Kemistin data-analysaattori
Määrittelydokumentti (1. iteraatio)

Taisto Gerdt
Timo Kääriäinen
Mika Miettinen
Pasi Rastas
Jari Tuominiemi

Ohjelmistotuotantoprojekti
Helsingin yliopisto
Tietojenkäsittelytieteen laitos
18. syyskuuta 2002

Versiohistoria:

Päiväys	Muutokset	Muutosten tekijä
16.9.2002	Alustava versio	Mika Miettinen
17.9.2002	Ryhmätapaamisessa 16.9 esitetyt muutokset	Mika Miettinen
18.9.2002	Ryhmäkokouksessa 17.9 esitetyt 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. Fragmenttitietokannan ylläpitoa varten ryhmä toteuttaa yksinkertaisen käyttöliittymän.

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.

3.2 Syötetiedosto

Ohjelman lukema syötetiedosto sisältää massaspektrometrin mittauksista saadun datan. Tiedoston alussa on tieto mittauksessa käytetystä energiatasosta. Tämän jälkeen tiedosto sisältää rivejä, joista kukin kuvaa yhtä massaspektrin piikkiä. Rivillä on siis kaksi tabulaattorein erotettua arvoa: molekyylin massa (oikeammin massa/varaus -suhde) ja sen intensiteetti. Jos kyseinen piikki kuvaa tutkittavan molekyylin fragmenttia, on rivillä myös kolmas luku: pieni kokonaisluku, joka ilmoittaa mistä isotopomeeristä kyseinen fragmentti on peräisin. Syötetiedoston muoto EBNF-notaatiolla esitettynä on siis (`NL` on rivinvaihto, `TAB` tabulaattori, `INTEGER` ei-negatiivinen kokonaisluku ja `EMPTY` tyhjä merkkijono)

```
input          ::= energy NL mass-spectrum
mass-spectrum ::= peak NL (peak NL)*
peak           ::= mass TAB intensity TAB mother
mass           ::= INTEGER
intensity      ::= INTEGER
mother         ::= INTEGER | EMPTY
```

3.3 Tulostiedosto

Ohjelman tuottama tulostiedosto sisältää laskennan tuloksena syntyneen isotopomeerijakauman puhtaana tekstitiedostona. Tulostiedosto sisältää aluksi tiedon tutkittavan molekyylin atomikaavasta. Tämän jälkeen tiedosto sisältää rivejä, joista kukin ilmoittaa jonkin positionaalisen isotopomeerin suhteellisen osuuden. Rivi sisältää ensin tiedon isotopomeeristä muodossa $[b_1 b_2 \ldots b_n]$ 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.

3.4 Fragmenttitietokanta

Isotopomeerijakauman laskemiseen käytettävän ohjelman lisäksi ryhmä toteuttaa fragmenttitietokannan. Fragmenttitietokanta sisältää seuraavat tiedot tutkittavista aminohappomolekyyleistä:

- Nimi
- Kemiallinen kaava
- Energia pilkkoutumistietoja varten (molekyyli pilkkoutuu eri tavoilla eri energioilla)
- Massaspektrometrinen menetelmä
- Molekyylin tuottamat fragmentit (kemiallinen kaava) ja suhteelliset osuudet kullakin energiamäärällä

Aluksi tietokanta toteutetaan yksinkertaisesti yhtenä tai useampana tekstitiedostona. Näitä tiedostoja käsittelevät funktiot (eli "tietokannan hallintajärjestelmä") toteutetaan Matlab-ohjelmointikielellä, kuten itse laskentamenetelmäkin. Koska tietokannan odotettavissa oleva maksimikoko on suhteellisen pieni (< 200 kt), luetaan tietokanta kokonaan keskusmuistiin. Samanaikaisuuden hallintaan liittyvät ongelmat jätetään huomiotta, sillä ryhmä olettaa että tietokannan tiedostoja ei muuteta ohjelman ollessa käytössä.

4 Käyttöliittymä

Ohjelmiston käyttöliittymä toteutetaan Matlabin käyttöliittymäkomponenttien avulla ja siitä pyritään tekemään mahdollisimman selkeä ja helppokäyttöinen. Käyttöliittymä mahdollistaa seuraavat asiat:

- Syöte- ja tulostiedostojen valitsemisen
- Automaattisesti tunnistettujen fragmenttien atomikaavojen muuttamisen käyttäjän toimesta
- Käyttäjä voi määritellä, onko syötetiedoston sisältämä massaspektri tuotettu full scan- vai daughter ion scan-menetelmällä
- Antaa käyttäjän määritellä raja-arvon, jonka alle jäävät syötetiedoston piikit jätetään laskennassa huomiotta.

Kuten edellä todettiin, ryhmä toteuttaa myös fragmenttitietokannalle yksinkertaisen käyttöliittymän. Sekin toteutetaan Matlabin avulla. Käyttöliittymä mahdollistaa seuraavat asiat:

- Tietojen lisäämisen tietokantaan
- Tietojen poistamisen tietokannasta
- Tietokannassa jo olevien tietojen muuttamisen

5 Käytettävyysvaatimukset

Ohjelman toiminta ei edellytä työasemalta erityisvaatimuksia. Mikä tahansa Windows 2000 -käyttöjärjestelmällä varustettu mikrotietokone, joka kykenee ajamaan Matlab-ohjelmistoa riittää. On mahdollista että ohjelma valmistuttuaan toimii myös muissa käyttöjärjestelmissä, joihin Matlab-ohjelmisto on saatavissa.

Ohjelman mahdollisista virhetoiminnoista käyttäjä saa välittömästi tiedon näytölle ja ohjeet virheen korjaamiseksi.

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.