Kemistin data-analysaattori

Määrittelydokumentti (2. iteraatio)

Taisto Gerdt

Timo Kääriäinen

Mika Miettinen

Pasi Rastas

Jari Tuominiemi

Ohjelmistotuotantoprojekti

Helsingin yliopisto

Tietojenkäsittelytieteen laitos

30. lokakuuta 2002

Versiohistoria:

Päiväys	Muutokset	Muutosten tekijä
28.10.2002	Alustava versio	Mika Miettinen
29.10.2002	Muutoksia kirjoitusasussa	Mika Miettinen
30.10.2002	Ryhmäkokouksessa 29.10 sovitut muutokset	Mika Miettinen

Sisältö

1 Johdanto

1.1 Projektin aihepiiri

Mikrobisolujen kasvua voidaan analysoida muun muassa mittaamalla solun tuottamia kemiallisia yhdisteitä. Erityisen tehokas menetelmä on niinkutsuttu isotooppijäljitys, jossa solulle syötetään hiili-13 -isotooppia sisältävää ravinnetta ja mitataan solun tuottamien yhdisteiden isotopomeerijakaumia, eli isotooppien jakaumia yhdisteen hiiliketjussa. Näiden isotopomeerijakaumien määrittämiseksi käytetään niinkutsuttua tandemmassaspektrometria, jossa voidaan mitata sekä koko molekyylin tuottama spektri että molekyylin palasten spektrit. Yhdistämällä eri spektrien tiedot isotopomeerijakaumasta saadaan tarkempi arvio kuin käyttämällä yhtä spektriä.

2 Ohjelmiston yleiskuvaus

2.1 Toimintaympäristö

Ohjelmisto toimii Windows 2000 -käyttöjärjestelmällä varustetuissa mikrotietokoneissa ja koostuu joukosta Matlab-ohjelmiston ohjelmointikielellä toteutettuja niinsanottuja m-tiedostoja. Myös ohjelman käyttöliittymä toteutetaan Matlabilla, joten kyseessä tulee olemaan puhdas Matlab-sovellus.

2.2 Ohjelmiston käyttäjät

Ohjelmiston lopullisena käyttäjänä on Raimo Ketola VTT:ltä.

2.3 Toteutettava ohjelmisto

Ryhmän toteuttaman ohjelmiston pääasiallinen käyttötarkoitus on molekyylin positionaalisten isotopomeerien jakauman laskeminen massaspektrometrin tuottaman datan perusteella. Laskenta tapahtuu laitoksella kehitetyn menetelmän [Ran02, RoR02] mukaan.

Tämän lisäksi ryhmä toteuttaa fragmenttitietokannan aminohappomolekyylien sekä näistä syntyvien fragmenttien ominaisuuksien tallettamiseksi. Tällä tietokannalla on oleellinen osa pääohjelman suorittamassa isotopomeerijakauman laskemisessa.

Kehitystyön pohjana ryhmä käyttää aiemmin laitoksella kehitettyä PIDC-nimistä Matlab-sovellusta [Ran01].

3 Ohjelmiston tarjoamat palvelut

3.1 Ohjelmiston toiminta

Pääpiirteissään ohjelmiston toiminta koostuu seuraavista askeleista:

1. Ohjelma lukee käyttäjän valitseman syötetiedoston.
2. Ohjelma nimeää syötetiedoston sisältämät piikistöt atomikaavoilla käyttäen apuna fragmenttitietokantaa.
3. Piikistöt nimettyään ohjelma laskee isotopomeerien jakauman.
4. Ohjelma tallentaa näin lasketun isotopomeerijakauman tekstitiedostoon.

4 Ulkoiset liittymät

4.1 Syötetiedosto

Ohjelman lukema syötetiedosto sisältää massaspektrometrin mittauksista saadun datan. Käsittelyn helpottamiseksi tiedoston oletetaan sisältävän vain yhden äitimolekyylin tiedot. Lisäksi oletetaan, että tiedoston ensimmäinen datarivi sisältää perusmassaisen äitimolekyylin tiedot.

