaleessa 3.3 on esitetty esimerkki rakenteellisten motiivien etsimisestä.

3.1. Proteiinien kolmiulotteisten rakennemallien koordinaattien kuvaaminen merkkijonoina

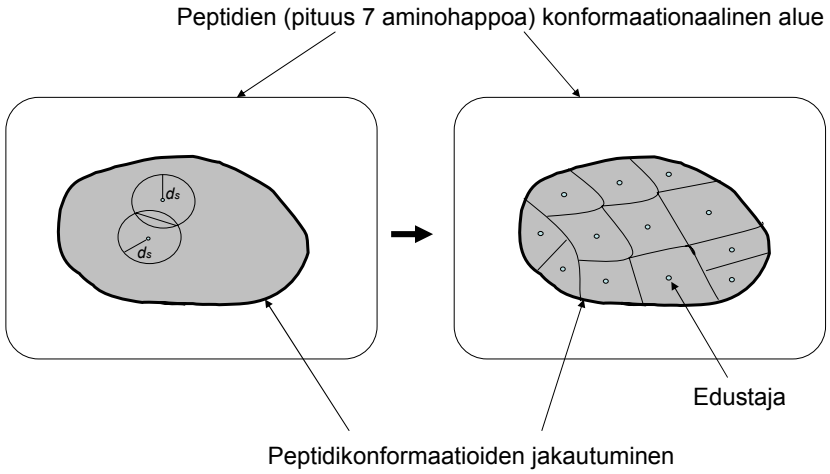
Matsuo-Kanehisa-algoritmi muuttaa proteiinin rakenteesta saatujen alfa-hiilien koordinaatit symbolimerkkijonoksi. Tämä menetelmä vastaa kvantisointiproseduuria, jota käytetään esimerkiksi digitaalisen signaalin prosessoinnissa ja kuvankäsittelyssä. Kvantisointiproseduuri muuttaa jatkuvan signaalin, kuten äänen ja kuvan digitaaliseksi signaaliksi, jotta signaalia voidaan käsitellä tietokoneella. Yhtälaisesti tämä kehitetty algoritmi tiivistää pienten peptidisegmenttien konformaatiot diskreeteiksi hahmoiksi. Jokainen hahmo esitetään symbolina ja täten koko polypeptidiketju esitetään symbolimerkkijonona. Toisin sanoen jokaiselle tietyn mittaiselle peptidisegmentille annetaan oma symboli sen konformaation

perusteella. Tällöin on mahdollista tunnistaa konformaationaaliset samanlaisuudet, joita ei voida erotella tavallisella sekundäärisiä rakenteita erottelevalla menetelmällä.

Algoritmin kehitykseen käytetty koordinaattidata 93 polypeptidiketjusta on peräisin PDB-proteiinitietokannasta (heinäkuu, 1989). Polypeptidiketjujen atomitarkkuus oli selvitetty alle 3.0 Å[1] tarkkuudella ja lisäksi polypeptidiketjulla ei ollut merkittävää sekvenssien homologiaa toisensa kanssa. [WSD99, MK93] Matsuo-Kanehisa-algoritmi valitsee proteiinisegmenttitiedostosta eri konformaation omaavia edustajia. Algoritmissa on määritelty kaksi parametria l ja d_s, missä l on segmentin pituus ja d_s näytteenottoväli. Peptidisegmentin pituus (l) valittiin kuvaamaan mitä tahansa proteiinirakenteen 'rakennuspalikkaa'. Tutkimuksien perusteella sekundäärirakenteen yksikön pienin pituus on neljä aminohappoa (esimerkki: yhden alfa-heliksin kierto on neljä aminohappoa). [WSD99] Koska algoritmin kehittämisessä oltiin kiinnostuneita rakenteellisten hahmojen osoittamisesta, valittiin peptidisegmentin pituudeksi seitsemän aminohappoa. Tällöin 93 polypeptidiä sisälsi yhteensä 15320 peptidisegmenttiä. [MK93]

