RNA:n sekundäärirakenteen ennustaminen

Anu Määttä
Anu.Maatta@cs.helsinki.fi

Tiedon louhinta biomolekyyliaineistoista
Helsingin yliopisto, tietojenkäsittelytieteen laitos
Raportti C–2003–09, s. 80–88, marraskuu 2003

Tiivistelmä

RNA eli ribonukleiinihappo toimii geneettisen informaation välittäjänä DNA:sta proteiiniksi sekä avustaa muiden molekyylien toimintaa toimimalla entsyyminä. RNA:n rakenteen selvitys on siis tärkeää solun toiminnan ymmärtämiseksi. RNA:n sekundäärirakenteella tarkoitetaan emäspariutumisten muodostamien osien ja niiden välisten silmukoiden muodostamaa kompleksia. Rakenteen ennustamisessa tärkeä merkitys on minimienergiamenetelmillä: rekursiivisella algoritmilla saadaan määritettyä ne silmukoiden ja pariutuneiden osien yhdistelmä, jossa molekyyli on stabiileimmassa muodossaan. Laajentamalla rekursiivista algoritmia saadaan laskettua myös biologisesti tärkeitä alitoptimaalisia rakenteita. Vertailevaa sekvenssianalyysia toiminnallisesti ja rakenteellisesti samankaltaisen RNA-molekyylien kesken käytetään myös yleisesti rakenteen määrityksessa. Yhdistämällä minimienergiamenetelmät vertailevaan sekvenssianalyysiin saadaan luotettavampia ennusteita RNA:n sekundäärirakenteesta.

1 Johdanto

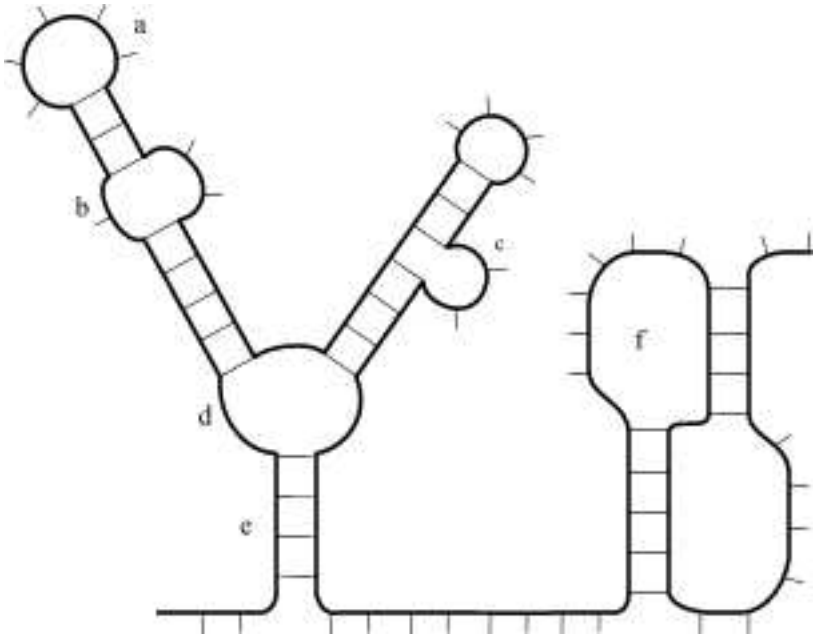
Johdannossa esittelen RNA:n rakenteeseen liittyviä käsitteitä.

RNA on ribonukleiinihappo, joka on muodostunut emäksistä A (adeniini), G (guaniini), C (sytosiini) ja U (urasiili, korvaa DNA:n tymiinin). Se toimii proteiinisynteesissä välivaiheena. RNA:lla on lisäksi entsymaattista aktiivisuutta, eli se edistää muiden molekyylien toimintaa.

