Supplementary file for “BPDA2d — A 2D global optimization based Bayesian peptide detection algorithm for LC-MS”

Youting Sun[1], Jianqiu Zhang*[2] , Ulisses Braga-Neto[1] and Edward R. Dougherty*[1,3,4]

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[2]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249, USA
[3]Computational Biology Division, Translational Genomics Research Institution, Phoenix, AZ 85004, USA
[4]Department of Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA

Email: Youting Sun - charonsun@tamu.edu; Jianqiu Zhang*- Michelle.Zhang@utsa.edu; Ulisses Braga-Neto - ulisses@ece.tamu.edu; Edward R. Dougherty*- edward@ece.tamu.edu;

*Corresponding author

Bayesian peptide detection

Let $\theta \triangleq \{\lambda_k, c_{k,ij}; k = 1, \ldots, N,\ i = 1, \ldots, cs,\ j = 0, \ldots, iso\}$ be the set of unknown model parameters. Given the observed denoised spectra $\mathbf{y}$, we apply Gibbs sampling [1] to determine the value of θ. Gibbs sampling uses the popular strategy of divide-and-conquer to sample a subset of parameters at a time while fixing the rest at the sample values from the previous iteration, as if they were true. In other words, for the l-th parameter group θ_l, we sample from the conditional posterior distribution $P(\theta_l|\theta_{-l}, \mathbf{y})$, where $\theta_{-l} \triangleq \theta \setminus \theta_l$, with values obtained from the previous iteration. After this sampling process iterates among the parameter groups for a sufficient number of cycles (i.e., the “burn-in” period), convergence is reached. The samples collected afterwards are shown to be from the marginal posterior distribution $P(\theta_l|\mathbf{y})$ which is independent of θ_{-l}, and thus these samples can be used to estimate the target parameters.

The Gibbs sampling process for the kth peptide candidate and the derivations of the conditional posterior distributions of model parameters are given below.

- **Sample the apex vector $\mathbf{c}_k \triangleq [c_{k,ij}; i = 1, \ldots, cs,\ j = 0, \ldots, iso]^T$ for the kth candidate**

  By the Bayesian principle, the conditional posterior distribution of $\mathbf{c}_k$ is proportional to the likelihood times the prior, that is,

$$P(\mathbf{c}_k\,|\mathbf{y}, \theta_{-\mathbf{c}_k}) \propto P(\mathbf{y}|\theta)\mathrm{Prior}(\mathbf{c}_k), \tag{1}$$

where $\theta_{-\mathbf{c}_k} \triangleq \theta \setminus \mathbf{c}_k$.

It is easy to show the likelihood satisfies

$$P(\mathbf{y}|\theta) \propto \exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{G}\lambda^{(0)} - \lambda_k \mathbf{g}_k)^T {\mathbf{\Sigma_e}}^{-1}(\mathbf{y} - \mathbf{G}\lambda^{(0)} - \lambda_k \mathbf{g}_k)\}, \tag{2}$$

where

$$\begin{aligned} \mathbf{y} = [&y(x_1,1), y(x_1,2), \ldots, y(x_1,T), y(x_2,1), y(x_2,2), \ldots, y(x_2,T), \ldots, \\ &y(x_M,1), y(x_M,2), \ldots, y(x_M,T)]^T \end{aligned} \tag{3}$$

is the observed denoised spectra vector.

$$\lambda^{(q)} \triangleq [\lambda_1, \ldots, \lambda_k = q, \ldots, \lambda_N]^T,\ q \in \{0,1\}, \tag{4}$$

is an indicator vector for peptide existence.

$$\mathbf{\Sigma_e} = \text{diag}\left([\sigma_1^2, \ldots, \sigma_T^2; \sigma_1^2, \ldots, \sigma_T^2; \ldots; \sigma_1^2, \ldots, \sigma_T^2]_{1\times MT}\right), \tag{5}$$

with σ_t^2 being the variance of the t-th spectrum.

$$\mathbf{G} = (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N), \tag{6}$$

whose k-th column is given by

$$\begin{aligned} \mathbf{g}_k = [&g_k(x_1,1), g_k(x_1,2), \ldots, g_k(x1,T), g_k(x_2,1), g_k(x_2,2), \ldots, g_k(x_2,T), \ldots, \\ &g_k(x_M,1), g_k(x_M,2), \ldots, g_k(x_M,T)]^T, \end{aligned} \tag{7}$$

which is a $MT \times 1$ vector with the entry $g_k(x_m,t) = \sum_{i=1}^{cs}\sum_{j=0}^{iso} c_{k,ij} l_k(t) I_{x_m=\alpha_{k,ij}}$, $m = 1, 2, \ldots, M$, $t = 1, 2, \ldots, T$, representing the signal at (x_m,t) generated by peptide candidate k.

