Adduct Clustering for MS Data

January 30, 2015

1 Introduction

Based on the group's previous work on MetAssign, we have seen how prior formula information, when available, can be used to model the inter-dependencies between peaks (based on the mass, RT and intensity values and the assignments to theoretical peaks). In this report, we propose a clustering model that do not use such formula information but instead takes into account the adduct relationships between peaks for clustering.

Later on, it might also be useful to incorporate the adduct clustering model proposed here here into HDP-Align. It might also be possible to come up with another model that combines the approach with formula prior (i.e. MetAssign) and without formula prior for the purpose metabolite identity annotations and precursor mass discovery.

2 Preliminary

2.1 Adduct Transformation

Given a set of metabolites (indexed by $i = 1, ..., I$) and a set of adducts (indexed by $k = 1, ..., K$), a precursor mass of a metabolite m_i will be observed with the addition of an adduct a_k. An adduct is encoded in the form:

$$a_k = nM + \sum_j h_j b_j \tag{1}$$

where n is the multiplicity of the molecule, b_j is the adduct part (e.g. O, H, C, N, etc.) and h_j is the count of such adduct part. For example using the adduct 'M+2H+Na': the multiplicity is n=1, the first adduct part is b_1=H, h_1=2 and the second adduct part b_2=Na, h_2=1. Each adduct part b_j has a corresponding known mass $g(b_j)$. The addition of an adduct a_k then defines a

linear transformation $f_k(m_i)$ from a precursor mass m_i to the group of observed peak masses, based on which k-th adduct has been applied. This is defined as

$$f_k(m_i) = \left(\frac{n}{|c|}\right) m_i + \left(\frac{-cE + \sum_j h_j g(b_j)}{|c|}\right) = d_k m_i + h_k \tag{2}$$

where c is the charge, E is the mass of an electron (0.00054857990924). The constants for the linear transformation $d_k m_i + h_k$ are $d_k = \frac{n}{|c|}$ and $h_h = \frac{-cE + \sum_j h_j g(b_j)}{|c|}$.

Following are some examples of adducts (indexed by k) and the corresponding transformation constants.

k	a_k	n	c	d_k	h_k	$f_k(m_i)$
1	M+H	1	1	1	$-E + g(H)$	$m_i - E + g(H)$
2	M+2H	1	2	0.5	$-E + g(H)$	$0.5m_i - E + g(H)$
3	M+Na	1	1	1	$-E + g(Na)$	$m_i - E + g(Na)$
4	M+K	1	1	1	$-E + g(K)$	$m_i - E + g(K)$
5	2M+H	2	1	2	$-E + g(H)$	$2m_i - E + g(H)$

Table 1: Some examples of adduct transformations, indexed by k

The atomic masses of the adduct parts are obtained from http://dx.doi.org/10.1351/pac200375060683 and summarised in below table:

b_j	$g(b_j)$	b_j	$g(b_j)$
O	15.9949146223	Na	22.98976966
H	1.0078250319	K	38.9637069
C	12.0	S	31.97207073
N	14.0030740074		

Table 2: Atomic masses of elements

2.2 Linear transformation of univariate Gaussian random variable

Given a univariate Gaussian random variable X normally distributed with mean μ_X and precision σ_X, we can apply some linear transformation $f(X) = dX + h$ to it and obtain a new random variable Y with mean $\mu_Y = d\mu_X + h$ and precision $\sigma_Y = \frac{1}{d^2} \cdot \sigma_X$.

$$\mu_Y = E\{Y\} = E\{dX + h\} = dE\{X\} + h = d\mu_X + h \tag{3}$$

$$
\begin{aligned}
\sigma_Y^{-1} &= E\{[Y-\mu_Y]^2\} \\
&= E\{[dX+h-d\mu_X+h]^2\} \\
&= E\{[d(X-\mu_X)]^2\} \\
&= d^2 E\{[X-\mu_X]^2\} \\
&= d^2\sigma_X^{-1} \\
\sigma_Y &= \frac{\sigma_X}{d^2}
\end{aligned} \tag{4}
$$