RNA ei kietoudu kaksoisjuosteeksi DNA:n tavoin vaan esiintyy yksijuosteisena. RNA voi emäspariutua itsensä kanssa. Emäspariutumisen säännöt ovat hieman erilaiset DNA:han verrattuna: perinteisten parien AU (DNA:ssa AT) ja GC lisäksi RNA:ssa esiintyy pari GU.

RNA:n rakennetta voi muiden biomolekyylien, kuten proteiinien tavoin tarkastella eri tasoilla. Mitä korkeammalle tasolle mennään, sitä lähempänä ollaan biologista ymmärrystä.

1. Primaarirakenne. RNA:n sekvenssi, esim. ACGGUGGGC...
2. sekundäärirakenne. RNA-molekyyli emäspariutuu itsensä kanssa. Muodostuu pariutuneita alueita ja niiden välisiä silmukoita (kuva 1).
3. tertiäärirakenne. Silmukkarakenne kietoutuu kolmiulotteiseksi rakenteeksi emäspariutumista heikompien interaktioiden avulla.
4. Kvarternaarinen rakenne. Eri RNA-molekyylit muodostavat keskenään ja muiden molekyylien kanssa monimutkaisempia komplekseja.



Kuva 1: RNA:n sekundäärirakenteen slimukkatyypit.

Tässä esityksessä keskitytään kuvaamaan sekundäärirakenteen ennustamista, sillä jo tertiäärirakenteen ennustamiseen ei nykyinen laskentateho riitä. sekundäärirakenne on ennustettavissa käyttämällä (lähes) pelkästään sekvenssiaineistoa, sillä sekundäärirakenteen muodostamiseen eivät vaikuta tertiäärirakenteen heikommat interaktiot. Erilaisia sekundäärirakenteita on arvioitu olevan 1.8^n kpl, missä n on sekvenssin pituus [ZuS84].

Kuvassa 1 on esitetty sekundäärirakenteen muodostavat perusyksiköt:
a. Hiusneulasilmukka (hairpin loop)
b. Sisäsilmukka (internal loop)
c. Ulkoneva silmukka (bulge loop)
d. Liitoskohta (junction)
e. Emäspariutunut alue.
f. Valesolmu (pseudoknot)

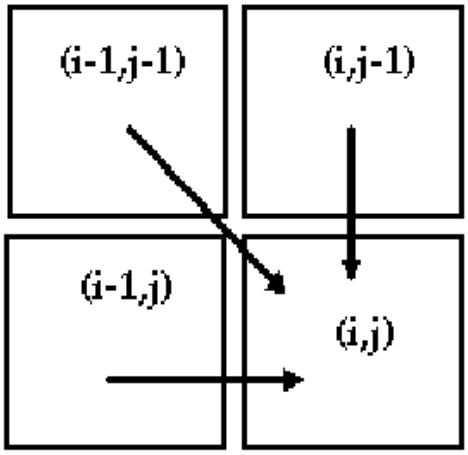
2 Energiamimialgoritmit

Tässä luvussa kerron vapaan energian minimoinnin avulla tapahtuvasta RNA:n sekundäärirakenteen ennustamisesta. Kappaleessa 2.1 kuvataan energiaminimin käsite ja sen merkitys rakenteen muodostumisessa. Kappaleessa 2.2 kerrotaan energiaminimin hakemisen perusalgoritmista. Kappaleessa 2.3 kappaleessa kuvataan vielä nykyisinkin käytössä olevan algoritmiin tehdyistä lisäyksistä perusmenetelmään verrattuna. Kappaleessa 2.4 kerrotaan energiaminimimenetelmien huonoista puolista.

2.1 Energiaminimi

Biomolekyylin energiaminimikonformaatio on se rakenne, jonka ylläpitäminen vaatii vähiten energiaa ja on siten stabiilein rakenne.

Nykyisin RNA:n rakenteen ennustamiseen käytetään pääasiassa kahdentyyppisiä algoritmeja: energiaminimin etsiviä ja sekvenssirinnastukseen perustuvia.



Kuva 2: Matriisin täyttö dynaamisessa algoritmissa.