Näytteenottoväli d_s kuvaa etäisyyttä kvantisoitaessa peptidisegmenttien konformaatioita (Kuva 2). Parametriarvo valittiin manuaalisesti, jotta saataisiin helposti käsiteltävä määrä eri konformaation omaavia edustajia ja saavutettaisiin sopiva homogeenisuuden taso, jolla kaikki saman kategorian edustajat olisivat samankaltaisia toisiinsa verrattuna. Näytteenottoväliksi määriteltiin 2.01 Å, jolloin tulokseksi saatiin helposti käsiteltävä määrä edustajia (Nr =37). Kahden segmentin väliseksi etäisyyksi määriteltiin minimoitu neliöllinen keskiarvoetäisyys (RMS etäisyys) kahden alfa-hiilen[2] välille. RMS-etäisyydet edustajajoukon sisällä olivat 1.28 ± 0.3 Å, jota verrataan näytteenottoväliin d_s (=2.01 Å). Toisin sanoen yhden edustajajoukon eli yhtä konformaatiota edustavien peptidisegmenttien välinen etäisyys on ≤ 2.01 Å. Tutkitut proteiinit (93 kpl) sisälsivät yhteensä 15320 seitsemän aminohapon mittaista peptidisegmenttiä, joista valittiin 37 edustajajoukkoa (Taulukko 1). Jokaiselle edustajajoukolle määriteltiin oma symboli (1-9, A-Z, @, #), joista yleisimpiä olivat heliksit ja β-säikeet eli symbolit N (HHHHHHH) ja # (EEEEEEE) [H=α-heliksi, E=β-säie]. Taulukon 1 toisessa sarakkeessa on eri segmenttien (15320 kpl) konformaatioiden jäsenmäärät ja osuudet, ja kolmas sarake ilmaisee sekundäärirakenteen. [WSD99, MK93]

Kuva 2 Kuvaus peptidisegmentti-konformaatioiden kvantisoinnista.



[1] Ångström (Å) mitta vastaa 10^{-10} metriä

[2] Alfa-hiili = Aminohappojen atomien nimeämisessä käytetän yleensä seuraavaa kaavaa: Pääketjun muodostavat aminoryhmä (NĐH),α-hiili (Cα) ja siihen liittynyt vety (Hα), sekä karbonyyliryhmä (C=O).

Matsuo-Kanehisa-algoritmi on esitetty pseudokoodina kuvassa 3. Algoritmissa valitaan proteiinisegmenttitiedostosta eri konformaation omaavia edustajia. Algoritmin syötteenä annetaan näytteenottoväli *(d_s)*, proteiinisegmenttien joukko *(S)* ja edustajienjoukko *(R)*. Joukosta *S* valitaan mielivaltaisesti yksi segmentti *s*. Segmentti *s* lisätään edustajienjoukkoon *R* ja samalla *s* poistetaan joukosta *S*. Tämän jälkeen käydään joukon *S* segmenttejä yksitellen läpi niin kauan kunnes joukko *S* on tyhjä eli kaikki segmentit on lisätty joukkoon R. Joukosta *S* otetaan mielivaltaisesti segmentti *u* ja jokaiselle joukon R edustajalle *r* lasketaan etäisyys segmentin *u* ja edustajan *r* välillä. Jos etäisyys segmentin *u* ja edustajan *r* välillä on pienempi kuin näytteenottoväli d_s, niin tällöin segmentti *u* kuuluu edustajan *r* joukkoon. Jos segmentin *u* ja edustajan *r* välinen etäisyys on suurempi kuin näytteenottoväli d_s, muodostetaan uusi edustaja joukkoon *R*..

Algoritmi ei ole deterministinen, mikä tarkoittaa sitä, että edustajienjoukko riippuu segmenttien valintajärjestyksestä. Edustajienjoukon tulee aina toteuttaa seuraavat ominaisuudet; a) kahden edustajan etäisyys täytyy olla suurempi kuin näytteenottoväli ja b) jokaisen segmentin minimietäisyys toiseen edustajaan samassa joukossa täytyy aina olla pienempi tai yhtäsuuri kuin näytteenottoväli. Sen jälkeen kun edustajienjoukot on määritelty voidaan jokainen tietojoukosssa oleva segmentti sijoittaa omaan luokkaan. [WSD99]

```
EDUSTAJIEN_VALINTA(ds, S, R)
/* Aluksi joukko S sisältää kaikkien proteiinien kaikki segmentit
/' ja edustajienjoukko R on tyhjä
  Valitse mielivaltaisesti segmentti s joukosta S;
  Lisää segmentti s edustajienjoukkoon R, ja poista se joukosta S;
/* käydään läpi kaikkien joukkoa S, jossa kaikki segmentit,
/* niin kauan kunnes se on tyhjä
while S ei ole tyhjä do
   begin
      Valitse mielivaltaisesti segmentti u joukosta S;
      dist(u) ← ∞;
      for jokaiselle edustajalle r joukossa R do
            begin
                  Laske etäisyys u ja r välillä, d(u,r);
                  if d(u,r) < dist(u) then dist(u) ← d(u,r);
            end;
      if dist(u) > ds then lisää u joukkoon R;
      Poista u joukosta S;
   End;
```

Kuva 3 Matsuo-Kanehisa-algoritmi pseudokoodina.

