Motiivien etsiminen MEME-algoritmilla

Paula Silvonen
Paula.Silvonen@vtt.fi
Tiedon louhinta biomolekyyliaineistoista
Helsingin yliopisto, tietojenkäsittelytieteen laitos
Raportti C-2003-52, s.21-30, marraskuu 2003

Tiivistelmä

Motiivilla tarkoitetaan bioinformatiikassa jotakin esimerkiksi DNA:sta tai proteiinista löytyvälle piirteelle ominaista sekvenssijaksoa tai rakennekomponenttia. Motiiveja analysoimalla voidaan saada tietoa esimerkiksi lääkkeiden kehittämiseen ja sairauksien syiden kartoittamiseen.

Tässä raportissa kuvataan EM-algoritmin (expectation maximization) käyttöä motiivien etsimisessä kohdistamattomasta proteiinisekvenssidatasta. Esimerkkinä käytetään MEME-ohjelmistoa, jossa on toteutettu laajennettu EM-algoritmi kaukaisesti toisilleen sukua olevien homologien löytämiseksi. Sekvenssejä sanotaan homologisiksi, jos ne ovat eriytyneet yhteisestä esi-isästä. Proteiinisekvenssidatassa motiivi voi olla esimerkiksi osasekvenssi, joka yhdistyy DNA:han, toisiin proteiineihin, steroideihin tai entsyymeihin. EM-algoritmi löytää motiiveja ohjaamattomalla oppimisella, mikä tekee siitä hyvin sopivan menetelmän suurten tietomassojen analysointiin. Löydettyjä motiiveja voidaan käyttää MAST- ja Meta-MEME -työkalujen syötteenä haettaessa proteiiniperheiden uusia jäseniä sekvenssitietokannasta.

1 Johdanto

Motiivi on bioinformatiikassa jollekin esimerkiksi DNA:sta tai proteiinista löytyvälle piirteelle ominainen sekvenssijakso tai rakennekomponentti. Motiiveja analysoimalla voidaan löytää samaa alkuperää olevia proteiiniperheitä, joiden avulla saadaan tietoa esimerkiksi lääkkeiden kehittämiseen ja sairauksien syiden kartoittamiseen. Proteiineja, jotka ovat eriytyneet yhteisestä esi-isästä, kutsutaan homologeiksi.

Useiden genomikartoitusprojektien tuloksena on syntynyt valtavia sekvenssitietopankkeja (esim. SWISS-PROT, Genpept98), joiden tehokkaaseen analysointiin tarvitaan uudenlaisia menetelmiä. Biomolekyylidatasta voidaan etsiä tiedonlouhinnalla tilastollisesti merkitseviä piirteitä, joiden löytäminen muilla keinoin olisi liian työlästä.

Suurien tietomassojen käsittelyssä yleisiä tehtäviä ovat datan luokittelu ja ryhmittely. Luokittelussa ohjelman tietoon saatetaan luokat, joihin syötteet halutaan jakaa; ryhmittelyssä tätä tietoa ei ole annettu.

Ohjattu oppiminen (supervised learning) tarkoittaa algoritmin opettamista opetusjoukon avulla [Koi02]. Opetusjoukon muodostavat joukko syötteitä ja niiden luokat tai joukko

opetusesimerkkejä, jotka on merkitty positiivisiksi tai negatiivisiksi. Ohjatun oppimisen menetelmiä ovat esimerkiksi päätöspuut, naiivi Bayes –luokittelu ja k-lähinaapuriluokittelu.

Ohjaamattomassa oppimisessa (unsupervised learning) syötedata ryhmitellään pelkkien datassa havaittujen ominaisuuksien perusteella ilman etukäteistietoa todellisista luokista ja niiden ominaispiirteistä. Tähän tarkoitukseen pitää määritellä jokin dataan sopiva etäisyysmitta, jonka perusteella toisiaan lähellä olevat havainnot voidaan sijoittaa samaan ryhmään. Menetelmästä riippuen ryhmien lukumäärä voi olla annettu ennalta tai se voidaan valita suorituksen aikana. Ohjaamatonta oppimista on yleensä vaikeampi toteuttaa kuin ohjattua; sen etuna on datan esikäsittelyn vähäisyys.