Minimienergia-algoritmeissa energiaparametrien tarkkuudella on keskeinen osa. Kullekin sekvenssin alirakenteelle (kuva 1) on biokemiallisesti määritetty ns. energiaparametrit (yksikkö kcal/mol), eli millaisia laskostumisenergioita rakenne saa eri emäsjärjestyksillä. Valesolmujen energiaparametrien määrittäminen on ollut hankalaa, joten niiden ennustaminen minimienergia-algoritmeilla on vaikeaa.

2.2 Rekursiiviset algoritmit

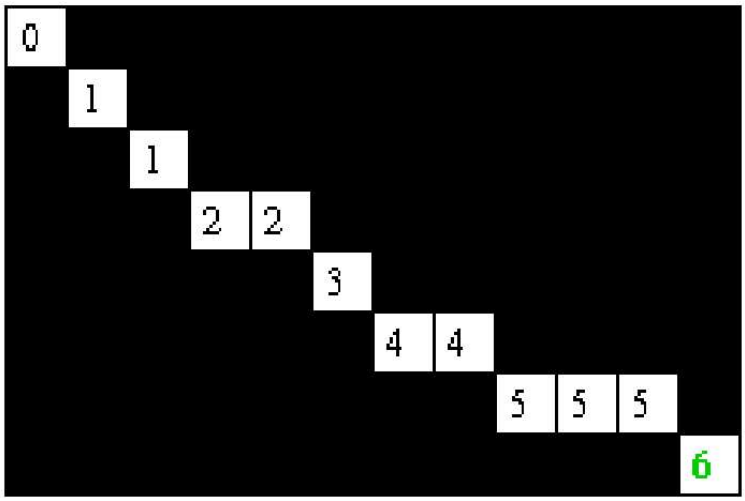
Ensimmäiset vapaan energian minimoimiseen tähtäävät algoritmit ottivat huomioon ainoastaan emäspariutumisten energian. Tällä liian yksinkertaisella oletuksella ei kuitenkaan voinut kovin varmasti ennustaa oikeaa sekundäärirakennetta. Näistä parannelluilla algoritmeilla otettiin energiaparametreja huomioon monipuolisemmin: yksittäisten emäsparien energian sijaan tarkasteltiin eri alirakenteiden (kuva 1) energioita kokonaisuuksina [ZuS81].

Rekursiivisissa laskostamisalgoritmeissa löydetään koko sekvenssin laskostumisen minimenergia yhdistämällä lyhyempien RNA-juosteiden laskostumisenergiat. Tämä tapahtuu dynaamisella ohjelmoinnilla, jossa optimiratkaisuun päädytään kahden vaiheessa. Ensimmäisessä vaiheessa taulukoidaan laskennan välituloksia matriisiin, jonka täytyttyä optimiratkaisu voidaan lukea viimeiseksi täytetystä ruudusta. Toisessa vaiheessa optimiratkaisuun johtanut polku jäljitetään palaamalla matriisissa takaisin alkuun ja tarkastelemalla samalla indeksejä.

RNA:n rakenteen ennustamisessa matriisiin talletetaan siis ensimmäiselle riville ja sarakkeelle lyhyiden, yleensä 5 nukleotidin mittaisten juosteiden laskostumisenergiat. Matriisia täytetään seuraavasti: kuhunkin matriisin alkioon (i,j) tuleva arvo on minimi sitä välittömästi edeltävien alkioiden arvoista (i-1,j-1), (i, j-1) ja (i-1, j) (kuva 2). Näin saadaan määritettyä yhä pidempien juosteiden ja lopulta koko sekvenssin laskostumisenergia. Matriisin täytyttyä voidaan viimeisestä alkiosta lukea optimiratkaisu eli laskostumisen absoluuttinen minimienergia (esim. kuvassa 3).

Ratkaisun jäljitysvaiheessa luetaan minimienergiarakenne aloittamalla matriisin vasemmasta alakulmasta edeten pitkin välituloksia (kuva 3). Samalla luetaan laskostumisen muodostavat emäsparit matriisin rivi- ja sarakeindekseistä.