Edustaja-joukko symboli	Jäsenien määrä (%)	Sekundääri-rakenne
N	2351 (16.52)	HHHHHHH
#	1979 (12.92)	EEEEEEE
C	1005 (6.56)	HHHHHHH
X	868 (5.67)	. S . EEEE
Z	775 (5.06)	EEEE . ST
V	747 (4.88)	T . . . EEE
@	740 (4.83)	EEE . S . .
5	511 (3.34)	. . HHHHH
W	468 (2.99)	ET . . . . .
H	415 (2.71)	EEE . TT .
A	389 (2.54)	. TT . . . E
D	385 (2.51)	E . . HHHH
E	375 (2.45)	. . . TT . .
U	341 (2.23)	. . SS . . .
2	339 (2.21)	H . TT . . .
1	302 (1.97)	HHHHHT .
B	271 (1.77)	. TT . . EE
G	269 (1.76)	E . . SSS .
T	229 (1.50)	ETS . . EE

Edustaja-joukko symboli	Jäsenien Määrä (%)	Sekundääri-rakenne
Y	197 (1.29)	. . S . . S .
O	202 (1.32)	HTTS . . E
6	194 (1.27)	HHHTT. .
L	184 (1.20)	HHTT . EE
R	182 (1.19)	TTS . TT .
M	180 (1.18)	HHHTTT .
9	159 (1.04)	. SSST . .
Q	157 (1.03)	T . . . S . .
F	151 (0.99)	EE . TTEE
8	133 (0.87)	. . TT . . E
7	126 (0.82)	E . TT . . E
S	115 (0.75)	. T . . TTH
K	107 (0.70)	E . TTS . E
4	100 (0.65)	. . SHHHH
3	90 (0.59)	E . TTT . .
I	87 (0.57)	EE . TTS .
P	14 (0.09)	SSSSSS .
J	13 (0.09)	HHTTTS .

Taulukko 1. Edustajienjoukko.
37 samankaltaisen peptidikonformaation omaavaa edustajaa (Nr). E=β-säie, H=α-heliksi, T=vetysidonnainen käännös, S='kaari'.

3.2. Rakenteellisten motiivien selvittäminen rakenne- ja sekvenssivertailulla

Edellä kuvattu Matsuo-Kanehisa-algoritmi muuttaa proteiinien kolmiulotteisten rakenteiden koordinaattitiedot symbolimerkkijonoiksi, joita voidaan verrata eli rinnastaa keskenään rakenteellisten motiivien identifioinnissa. Merkkijonojen rinnastuksen tarkoituksena on heijastaa rakenteellisesti samoja tai samankaltaisia alueita, jolloin merkitykselliset rakenteelliset motiivit voidaan havaita toiminnaltaan sukulaissekvenssien monirinnastetuksesta. Rinnastukseen on kehitetty erilaisia rinnastusalgoritmeja. [WSD99] Rinnastuksessa on tärkeää käsitellä oikein identtiset parit, ei-identtiset parit, aukot ja lisäksi käytettävä samanlaisuuspisteytysmatriisi (= mutaatiomatriisi tai substituutiomatriisi) aminohapposekvenssien vertailuihin. Tämä yksinkertaisimmillaan tarkoittaa sitä, että jokaisesta identtisestä parista annetaan piste ja jokaisesta ei-identtisestä parista nolla pistettä. Nykyisin käytetyimpiä ja tärkeimpiä ovat Dayhoffin (1978) PAM-matriisit ja Henikoff & Henikoff (1992) kehittämät BLOSUM-matriisit. Näistä kahdesta PAM-matriisit perustuvat vahvasti evoluutioteoriaan ja BLOSUM-matriisit puolestaan konservoituneiden proteiinisekvenssien samankaltaisuuksiin. [WSD99, TU03].