SOM (Self Organizing Map), *pääkomponenttianalyysi*, *vektorikvantisointi*, *assosiaatiosäännöt* ja *EM-algoritmi* (expectation maximization, odotuksen maksimointialgoritmi) ovat esimerkkejä erilaisista ohjaamattoman oppimisen menetelmistä. Ohjaamatonta oppimista käytetään bioinformatiikan lisäksi monilla muilla aloilla, kuten kuvantunnistuksessa ja luonnollisen kielen käsittelyssä.

Tässä raportissa kuvataan EM-algoritmia motiivien etsinnässä. Esimerkkinä käytetään MEME-ohjelmistoa. Luku 2 määrittelee odotuksen maksimointialgoritmin. Luvussa 3 on yleiskuvaus MEME-ohjelmistosta ja sen käyttötarkoituksista. Luku 4 on yhteenveto.

2 EM-algoritmi

Dempsterin [DLR77] alun perin esittämä EM-algoritmi (expectation maximization) on iteratiivinen optimointimenetelmä hierarkkisille tilastollisille malleille, jotka koostuvat epätäydellisestä havaintodatasta δ, havaitsemattomista arvoista eli piilomuuttujista (hidden variable) h ja mallin parametreista α [Uus03]. Algoritmilla estimoidaan puuttuvan datan arvoja ja mallin parametreja havaintodatan perusteella. EM-algoritmia käytetään monilla aloilla tilanteissa, joissa tunnemme havaitsemamme arvot tuottaneen yleisen tilastollisen mallin, mutta havaintomme ovat epätäydellisiä ja mallin parametrit tuntemattomia. Aloja joilla EM-algoritmia ja sen muunnelmia on käytetty, ovat esimerkiksi kieliteknologia, tähtitiede ja taloustiede.

Algoritmissa on tavoitteena oppia prosessin todennäköisimmät parametrit havaintodataa tarkkailemalla. Maksimoitavan todennäköisyyden mittana käytetään yleisesti *logaritmista uskottavuusfunktiota* (log likelihood). Logaritminen uskottavuusfunktio on kätevä mitta, koska logaritmien yhteenlasku vastaa kertolaskua lineaarisessa avaruudessa, ja näin pääsemme eroon merkitsevyyden häviöstä, joka voi syntyä todennäköisyysarvojen (määritelmän mukaan <1) kertolaskusta [JuM00].

Optimointiongelman voi esittää formaalisti seuraavalla tavalla: Olkoon δ kiinnitetty. Etsi arvot α siten, että todennäköisyys

$$P(\delta \mid \alpha) = \sum_{h} P(\delta, h \mid \alpha)$$

maksimoituu.

EM-algoritmi muodostuu kahdesta iteraatioaskeleesta, ennustamisesta (expectation step, E-askel) ja mallin parametrien estimoinnista (maximization step, M-askel), joita toistamalla malli todistettavasti suppenee (lokaaliin) optimitulokseen. E-askeleessa estimoidaan piilomuuttujien *h* arvot laskemalla niiden odotusarvo annettuna havaintodata δ ja mallin sen hetkinen parametriestimaatti α^P:

$$\hat{h} = E[h \mid \delta, \alpha^P]$$

M-askeleessa käytetään piilomuuttujien *h* viimeisintä estimaattia, jotta saadaan parannettua mallin parametrien estimaattia. Eli valitaan α^{P+1} siten että yhteistodennäköisyys $P(\delta, \hat{h})$ maksimoituu:

$$\alpha^{P+1} = \{ \alpha \mid P(\delta, \hat{h} \mid \alpha) \text{ maksimoituu} \}.$$

Voidaan osoittaa, että EM-algoritmi ei huononna parametriestimaattia iteraatiosta seuraavaan siirryttäessä (monotonisuus), joten se suppenee (lokaaliin) maksimiin. Kuten monissa muissakin hill-climbing algoritmeissa, riippuu aloituskohdan valinnasta, päädytäänkö lokaaliin vai globaaliin maksimiin.