Dynaamisessa ohjelmoinnissa syntyvästä matriisista saadaan jäljitettyä ainoastaan absoluuttisen energiaminimin omaava rakenne. Tämä ei kuitenkaan ole kovin perusteltua, sillä termodynamiikan sääntöjen mukaan biomolekyylin rakenne on tiheysfunktion määrittämä



Kuva 3: Rakenteen jäljitys optimiratkaisusta lähtien.

konsensus monesta, energiaminimiä hyvin lähellä olevasta rakenteesta [Ped00]. Vaihtoehtoisia rakenteita saadaan tällä menetelmällä aikaan vain energiaparametreja muuttamalla, mikä kasvattaa laskostamisen aikavaatimusta epäkäytännöllisen suureksi.

2.3 Alioptimaalisten ratkaisujen etsiminen

Zuker [Zuk89] ehdotti algoritmiin parannusta, joka ottaa huomioon optimaalisen rakenteen lisäksi myös muut, energiaminimiä lähellä olevat rakenteet.

Zukerin menetelmässä on perusalgoritmiin lisätty seuraava oletus: Lineaarinen RNA-molekyyli muunnetaan kehämäiseksi yhdistämällä molekyylin päät. Kehämäisessä molekyylissa laskostamisen voi aloittaa ottaen lähtökohdaksi minkä hyvänsä emäsparin, merkitään tätä paria (i,j). Kehämäinen molekyyli voidaan jakaa kahteen osaan pariutuneiden emästen i ja j välillä:

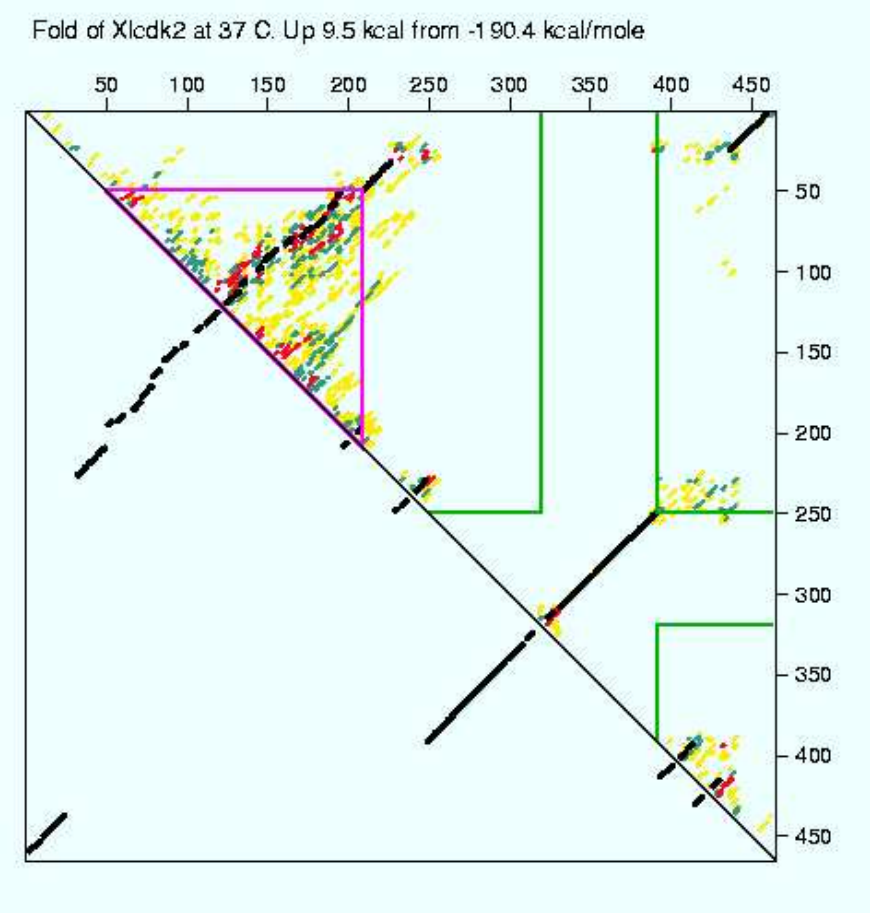
- laskostumisen muodostava osa i:stä j:hin, (i, j)
- laskostumisen ulkopuolelle jäävä osa j:stä molekyylin päiden kautta (nyt yhdistettyinä) i:hin, (j, i)