Syötetiedosto sisältää rivejä, jotka ovat joko kommenttirivejä tai kuvaavat yhtä massaspektrin piikkiä. Kommenttirivit alkavat prosenttimerkillä ja sisältävät tämän jälkeen mitä tahansa merkkejä jatkuen aina seuraavaan rivinvaihtomerkkiin saakka. Kukin datarivi sisältää kolme tabulaattorein (tai välilyönnein) erotettua arvoa: ensimmäisenä molekyylin massa (oikeammin massa/varaus -suhde) ja toisena sen intensiteetti. Kolmannelle sarakkeelle on kolme mahdollisuutta: se voi olla tyhjä, sisältää pienen kokonaisluvun tai sisältää M-kirjaimen. Jos sarake sisältää M-kirjaimen, kuvaa kyseinen piikki äitimolekyyliä. Jos sarake taas on tyhjä, on massaspektri tuotettu “Full scan”-menetelmällä ja piikki kuvaa tytärionia. Jos massaspektri on tuotettu “Daughter ion scan”-menetelmällä, on sarakkeessa kokonaisluku. Tällöin kyseinen piikki kuvaa tytärionia ja luku kertoo mistä äitimolekyylin isotopomeeristä kyseinen fragmentti on peräisin.

Heti tiedoston alussa olevat kommenttirivit tulkitaan liittyvän itse mittaukseen ja ne liitetään sellaisenaan mukaan tulostiedostoon. Syötetiedoston muoto EBNF-notaatiolla esitettynä on siis (`NL` on rivinvaihto, `CHARACTER` mielivaltainen merkki, `TAB` tabulaattori, `SPACE` välilyönti, `INTEGER` ei-negatiivinen kokonaisluku ja `EMPTY` tyhjä merkkijono)

```
input          ::= input-line NL (input-line NL)*
input-line     ::= comment | peak
comment        ::= '%' CHARACTER*
peak           ::= mass ws intensity ws mother
mass           ::= INTEGER
intensity      ::= INTEGER
```

```
mother          ::= INTEGER | EMPTY | 'M'
ws              ::= TAB | SPACE
```

4.2 Tulostiedosto

Ohjelman tuottama tulostiedosto sisältää laskennan tuloksena syntyneen isotopomeerijakauman puhtaana tekstitiedostona. Tulostiedoston alussa on tiedot laskennan päiväyksestä, sekä mittauksessa käytetystä energiamäärästä ja laitteesta. Näiden jälkeen tiedostossa on syötetiedoston alussa olleet kommenttirivit ja laskennassa käytetyt syötetetiedoston datarivit referoituna.

Tämän jälkeen tiedosto sisältää rivejä, joista kukin ilmoittaa jonkin positionaalisen isotopomeerin suhteellisen osuuden. Rivi sisältää ensin tiedon isotopomeeristä muodossa $[b_1 b_2 ... b_n]$ (luvut b_i ovat bittejä jotka ilmoittavat onko paikassa i sijaitseva hiili leimattu vai ei) ja tämän jälkeen tabulaattorilla erotettuna kyseisen isotopomeerin suhteellisen osuuden. Jos ohjelma ei saanut isotopomeerijakaumaa täysin ratkaistua (kuten yleensä käy), jotkut rivit sisältävät yhden positionaalisen isotopomeerin sijasta summan useammista positionaalisista isotopomeereistä.

Näin muotoiltuna tulostiedosto on luettavissa yleisimmillä taulukkolaskentaohjelmilla.

4.3 Fragmenttitietokanta

Isotopomeerijakauman laskemiseen käytettävän ohjelman lisäksi ryhmä toteuttaa fragmenttitietokannan. Tietokanta toteutetaan tekstitiedostopohjaisena siten, että kutakin massaspektrometrissä käytettyä energiaa ja laitteistoa tai teknistä järjestelyä kohden on oma tiedostonsa. Tiedostossa esitellään sarakkeittain seuraavat tiedot:

- ioniin liittyvä äitimolekyylin nimi
- ionin asema (M-äitimolekyyli-ioni, D=tytärioni äitimolekyylille)

- ionin kemiallinen kaava
- ionissa mukana olevat hiiliatomit
- ionin kaavapaino
- ionin suhteellinen osuus (vain tytärioneille)

Äitimolekyylin kemiallinen kaava esitetään ionimuodossa siten, kuin se massaspektrissä esiintyy. Näin esimerkiksi kaavapainoltaan 89 oleva alaniini (C3H7NO2) esitetään tiedostossa muodossa C3H8NO2 (kaavapaino 90). Tiedostossa esiintyvä kaavapaino voitaisiin laskea myös suoraan kaavasta, mutta se on otettu mukaan tiedostoon helpottamaan tiedonhakua fragmenttitietokannasta. Viimeisessä sarakkeessa esitetään tytärionin piikin suhteellinen voimakkuus kyseisellä energialla. Jos rivi kuvaa äitimolekyyliä, on sarake tyhjä.