Tunnetuimpia sekvenssien rinnastamisalgoritmeja ovat dynaamiseen ohjelmointiin perustuvat Smith-Waterman ja Needleman-Wunsch-algoritmit [APS99]. Smith-Waterman-

algoritmi etsii kahden sekvenssin välisen parhaan osittais- eli paikallisrinnastuksen. Sen sijaan Needleman-Wunsch-algoritmi etsii kahden sekvenssin välisen parhaan mahdollisen kokonaisrinnastuksen ja tällöin sekvenssit rinnastetaan koko pituudeltaan optimaalisesti. [NW70, SW81]

Matsuo ja Kanehisa käyttivät modifoitua Goad-Kanehisa-algorimia, mikä on yleistys Needleman-Wunsch-algoritmista, merkityksellisten paikallisten samanlaisuuksien etsimiseen laskostumishahmoista.[MK93] Rakenteellisten motiivien etsimiseen käytetty Goad-Kanehisa-algoritmi toimii kuten Needleman-Wunsch-algoritmi, mutta se löytää pitkiä sekvenssejä rinnastamalla myös kaikki alisekvenssit, jotka muistuttaa toisiaan paikallisesti [GK82].

Matsuo-Kanehisa-algoritmin kehityksessä käytetyn 15320 segmenttiä sisältävän tiedoston perusteellisen vertailun seuraksena saavutettiin 858 merkittävää samankaltaisuusparia, jotka saavuttivat halutut kriteerit. Rakennevertailun tuloksena löydettiin muun muassa nukleotidin sitovat motiivit (Rossman-laskos) ja kalsiumia sitovat motiivit (EF-käsirakenne), joista molemmat osoittivat hyvää rakenteellisen hahmon, sekvenssihahmon ja toiminnallisen merkityksen korrelaatiota. [WSD99].

3.3. Esimerkki proteiinien rakenteellisten motiivien etsimisestä

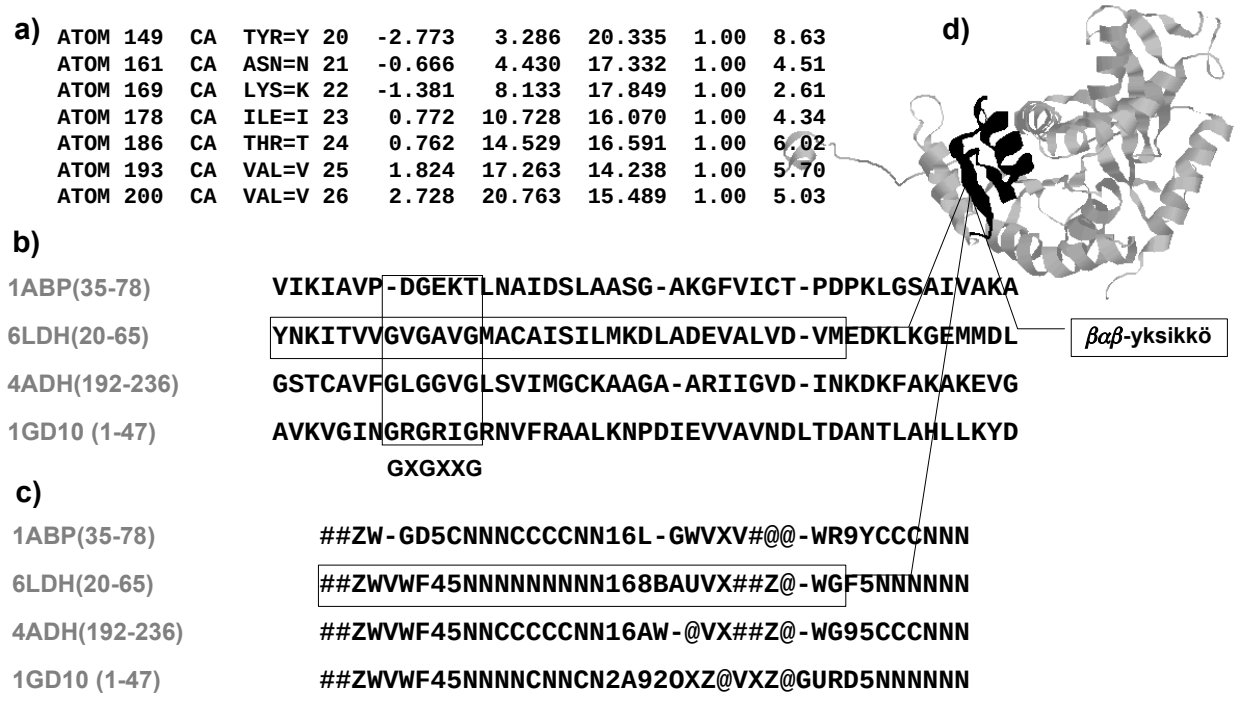
Esimerkkiin on otettu mukaan neljän eri proteiinin osasekvenssiä (Taulukko 2).