Molempien osien laskostumisenergiat lasketaan erikseen dynaamisella algoritmilla ja merkitään V(i,j) sekä V(j,i). Koko molekyylin minimilaskostumisenergia, E_{min} on siten $E_{min} = V(i,j) + V(j,i)$

Laskostamisen lähtökohdaksi otetaan vuorollaan kaikki mahdolliset emäsparit. Alioptimaalisen ratkaisun etsimisessä otetaan huomioon kaikki emäsparit, joilla E_{min} on tarpeeksi lähellä aitoa energiaminimiä.

P-optimaaliseksi emäspariksi kutsutaan paria, jolle $V(i,j)+V(j,i) \geq (1-P/100)*E_{min}$. P-arvolla ilmaistaan kuinka monta prosenttia ratkaisujoukkoon hyväksytyt laskostumiset saavat poiketa aidosta energiaminimistä. P-arvoa käytetään absoluuttisen energiamäärän sijaan, koska energiaparametreja ei ole määritelty tarpeeksi hyvin.

Ratkaisujoukkoa tarkastellaan energiamatriisin (kuvassa 4 avulla. Siinä piirretään jokainen P-optimaalinen emäspari i x j -matriisiin, josta nähdään kunkin emäsparin sijoittuminen energiaminimin suhteen. Kukin energiamatriisissa tarpeeksi lähellä absoluuttista minimienergiaa sijaitseva emäspari valitaan vuorollaan laskostamisen lähtökohdaksi, jolloin saadaan laskettua kaikki P-optimaaliset laskostumiset.



Kuva 4: Energiamatriisi.

Koska lineaarinen RNA-molekyyli oletettiin kehämäiseksi, on molekyylin päät sisältävää silmukkaa tarkasteltava erikseen laskostumisissa. Tällainen silmukka katkaistaan molekyylin päätöskohdasta, jolloin silmukkaan jää pariutumattomia emäksiä.

Energiaminimialgoritmeja kutsutaan yleisesti MFOLD-algoritmeiksi. Ne ovat hyvin yleisesti tutkijoiden käytössä www-palvelimilla [1]. Energiaparametreja on nykyisin käytössä oleviin MFOLD-palvelimiin parannettu [Mat99], mutta niillä saadaan ennustettua oikea rakenne ainoastaan 73 prosentin varmuudella [Mat02].

2.4 Energiamimimialgoritmien rajoitukset

Energiaminimialgoritmit ovat nykyäänkin yleisesti käytössä. Niillä on kuitenkin rajoituksia:

- Valesolmuja ei pystytä löytämään. Valesolmujen poisjättö sekundäärirakenteisen ennustamisessa on kuitenkin hyvin keinotekoista, sillä ei ole biokemiallistä syytä miksei niitä RNA-molekyylirakenteessa voisi esiintyä. Valesolmujen ennustaminen on niiden energiaparametrien puutteellisuudesta johtuen laskennallisesti vaikeaa. Konetehon kasvaessa on kuitenkin kehitetty myös niiden ennustamiseen kykeneviä menetelmiä [RiE99].

- Minimienergia-algoritmeja voi dynaamisen algoritmin aikavaatimusten vuoksi käyttää vain suhteellisen lyhyille sekvensseille.

- Minimienergian laskeminen vaatii tarkkaa tietoa energiaparametrien oikeista arvoista. Niiden arvioiminen on kuitenkin hankalaa eikä ehdottoman tarkkaa ratkaisua vielä ole.