Kuten koko ohjelmakin, myös näitä tiedostoja käsittelevät funktiot (eli "tietokannan hallintajärjestelmä") toteutetaan Matlab-ohjelmointikielellä. Tietokannan käsittely toteutetaan siten, että käyttäjän käyttöliittymässä valitsemaa energiaa ja laitetta vastaava tiedosto luetaan kokonaan keskusmuistiin. Tästä tietorakenteesta karsitaan myöhemmin ne alkiot joita laskennassa ei tarvita. Mahdolliset samanaikaisuuden hallintaan liittyvät ongelmat jätetään huomiotta, sillä ryhmä olettaa että tietokannan tiedostoja ei muuteta ohjelman ollessa käytössä.

5 Käyttöliittymä

Ohjelmiston käyttöliittymä toteutetaan Matlabin käyttöliittymäkomponenttien avulla. Asiakkaat pitivät ohjelmiston ensimmäisen version käyttöliittymää selkeänä ja helppokäyttöisenä, joten käyttöliittymää ei suunnitella uudestaan eikä suuria muutoksia tehdä. Asiakkaan toivomusten mukaisesti liittymän kieleksi vaihdetaan kuitenkin englanti. Käyttöliittymä mahdollistaa seuraavat asiat:

- syöte- ja tulostiedostojen valitsemisen
- mittauksessa käytetyn energiamäärän (eV) määrittelemisen
- mittauksessa käytetyn laitteen valitsemisen
- automaattisesti tunnistettujen fragmenttien atomikaavojen muuttamisen käyttäjän toimesta

Käyttäjän valittavissa olevat laitteet ja energiamäärät päätellään olemassaolevien fragmenttitietokannan tiedostojen perusteella. Ensimmäisen iteraation määrittelydokumentissa mukana ollut kynnysraja-arvon määrittelеminen jätetään ohjelmiston lopullisesta versiosta pois, sillä Juho Rousu ja Ari Rantanen totesivat tämän asian järkevän toteutuksen olevan vaikeaa ja vaativan lisää tutkimusta.

Fragmenttitietokannan ylläpitoon ei kehitetä varsinaista käyttöliittymää, sillä asiakkaiden mielestä tietokannan päivittäminen suoraan tekstieditorin avulla on riittävä ratkaisu. Halutessaan päivittää tietokantaa käyttäjä painaa ohjelman pääikkunassa olevaa painonappia, joka avaa tietokannan oikean tiedoston Matlabin editoriin. Tätä editoria käyttämällä käyttäjä voi poistaa ja lisätä sekä muokata tietokannan tietoja.

6 Käytettävyysvaatimukset

Ohjelma toimii missä tahansa Windows 2000 -käyttöjärjestelmällä varustetussa mikrotietokoneessa johon on asennettu Matlab-ohjelmiston versio 6. Ohjelma toimii todennäköisesti myös muissa käyttöjärjestelmissä, joihin Matlab-ohjelmisto on saatavissa.

Ohjelman suorituksen aikana tapahtuvien virheiden virheilmoitukset tulevat näkyviin ohjelman pääikkunan alalaitaan. Tämä antaa käyttäjälle mahdollisuuden korjata asia ja yrittää uudestaan.

Lähteet

Ran01 Rantanen, A., PIDC (Positional Isotopomer Distribution Calculator), http://www.cs.helsinki.fi/research/icomic/PIDC/PIDC.zip

Ran02 Rantanen, A. et al., Computing Positional Isotopomer Distributions from Tandem Mass Spectrometric Data. *Metabolic Engineering*, hyväksytty.

RoR02 Rousu, J., Rantanen, A., Improved Computation of ^{13}C-Isotopomer Distributions from Tandem-MS Data.