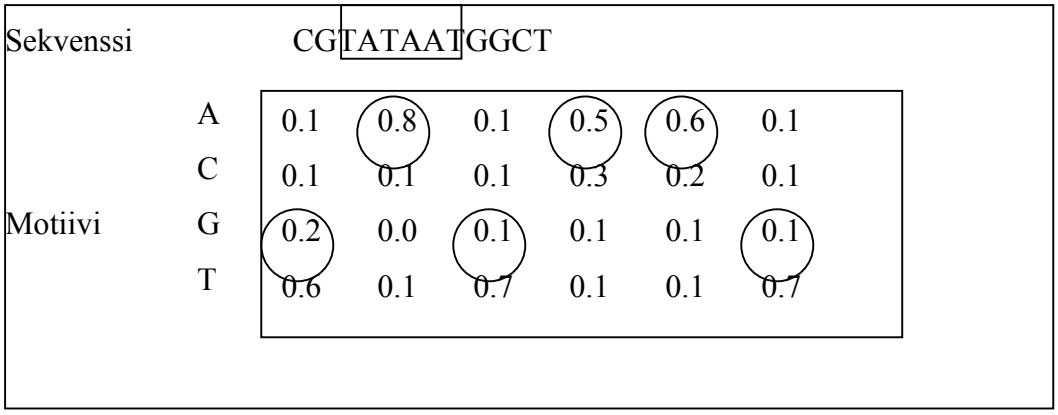
3 MEME, MAST ja Meta-MEME

MEME on ohjelmisto, joka tunnistaa motiiveja, kuten geenien säätelyosia ja proteiinien toiminnallisia domeeneja (proteiinirakenteen itsenäisesti poimuttuva yksikkö). MEME:lle annetaan syötteenä kohdistamatonta sekvenssidataa sisältäviä ryhmiä, joista motiivit etsitään odotuksen maksimoinnilla. Ongelmaa voidaan kuvata ajattelemalla joukkoa merkkijonoja, joista halutaan löytää likimäärin samanlaisia osamerkkijonoja, ja ryhmitellä nämä merkkijonot yhteisten osien mukaisiin luokkiin. Löydetyistä motiiveista voidaan valita kiinnostavimmat, ja niiden avulla voidaan hakea sekvenssitietokannoista luokkien uusia jäseniä. MAST ja Meta-MEME ovat ohjelmistoja, joilla tällaista hakua voidaan tehdä.

3.1 Motiivien etsiminen – MEME (Multiple EM for Motif Elicitation)

MEME:ssä [BaE95] motiivi on tilastollinen malli, jossa annetaan todennäköisyys jokaiselle N:n merkin pituiselle osamerkkijonolle. Jos osamerkkijono sopii malliin, sitä kutsutaan motiivin instanssiksi mallin määrittämällä todennäköisyysasteella. Malli kuvataan matriisina, jossa sarakkeiden lukumäärä on motiivin pituus N; rivien lukumäärä on aakkoston koko, joka on proteiinien tapauksessa 20 (aminohapot, joista proteiini muodostuu). Matriisin alkio (i, j) on todennäköisyys, että aakkonen i on motiivin instanssin kohdassa j. Kuvassa 1 on esitetty DNA-sekvenssi ja motiivi, johon on merkitty osasekvenssin aakkosten todennäköisyydet. DNA-

sekvenssin tapauksessa aakkosia on neljä. Motiivimatriisi esittää suhteellisia frekvenssejä, jotka havaittaisiin jokaisessa sarakkeessa kaikkien mahdollisten hahmoa kuvaavien esimerkkisekvenssien rinnastuksessa.