Koska minimienergia-algoritmeissa otetaan huomioon ainoastaan yhden sekvenssin emäsjärjestys, ei sukulaismolekyylien rakennetta voi mitenkään ottaa huomioon. Tämä kuitenkin säästäisi vaivaa ja ennusteiden paikkansapitävyys olisi varmempaa, sillä biologiassa samankaltaisuutta kahden molekyylin välillä esiintyy yleensä usealla tasolla: jos primäärirakenne eli sekvenssi on samankaltainen, ovat myös sekundäärirakenne ja sitä korkeammat rakenteet mitä luultavimmin samankaltaisia.

3 Muita menetelmiä sekundäärirakenteen ennustamiseen

Tässä luvussa kuvataan lyhyesti muita laskostamisen menetelmiä. Kappaleessa 3.1 esitellään vertailevan sekvenssianalyysin periaate. Kappaleessa 3.2 kerrotaan vertailevan sekvenssianalyysin ja energiaminimimenetelmän yhdistävästä algoritmista.

3.1 Vertaileva sekvenssianalyysi

Vertailevassa sekvenssianalyysissa rinnastetaan (esim. monen sekvenssin rinnastusalgoritmilla, CLUSTALW [Hig94]) homologisia, eli samaa alkuperää olevia ja saman biokemiallisen toiminnan omaavia RNA-sekvenssejä keskenään. Sekvensseistä etsitään sitten sellaisia alueita, jotka

[1] esim. CSC:n palvelin, http://www.csc.fi/molbio/progs/mfold/

a) kykenevät pariutumaan keskenään, esim. ACCCUGUGGGU

```
     U
ACCC
       G
UGGG
     U
```

b) voivat kovarioida. Kovarianssissa jotkin emäkset sekvenssissä voivat vaihtua toisiksi, kunhan niiden pariutumismahdollisuus säilyy. Esim. alla olevat sekvenssit eivät ole täysin identtisiä, mutta ne voivat silti muodostaa rakenteeltaan samankaltaisen silmukan.

```
ACCCUGUGGGU sekvenssi 1
AACCUGUGGUU sekvenssi 2
 x       x
```

x = sekvensseissä 1 ja 2 kovarioivat emäkset.

Emäsparin vaihtumisesta (x:llä merkityt kohdat) huolimatta silmukan rakenne säilyy muuttumattomana.

```
sekvenssi 1      sekvenssi 2
     U                U
ACCC             AACC
       G                G
UGGG             UUGG
 x   U            x   U
```

Sekvenssien sekundäärirakenteen kannalta samana pysyneen alueen löytäminen on vaikeaa ja virhealtista RNA-aakkoston pienen koon ja kovarianssin sallimisen takia [ZuS84].

3.2 Dynalign

Dynalign [Mat02] saa parametrinaan kaksi sekvenssiä, jotka se rinnastaa keskenään. Pariutuvat alueet rinnastetaan täsmälleen, ei-pariutuvien rinnastuksen ei tarvitse olla täsmällinen. Näistä ennustetaan yhteinen sekundäärirakenne energiaminimimenetelmällä. Yhteisessä rakenteessa sallitaan emäspariutuminen vain, kun linjauksen samassa kohdassa sekvenssejä olevat emäsparit voivat kovarioida keskenään. Yhteisen rakenteen minimierengia lasketaan seuraavasti:

$E_{min} = E_{sekvenssi\ 1} + E_{sekvenssi\ 2} + (E_{linjauksen\ aukko}$ * aukkojen lukumäärä).