PDB id	Proteiinin nimi	Sekvenssialue
1ABP	L-Arabinose-Binding Protein Complex With L-Arabinose	35-48
6LDH	M_4 Apo-Lactate Dehydrogenase	20-65
4ADH	Apo-Liver Alcohol Dehydrogenase	192-236
1GD1O	holo-D-Glyceraldehyde-3-Phosphate Dehydrogenase-Chain O	1-47

Taulukko 2: Esimerkissä olevien proteiinien PDB (Protein Data Bank) numero (id), nimi ja sekvenssialue.

Taulukossa 2 esitetyistä proteiineista ensimmäinen on arabinoosia sitovat proteiini (1ABP), ja kolme viimeisintä dehydrogenaaseja, joiden sekvenssi sisältää $\beta\alpha\beta$-yksikön eli nukleotidia sitovan domeenin.

Kuvan 4 kohdassa a) on proteiinin 6LDH aminohappojen 20-26 alfa-hiilien atomikoordinaatit. Neljän mainitun proteiinien aminohapposasekvenssit on rinnastettu kohdassa b) yleistettyllä Needleman-Wunsch-algoritmilla. Monirinnastuksessa voidaan erottaa konservoitunut alue GXGXXG, joka on tyypillinen dehydrogenaaseille mutta ei arabinoosia sitoville proteiineille (1ABP). Matsuo-Kanehisa-algoritmilla muutetaan proteiinien kolmiulotteisen rakenteen koordinaattitiedot symbolimerkkijonoksi, joita rinnastetaan Needleman-Wunsch-algorimilla (kohta c). Symbolimerkkijonorinnastuksesta huomataan konservoituneita alueita kaikkien proteiinien välillä enimmäkseen merkkijonorinnastuksen alkupäässä. Dehydrogenaasien välillä samanlaisuutta on enemmän 6LDH ja 4ADH kesken, toisaalta 1GD1O näyttää kovin erilaiselta rinnastuksen keskivaiheilla, vaikka näillä kaikilla kolmella hydrogenaasilla on samanlainen rakenne eli $\beta\alpha\beta$-yksikkö, joka on kuvattu mustalla 6LDH proteiinin kolmiulotteiseen rakenteeseen (kohta d).



Kuva 4: Dehydrogenaasit (sis. $\beta\alpha\beta$-yksiköt).
a) Osa 6LDH proteiinin koordinaattitiedostoa, missä on esitetttynä vain alfa-hiilien (CA) koordinaatteja b) Dehydrogenaasien aminohapporinnastus yleistetyllä Needleman-Wunsch-algoritmilla c) Matsuo-Kanehisa-algoritmilla aikaansaatujen seitsemän aminohapon mittaisten polypeptidisegmenttien symboliesitysten rinnastus, d) 6LDH proteiinin kolmiulotteinen molekyylimalli ($\beta\alpha\beta$-yksikkö mustalla) [PDB03].

4. Yhteenveto

Lyhyiden peptidisegmenttien kolmiulotteisten rakennemallien koordinaattien muuttaminen symboleiksi on käsitteellisesti jatkumo sekundäärirakenteiden analyysistä. Menetelmän avulla heijastuu ainoastaan proteiinisegmenttien paikalliset rakenteet ja tämän vuoksi se ei ole saavuttanut laajaa hyväksyntää proteiinien laskostumisen analyysissä. Proteiinirakenteiden symbolimerkkijonoesitys soveltuu paikallisten rakenteellisten motiivien selvittämiseen. Yleisesti voidaan todeta, että paikalliset rakenteelliset motiivit korreloivat hyvin aminohapposekvenssien motiiveihin [WSD99].