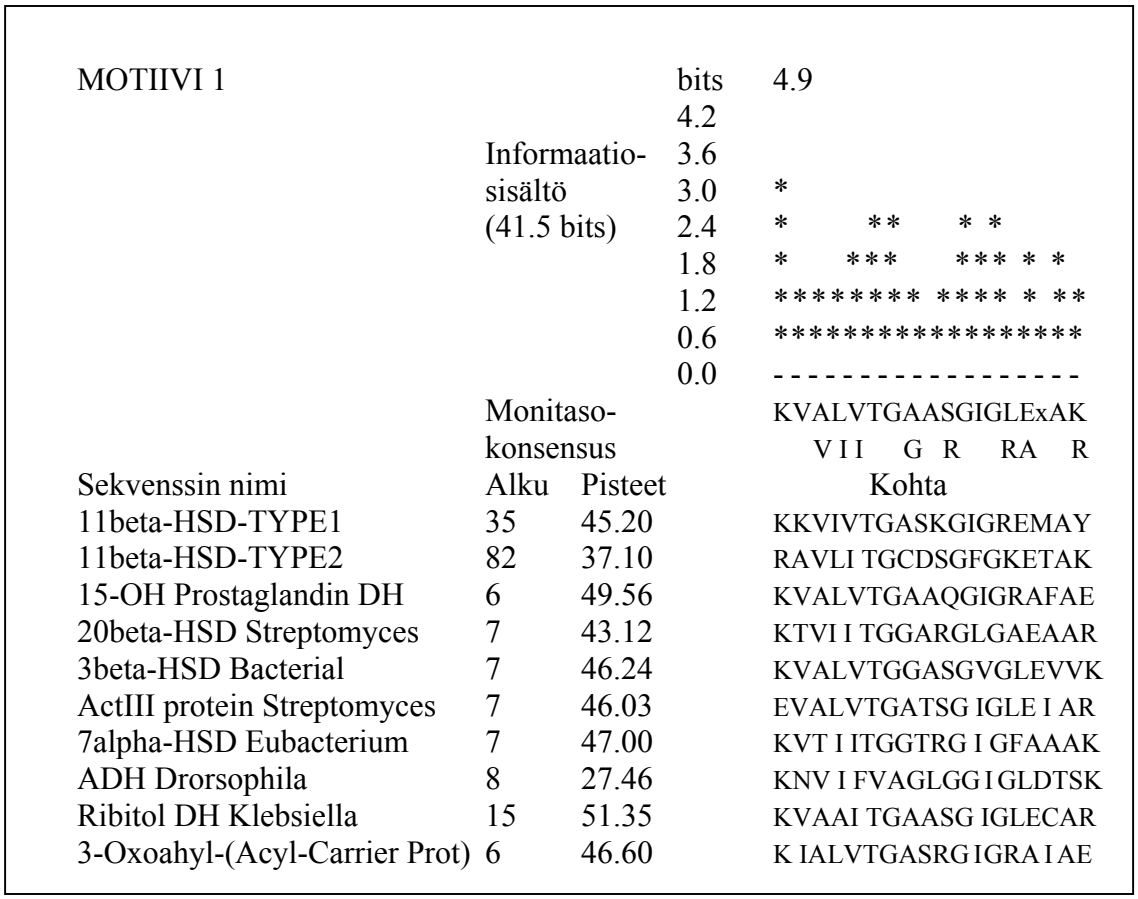


Kuva 1 MEME:n motiivimatriisi ja DNA-sekvenssi. Motiivimatriisissa esitetään aakkosten suhteelliset frekvenssit motiivin jokaisessa positiossa [BaE95].

MEME saa syötteenä joukon DNA- tai proteiinisekvenssejä ja tuottaa joukon tilastollisia sekvenssimalleja, joista jokainen kuvaa yhden motiivin, jonka parametrit on estimoitu EM-algoritmilla. MEME etsii yhden motiivin kerrallaan analysoitavasta sekvenssiryhmästä. Jokaiselle löytämälleen motiiville MEME maksimoi heuristisen funktion mallin parametrien löytämiseksi. Motiivista raportoidaan

- matriisi, jossa esitetään jokaisen aakkosen todennäköisyys motiivin jokaisessa positiossa,
- motiivin todennäköisin esiintymiskohta jokaisessa datajoukon sekvenssissä,
- diagrammi informaatiosisällöstä motiivin jokaisessa positiossa,
- tiivistelmä motiivin konsensussekvenssistä ja
- positioriippuvainen pisteytysmatriisi. [BBEG99]

Esimerkki motiiviraportista on esitetty kuvassa 2.