$E_{sekvenssi1}$ ja $E_{sekvenssi2}$ lasketaan algoritmilla, jossa ainoastaan kunkin emäksen lähin naapuri sekvenssissä otetaan huomioon [Mat99].

```
ACCCUGUGGGU sekvenssi 1
AACCG--GGUU sekvenssi 2
xxxx   xxxx
```

x = pariutumaan kykenevät alueet
- = aukko linjauksessa

Algoritmi yhdistää energiaminimin laskemisen ja vertailevan sekvenssianalyysin. Sillä on mahdollista saada luotettavampi ennuste RNA-molekyylin rakenteesta kuin pelkästään energiaminimimenetelmää käyttämällä. Sen aikavaativuus kahdella sekvenssillä on $O(M^3N^3)$, missä M on lyhemmän sekvenssin pituus. Kolmella sekvenssillä aikavaatimus kasvaa luokkaan $O(M^6N^3)$, mikä estää algoritmin käytön kovin pitkille sekvensseille nykyisellä laskentateholla.

4 Yhteenveto

Energiaminimialgoritmeissa RNA-rakenteen ennustamiseen riittää kun käytössä on ennustettava sekvenssi sekä tiedot energiaparametreista.

Vertaileva sekvenssianalyysi vaatii aina usean sekvenssin rinnastusta. Rinnastus on RNA-rakenteen vaatimien oletuksien takia virhealtista ja perinteisesti osa vertailevalla sekvenssianalyysilla tehtävästä rakenne-ennustuksesta on jouduttu tekemään käsin, mikä hidastaa laajempien aineistojen käsittelyä. Vertaileva sekvenssianalyysi ottaa huomioon biologiassa tärkeän rakenteen ja toiminnon välisen suhteen ja sillä saadaan luotettavia ennusteita rakenteesta kunhan käytössä on toisilleen tarpeeksi läheisiä sekvenssejä.

RNA:n sekundäärirakenteen ennustaminen on haastava ja monipuolinen ongelma. Vielä ratkaistavia asioita ovat biokemiallisin menetelmin tehtävä energiaparametrien optimointi sekä valesolmujen ja aitojen solmujen salliminen rakenteissa sekä sekundäärirakennetta korkeammat rakennetasot. Konetehon lisääntyminen auttaa osaltaan laskemaan rakenteita yhä pidemmille sekvensseille myös korkeammilla rakennetasoilla, mutta RNA:n rakenteen ennustamisessa riittää vielä pohdittavaa.

5 Viitteet

[Hig94]
Higgins D., Thompson J., Gibson T.Thompson J.D., Higgins D.G., Gibson T.J., CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-80, 1994.

[JTZ89]
Jaeger, J. A., Turner, D. H., Zuker, M., Improved predictions of secondary structures for RNA. Proceedings of the National Academy of Sciences 86, 7706-7710, 1989.

[Mat99]
Mathews, D. H., Sabina, J., Zuker, M., Turner, D. H., Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. Journal of Molecular Biology 288, 911-940, 1999.

[MaT02]
Mathews, D. H., Turner, D. H., Dynalign: An Algorithm for Finding the Secondary Structure Common to Two RNA Sequences. Journal of Molecular Biology 317, 191-203, 2002.

[Mou01]
Mount, D. W., Bioinformatics - Sequence and Genome analysis. Cold Spring Laboratory Press, 2001.

[MZT99]
D.H. Mathews, J. Sabina, M. Zuker & D.H. Turner Expanded Sequence Dependence of Thermodynamic Parameters Provides Robust Prediction of RNA Secondary Structure Journal of Molecular Biology,1999

[Ped00]
Pedersen, C., Algorithms in Computational Biology, 2000.

[RiE99]
Rivas, E., Eddy, S.R A dynamic programming algorhitm for RNA structure prediction including pseudoknots. Journal of Molecular Biology 285, 2053-2068, 1999.

[Zuk89]
Zuker, M., On Finding All Suboptimal Foldings of an RNA Molecule. Science 244, 48-53, 1989.

[ZuS81]
Zuker, M., Stiegler, P., Optimal computer folding of large RNA sequences using thermodynamics and auxuliary information, Nucleic Acids Research 9, 133-148, 1981.

[ZuS84]
Zuker, M., Sankoff, D., RNA secondary structures and their prediction. Bulletin of Mathematical Biology, 46(4):591-621, 1984.