3 A Generative Model

The observed data is the set of N peak masses $X = \{x_1, x_2, ..., x_n\}$, which may be pooled together from different runs (files). We will ignore the chromatography part for now, so, we are not considering retention time and chromatographic peak shapes etc. and assume that all the peaks already share similar RT values (have been aligned). The following generative process is assumed:

1. A run contains multiple metabolites. This is modeled as a Dirichlet Process mixture model, where each mixture component in the DP mixture corresponds to a metabolite. Metabolites are indexed from $i = 1, ..., I$. Each metabolite has exactly one precursor molecular mass m_i, drawn from the base Gaussian distribution with mean μ_0 and precision σ_0

$$m_i|\mu_0, \sigma_0 \sim \mathcal{N}(\mu_0, \sigma_0^{-1}) \tag{5}$$

2. For each metabolite i, we generate exactly K adduct clusters, corresponding to the possible K adduct transformations. The number of adducts (K) and the exact form of adduct transformations (the f_ks in Table 1) must be provided by user as input to the model. The adducts are modeled as a finite mixture of K Gaussian components linked to metabolite i, with uniform prior on the mixture proportion $\boldsymbol{\pi} = (\frac{1}{K}, \frac{1}{K}, .., \frac{1}{K})$.

3. For each adduct cluster a in metabolite i, we then assign it a linear transformation k. The categorical variable $\varphi_{ia} = k \in \{1, ..., K\}$ denotes the pre-determined one-to-one assignment of adduct component a in metabolite i to the linear transformation k. Observed peaks are then generated by sampling from the transformed distribution $f_k(\mathcal{N}(m_i, \lambda^{-1}))$, where m_i is the initial metabolite mass and λ some fixed precision. Given the assignment of peak n to adduct a in metabolite i ($z_{nia} = 1$) and the assignment of adduct transformation k to adduct a in metabolite i ($\varphi_{ia} = k$), we generate peak masses (x_n) by sampling from

$$x_n|z_{nia} = 1, \varphi_{ia} = k \quad \sim \quad f_k(\mathcal{N}(m_i, \lambda^{-1})) \tag{6}$$

Since we know the constants for every transformation k (the d_ks and h_ks in Table 1), the above equation simplifies to

$$x_n|z_{nia} = 1, \varphi_{ia} = k \sim \mathcal{N}(d_k m_i + h_k, d_k^2 \lambda^{-1}) \tag{7}$$

There can be multiple peaks generated here under each transformed cluster. We assume these are peaks coming from different runs and sharing similar RT values (not considered in the model).

4 Inference

Inference is performed via Gibbs sampling. In each iteration of Gibbs sampling, we randomly loop through all peaks, remove the current peak from the model and update the assignment of peak n to metabolite i adduct a. The conditional probability of $p(z_{ni.} = 1|)$ given any other parameter is the following

$$p(z_{ni.} = 1|x_n, ...) \propto \begin{cases} c_i \cdot p(x_n|z_{ni} = 1, ...) \\ \alpha \cdot p(x_n|z_{ni^*} = 1, ...) \end{cases} \tag{8}$$

where:

- c_i is the count of peaks currently assigned to metabolite i
- α is the DP concentration parameter
- $z_{ni} = 1$ denotes the assignment of peak n to an existing metabolite i
- $z_{ni^*} = 1$ denotes the assignment of peak n to a new metabolite i^*.

We consider the top part of eq. (8). The likelihood of peak x_n to be in an existing metabolite i is given by the sum over the finite mixture components of the transformed adducts.

$$\begin{aligned} p(x_n|z_{ni} = 1, ...) &= \frac{1}{K} \sum_k p(x_n|m_i, \lambda, ...) \\ &= \frac{1}{K} \sum_k \mathcal{N}(x_n|d_k m_i + h_k, d_k^2 \lambda^{-1}) \end{aligned} \tag{9}$$

Then we consider the bottom part of eq. (8). The likelihood of peak x_n to be in a new metabolite i^* is obtained by marginalising over all possible values of m_i.