Kuva 2 Esimerkki MEME-motiivista. Dehydrogenaasin ominaispiirteita kuvaava motiivi. Informaatiosisältöä kuvaava diagrammi osoittaa informaation säilyneisyyden motiivin jokaisessa positiossa. Konsensussekvenssi diagrammin jokaisen pylvään alla kuvaa kohtia joissa tietty aminohappo esiintyy vähintään 20% todennäköisyydellä [BBEG99].

Toisin kuin motiivin kuvaamiseen käytetty todennäköisyysmatriisi, pisteytysmatriisi on logaritminen uskottavuusmatriisi, joka ottaa myös huomioon todennäköisyyden, jolla aminohappo esiintyy motiivin ulkopuolella. Pisteytysmatriisia käytetään motiiviin sopivien sekvenssien etsinnässä. Esimerkki pisteytysmatriisista on kuvassa 3.

MEME-algoritmi on yhdistelmä EM-algoritmista, EM-pohjaisesta heuristiikasta, jolla algoritmin aloituspiste valitaan, heuristiikasta, jolla mallin vapaiden parametrien lukumäärä estimoidaan, sekä ahneesta hausta useiden motiivien löytämiseksi [BaE95b]. Estimoitavat arvot ovat motiivin alkupositioista muodostettu matriisi (piilomuuttujat) ja kirjainten todennäköisyydet motiivin jokaisessa positiossa.

Alla on esitetty MEME-algoritmi pseudokoodilla [BaE95]. Algoritmissa esiintyvä PASSES on käyttäjän määräämä etsittävien motiivien maksimilukumäärä.

```
for i=0 to PASSES {
        jokaiselle syötedatan osamerkkijonolle {
                suorita EM kerran jokaisella osamerkkijonosta johdetulla aloituspisteellä
                valitse todennäköisin jaetun motiivin malli
                suorita EM suppenemiseen asti samalla aloituspisteellä
                tulosta jaetun motiivin malli
                poista motiivin esiintymät datajoukosta
        }
}
```

EM-algoritmi suoritetaan kerran eri aloituskohdille, jotka on johdettu syötedatan osasekvensseistä. Jokainen ajo tuottaa todennäköisyysmallin, joka kuvaa mahdollista datasta löytyvää motiivia. Aloituskohta joka tuottaa parhaan todennäköisyysmallin (yhden ajon perusteella), valitaan uudeksi aloituskohdaksi, josta algoritmi ajetaan suppenemiseen asti. Aloituskohdalla tarkoitetaan tässä alkuperäisiä kirjainten todennäköisyysmatriiseja. EM-algoritmi löytää yhtäaikaisesti motiivimallin ja estimoi syötedatan sekvensseistä jokaisen mahdollisen motiivin aloituskohdan todennäköisyyden.

MEME-algoritmissa logaritminen uskottavuusfunktio on määritelty

$$\log(\text{likelihood}) = N\sum_{j=1}^{W}\sum_{l\in L} f_{lj}\log(\rho_{lj}) + N(P-W)\sum_{l\in L} f_{l0}\log(\rho_{l0}) + N\log\left(\frac{1}{P-W+1}\right),$$

missä N on datajoukossa olevien sekvenssien lukumäärä, P sekvenssien pituus, W jaetun motiivin pituus, L sekvenssien aakkosto, ρ_{lj} kirjaimen l motiivin kohdassa j oleva (tuntematon) todennäköisyys, ρ_{l0} kirjaimen l kaikissa ei motiivin positioissa oleva (tuntematon) todennäköisyys, f_{lj} kirjaimen l motiivin kohdassa j oleva havaittu frekvenssi ja f_{l0} kirjaimen l kaikissa ei motiivin positioissa havaittu frekvenssi. Todennäköisin aloituspiste valitaan logaritmisen uskottavuusfunktion perusteella; toinen mahdollisuus olisi käyttää informaatiosisällön mittaa.