Tässä raportissa kuvattiin proteiinien rakenteellisten motiivien selvittäminen Matsuo-Kanehisa–algoritmia ja yleistettyä Needleman-Wunsch-algoritmia käyttämällä. Tämä on kuitenkin vain yksi menetelmä proteiinien rakenteellisten motiivien selvittämisessä. Tämän menetelmän suurin hyöty lienee siinä, että proteiinien kolmiulotteisten rakenteiden

atomikoordinaatit voidaan esittää symbolimerkkijonona, jolloin on mahdollista tehdä proteiinien rakennevertailuja tyypillisiä sekvenssien rinnastusmenetelmiä käyttäen. Päinvastainen menetelmä rakenteellisten motiivien selvittämiseen olisi aloittaa tutkiminen proteiinien aminohapposekvensseistä, rinnastaa nämä ja konservoitujen aminohappojen avulla tunnistaa sekvenssimotiiveja, jotka voisivat korreloida yleisten rakenteellisten hahmojen ja konservoitujen toimintojen kanssa. [WSD99]

Auttaako proteiinien kolmiulotteisten rakennetietojen ja aminohapposekvenssien järjestelmällinen analyysi ymmärtämään proteiinien toiminnan, sekvenssien ja rakenteiden suhdetta? Monet kiinnostavimmat proteiinien toiminnalliset ja evolutionaaliset suhteet ovat muinaisia, jolloin niitä ei voida luotettavasti havaita sekvenssivertailulla. Tämänkaltaiset suhteet voidaan havaita ainoastaan proteiinien tertiäärirakenteita vertailemalla. Matsuo ja Kanehisan kehittämän menetelmän on todettu sopivan hyvin tertiäärirakenteisiin, mutta algoritmia käytetään kuitenkin harvoin. [JECT02, WSD99]

Viitteet

[APS99] Attwood T.K. and D.J Parry-Smith. *Introduction to Bioinformatics*. Addison Wesley Longman Limited. 1999.

[GK82] Goad W.B., M.I. Kanehisa. Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucleic Acids Res*. 11;10(1):247-63, 1982.

[HKM+02] Heikkinen Erja, Marjo Korpi, Kimmo Mattila, Ilkka Porali, Mauno Vihinen. Lyhyt bioinformatiikan sanasto, 2002.

[HMC99] Hudak J. and M.A. McClure. A Comparative Analysis of Computational Motif-Detection Methods. *Pacific Symposium on Biocomputing* 4:138-149 1999.

[JECT02] Jonassen Inge, Ingvar Eidhammer, Darrell Conklin, and William R. Taylor. Structure motif discovery and mining the PDB. *Bioinformatics* 18: 362-367, 2002.

[VJ03] Vuorinen Jukka. Entsyymielektroforeesin perusteet. http://www.joensuu.fi/biologia/vuorinen/ef/efkalvot03.pdf, 2003.

[NW70] Needleman S. B., C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol Biol* 48(3):443-53, 1970.

[MK93] Y Matsuo and M Kanehisa. An approach to systematic detection of protein structural motifs. *Comput. Appl. Biosci.* 9: 153-159, 1993.

[P03] Prosite. http://ca.expasy.org/cgi-bin/prosite-list.pl

[PDB03] PDB documentation and information. http://www.rcsb.org/pdb/info.html

[KS00] Kilpeläinen Seppo, Oulun yliopisto. Lyhyt johdatus molekyylimalleihin. http://www.biochem.oulu.fi/~skilpela/kurssi/kurssi.html

[SW81] Smith T. F. and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.* 25;147(1):195-7., 1981.

[TU03] Tuimala Jarno ja Pekka Uimari. Geneettinen bioinformatiikka –luentomoniste. Helsingin yliopisto, perinnöllisyystieteen laitos, 2003.

[WSD99] Wang Jason T. L., Bruce A. Shapiro, and Dennis Shasha (edition). *Pattern Discovery in Biomolecular Data: Tools, Techniques, and Applications*. Chapter 6: Kentaro Tomii and Minoru Kanehisa. Systematic Detection of Protein Structural Motifs. New York, Oxford University Press, pp. 97-110, 1999.

[ZPV94] Zubay G., W. Parson, D. Vance. *Principles of Biochemistry* - Protein Structure and Function. Wm. C. Brown Publishers Dubuque, Iowa – Melbourne, Australia – Oxford, England, pp. 47-132, 1994.