$$\begin{aligned} p(x_n|z_{ni^*} = 1, ...) &= \int p(x_n|m_i) \cdot p(m_i|\mu_0)\, dm_i \\ &= \int \left[\frac{1}{K} \sum_k \mathcal{N}(x_n|d_k m_i + h_k, d_k^2 \lambda^{-1}) \right] \cdot \mathcal{N}(m_i|\mu_0, \sigma_0^{-1})\, dm_i \\ &= \textbf{???} \end{aligned} \tag{10}$$

Also we need to update the metabolite mass m_i given every other parameter.

$$\begin{aligned} p(m_i|x_n,...) &\propto \left[\prod_{n\in i} p(x_n|m_i)\right] p(m_i|\mu_0,\sigma_0) \\ &\propto \left[\prod_{n\in i} \frac{1}{K}\sum_k \mathcal{N}\left(x_n|d_k m_i + h_k, d_k^2\lambda^{-1}\right)\right] \mathcal{N}(m_i|\mu_0,\sigma_0^{-1}) \\ &\propto \left[\prod_{n\in i} \mathcal{N}\left(x_n|\sum_k d_k m_i + h_k, \sum_k d_k^2\lambda^{-1}\right)\right] \mathcal{N}(m_i|\mu_0,\sigma_0^{-1}) \end{aligned}$$

Written in the form of the density functions

$$p(m_i|x_n,...) \propto \left[\prod_{n\in i} exp\left(\frac{-(\sum_k d_k^2\lambda^{-1})}{2} x_n - \sum_k d_k m_i + h_k)^2\right)\right] exp\left(\frac{-\sigma_0}{2}(m_i-\mu_0)^2\right)$$

$p(m_i|x_n,...)$ is proportional to the products of Gaussian, which is a Gaussian.
We equate this to $N(\mu_i,\sigma_i^{-1})$. Then

$$exp\left(\frac{-\sigma_i}{2}(m_i-\mu_i)^2\right) \propto exp\left(\sum_{n\in i}\left[\frac{-(\sum_k d_k^2\lambda^{-1})}{2}(x_n - \sum_k d_k m_i + h_k)^2\right]\right) exp\left(\frac{-\sigma_0}{2}(m_i-\mu_0)^2\right)$$

Let $\sum_k d_k^2\lambda^{-1} = r$ for clarity, then

$$exp\left(\frac{-\sigma_i}{2}(m_i-\mu_i)^2\right) \propto exp\left(\sum_{n\in i}\left[\frac{-r}{2}(x_n - \sum_k d_k m_i + h_k)^2\right]\right) exp\left(\frac{-\sigma_0}{2}(m_i-\mu_0)^2\right)$$

Collecting the quadratic terms for m_i from both sides, we can solve for σ_i

$$\begin{aligned} \sigma_i m_i^2 &= \sum_{n\in i}\left[r(\sum_k d_k)^2 m_i^2\right] + \sigma_0 m_i^2 \\ \sigma_i &= c_i\left[r(\sum_k d_k)^2\right] + \sigma_0 \end{aligned} \tag{11}$$

where c_i is the count of peaks under metabolite i. Collecting the linear terms for m_i from both sides, we can solve for μ_i

$$
\begin{aligned}
\frac{-\sigma_i}{2}(-2m_i\mu_i) &= \sum_{n\in i}\left[\frac{-r}{2}(-2x_n(\sum_k d_k m_i + h_k))\right] + \frac{-\sigma_0}{2}(-2m_i\mu_0) \\
\sigma_i(-2m_i\mu_i) &= \sum_{n\in i}\left[r(-2x_n(\sum_k d_k m_i + h_k))\right] + \sigma_0(-2m_i\mu_0) \\
\sigma_i(m_i\mu_i) &= \sum_{n\in i}\left[rx_n(\sum_k d_k + h_k))\right] + \sigma_0(m_i\mu_0) \\
\sigma_i\mu_i &= r(\sum_k d_k + h_k)\sum_{n\in i} x_n + \sigma_0\mu_0 \\
\mu_i &= \frac{1}{\sigma_i}\left[r(\sum_k d_k + h_k)\sum_{n\in i} x_n + \sigma_0\mu_0\right]
\end{aligned}
$$

References