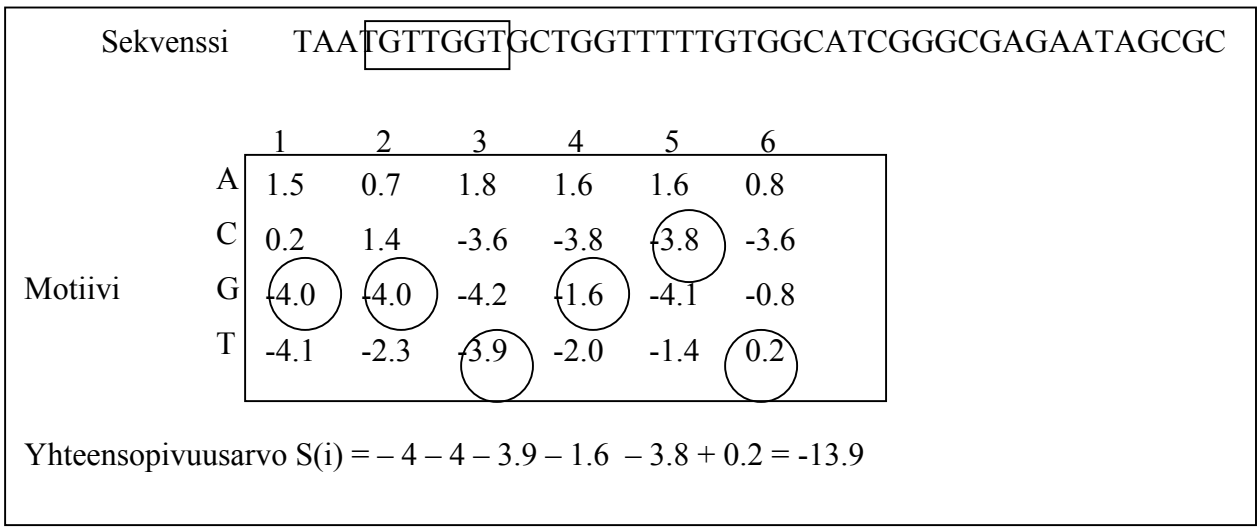
Kirjainten frekvenssimatriisin alkuarvoiksi asetetaan valitun aloituspisteen kirjaimille arvo X, siten että $0<X<1$, ja muille mahdollisille kirjaimille arvo $(1-X)/(L-1)$, jossa L on aakkoston koko. Näin saadaan sarakkeen frekvensseille summaksi 1.0 ja X lähelle arvoa 1.0. Jos asetettaisiin X=1.0 ja muille arvot 0, EM ei konvergoisi. X:n arvot 0.4–0.8 ovat osoittautuneet testeissä suunnilleen yhtä hyviksi [BaE95]. Sekvenssistä johdettu aloituskohta on perusteltu, jos motiivimalleissa ei sallita lisäyksiä ja poistoja; tällöin motiivin täytyy vastata melko tarkasti syötteessä olevia todellisia sekvenssejä. Näin kasvatetaan todennäköisyyttä, että EM konvergoi globaaliin maksimiin.

3.2 Jatkokäsittely – MAST (Motif Alignment and Search Tool) ja Meta-MEME

MAST on työkalu, jolla sekvenssitietokannasta voidaan etsiä MEME:n löytämän motiivin toteuttavia sekvenssejä [BaG98]. Jokaiselle tutkittavan sekvenssin osasekvenssille lasketaan yhteensopivuusarvo MEME:n pisteytysmatriisin avulla. Algoritmille annetaan kynnysarvo, jonka ylittävien osasekvenssien katsotaan olevan motiivin kanssa riittävän yhteensopivia.

MAST käyttää QFAST-algoritmia sekvenssien ja motiivien yhteensovittamiseen. Algoritmissa lasketaan merkitsevyystasojen (p-arvojen) tulon jakauma, jonka perusteella sekvenssin ja syötteenä annettujen motiivien yhtenevyyden tilastollinen merkitsevyys määritellään.

MAST saa syötteenä joukon motiiveja ja vertaa joukkoa jokaiseen sekvenssikannan sekvenssiin. Jokaiselle motiiville etsitään sekvenssistä parhaiten sopiva aloituskohta pisteytysmatriisin avulla, lasketaan osuman p-arvo, joka normalisoidaan sekvenssin pituuden suhteen. Jokaisen sekvenssin normalisoidut p-arvot kerrotaan keskenään, ja tulon p-arvo otetaan tilastollisen merkitsevyyden mittariksi. Aloituskohdan pisteytyksen laskeminen on esitetty kuvassa 3. Kaikki positiot sekvenssissä, joiden arvo on yli algoritmille annetun kynnysarvon, palautetaan tuloksena.



Kuva 3 Motiivin ja sekvenssin yhteensovitus MAST-ohjelmassa. Yhteensopivuusarvo lasketaan summaamalla motiivin pisteytysmatriisista sekvenssin aakkosia vastaavat lukuarvot.

MAST-ohjelmassa ei ole otettu huomioon, että motiivit esiintyvät tyypillisesti tietyssä järjestyksessä yhden sekvenssiperheen sisällä. Tämän tyyppistä informaatiota mallinnetaan usein kätketyillä Markov-malleilla (Hidden Markov Models, HMM), joita on kuvattu toisaalla tässä teoksessa[Koh03]. MEME-ohjelmistosta on kehitetty laajennus, Meta-MEME, joka samoin kuin

MAST ottaa syötteenä MEME:n löytämän motiivijoukon [BBEG99]. Meta-MEME rakentaa joukosta HMM-mallin, jota voidaan käyttää homologien etsintään sekvenssitietokannoista. Algoritmi laskee kannan jokaiselle sekvenssille pisteytyksen (logaritminen uskottavuusmatriisi), joka on verrannollinen todennäköisyydelle, että sekvenssi oli annetun mallin generoima. Samoin kuin MAST, Meta-MEME saa todennäköisyydelle kynnysarvon, jonka ylittävien sekvenssien ajatellaan kuuluvan motiivin kuvaamaan sekvenssiperheeseen.

3.3 Tuloksia

MEME, MAST ja Meta-MEME –työkaluja testattiin pienellä opetusjoukolla, johon kuului 10 dehydrogenaasia ja 4 seriiniproteaasia, jotka kuuluvat eri proteiinisuperperheeseen. [BBEG99] Seriiniproteaasit ovat entsyymejä, jotka liittyvät esimerkiksi proteiinien aineenvaihduntaan ja kasvainten etäpesäkkeiden muodostumiseen. Dehydrogenaasit ovat lyhytketjuisia alkoholeja, jotka säätelevät esimerkiksi ihmisten androgeenin, estrogeenin ja adrenaliinisteroidien pitoisuuksia.

MEME:n löytämistä 12 ensimmäisestä motiivista vain dehydrogenaaseilla oli motiivit 1-3 ja 6-9, ja vain seriiniproteaaseilla motiivit 4 ja 5 sekä 10-12. Tämä antaa viitteitä siitä, että MEME pystyy onnistuneesti erottamaan proteiiniperheitä joukosta, jossa on muihinkin superperheisiin kuuluvia sekvenssejä.

MEME:n löytämät seitsemän dehydrogenaasimotiivia ja viisi seriiniproteaasimotiivia annettiin seuraavaksi MAST:in syötteeksi. Haku tehtiin Genpept98-sekvenssikantaan. Tulokset osoittivat, että ensimmäinen motiivijoukko kuului dehydrogenaaseja kuvaaviin motiiveihin, ja toinen joukko kuvasi seriiniproteaaseja. Motiivit eivät esiinny tilastollisesti merkitsevinä muissa proteiiniperheissä.

Meta-MEME:lle muodostettiin HMM-mallit seitsemästä edellä mainitusta MEME:n löytämästä dehydrogenaasimotiivista ja neljästä seriiniproteaasimotiivista. Tämän jälkeen niitä käytettiin haussa Genpept98-kantaan. Meta-MEME löysi yli 600 dehydrogenaasihomologia ja yli 600 seriiniproteaasia Genpeptin yli 200000 proteiinin joukosta.

MEME ja MAST ovat pystyneet identifioimaan kaukaista sukua olevia proteiinihomologeja, joita muilla sekvenssianalyysityökaluilla, kuten BLAST, Fasta ja PROSITE, ei ole onnistuttu löytämään. Meta-MEME:n tulokset ovat olleet vielä lupaavampia kuin MAST:in. Esimerkiksi sokeriepimeraasien kuulumisesta lyhytketjuisten dehydrogenaasien perheeseen on saatu tilastollista näyttöä. MEME:n löytämät motiivit ovat sopusoinnussa 3D-malliltaan tunnettujen dehydrogenaasien kanssa, joten motiivien järjestystä voidaan käyttää kuvauksena kolmiulotteisesta informaatiosta yksiulotteiseen sekvenssianalyysiin.

4 Yhteenveto

Motiivien löytäminen DNA- ja proteiinisekvenssidatasta antaa arvokasta tietoa monista biologisista prosesseista, ja auttaa kehittämään monia bioteknologian sovelluksia, kuten lääkkeitä sekä kasvien ja eläinten malleja.

EM-algoritmi on monella alalla käytetty ohjaamattoman oppimisen menetelmä, jolla voidaan estimoida havaintodatan perusteella havainnot tuottaneen mallin parametreja ja havaintodatasta puuttuvia arvoja. Menetelmää on käytetty MEME-ohjelmistossa, josta tutkimukset ovat osoittaneet sen pystyvän luokittelemaan sekvenssidataa ja löytämään proteiiniperheitä kuvaavia motiiveja. MEME:n löytämiä motiiveja voidaan käyttää syötteenä MAST ja Meta-MEME -työkaluille, kun tehdään sekvenssitietokantahakua. Näin voidaan löytää proteiiniperheiden kaukaisia homologeja.

Käyttämällä ohjaamattomasti oppivaa järjestelmää motiivien etsinnässä vältytään ihmistarkkailijan aiheuttamalta vinoumalta esimerkiksi motiivin aloituskohdan valinnassa. Sekvenssidatan valtavan määrän takia myös tiedon esikäsittelyn määrän väheneminen on suuri etu.

Viitteet

[BaE95] Timothy L. Bailey, Charles Elkan, Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. *Machine Learning Journal*, 21, 51-83 (1995).

[BaE95b] Timothy L. Bailey and Charles Elkan, The value of prior knowledge in discovering motifs with MEME. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pp. 21-29, AAAI Press, Menlo Park, California, 1995.

[BaG98] Timothy L. Bailey, Michael Gribskov, Combining Evidence Using p-values: Application to Sequence Homology Searches. *Bioinformatics*, 14(48-54), 1998.

[BBEG99] Timothy L. Bailey, Michael E. Baker, Charles P. Elkan, William N Grundy, MEME, MAST, and Meta-MEME: New Tools for Motif Discovery in Protein Sequences. In *Pattern Discovery in Biomolecular Data*. Oxford University Press, 30-54.

[DLR77] Dempster, A.P., Laird, N.M., and Rubin, D.B., Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Ser. B., 39, 1 -38, 1977.

[JuM00] Jurafsky, D., Martin, J.H., Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall, Inc., 2000.

[Koh03] Jukka Kohonen. Geenien etsintä. Teoksessa *Tiedon louhinta biomolekyyli-aineistoista*, 89-99. Raportti C-2003-52, Helsingin yliopisto, Tietojenkäsittely-tieteen laitos, marraskuu 2003.

[Koi02] Petri Koistinen, Tilastollinen hahmontunnistus. Kurssimoniste, Rolf Nevanlinna -instituutti. 2002.

[Uus03] Esa Uusipaikka, Hierarkinen tilastollinen malli ja EM-algoritmi. http://users.utu.fi/esauusi/kurssit/bioinformatiikka/