The heights of the isotopic peaks of peptide candidate k at charge state i follow a multinomial distribution [2], which by the Central Limit Theorem can be approximated by a Gaussian distribution as below:

$$\begin{aligned} P(c_{k,ij}, j = 0, \ldots, iso\,|a_k, \eta_{k,i}, \pi_k\,) &= MN(a_k \eta_{k,i}, \pi_k) && (8) \\ &\approx N(a_k \eta_{k,i} \pi_k, a_k \eta_{k,i}[\text{diag}(\pi_k) - \pi_k^T \pi_k]), && (9) \end{aligned}$$

where a_k is the total apex intensity of candidate k, $\eta_k \triangleq [\eta_{k,1}, \eta_{k,2}, \ldots, \eta_{k,cs}]^T$ denotes the candidate's charge state distribution, and $\pi_k \triangleq [\pi_{k,0}, \pi_{k,1}, \ldots, \pi_{k,iso}]^T$ is the theoretical isotopic distribution estimated by the Averagine approach [3, 4].

Thus the prior distribution of the peak height vector $\mathbf{c}_k$ is given by:

$$\text{Prior}(\mathbf{c}_k) = P(\mathbf{c}_k \,|a_k, \eta_k, \pi_k\,) \approx N(\mu_{\mathbf{c}_k}, \mathbf{\Sigma}_{\mathbf{c}_k}), \tag{10}$$

where

$$\begin{aligned} \mu_{\mathbf{c}_k} &= [a_k\eta_{k,1}\pi_k^T, a_k\eta_{k,2}\pi_k^T, \ldots, a_k\eta_{k,cs}\pi_k^T]^T, & (11)\\ \mathbf{\Sigma}_{\mathbf{c}_k} &= \text{diag}(\Sigma_i), & (12) \end{aligned}$$

with

$$\Sigma_i = a_k\eta_{k,i}[\text{diag}(\pi_k) - \pi_k^T\pi_k],\; i = 1, 2, \ldots, cs. \tag{13}$$

Substituting Eq. 2 and Eq. 10 into Eq. 1 and it can be shown by algebraic manipulations [5] that the conditional posterior distribution of $\mathbf{c}_k$ is also Gaussian, with the mean vector and covariance matrix given below:

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{c}_k|\mathbf{y},\theta_{-\mathbf{c}_k}} &= (\mathbf{I} - \mathbf{K}\mathbf{H}_k)\mathbf{\Sigma}_{\mathbf{c}_k}, & (14)\\ \mu_{\mathbf{c}_k|\mathbf{y},\theta_{-\mathbf{c}_k}} &= \mu_{\mathbf{c}_k} + \mathbf{K}(\mathbf{y} - \mathbf{G}\lambda^{(0)} - \mathbf{H}_k\mu_{\mathbf{c}_k}), & (15) \end{aligned}$$

where $\mathbf{H}_k = [h_{ms,(i-1)\times(iso+1)+j+1}]_{MT\times cs(iso+1)}$ is the elution profile matrix of candidate k. The $[(i-1)\times(iso+1)+j+1]$th column contains the normalized elution profile of candidate k at charge state i and isotopic number j which has been estimated in preprocessing steps. And $\mathbf{K} \triangleq \mathbf{\Sigma}_{\mathbf{c}_k}\mathbf{H}_k^T\left(\mathbf{H}_k\mathbf{\Sigma}_{\mathbf{c}_k}\mathbf{H}_k^T + \mathbf{\Sigma}_e\right)^{-1}$ is known as the Kalman gain matrix [6].
Note that the matrices involved in the above equations have huge dimensions which make the calculation almost infeasible. Thus, to update each peptide's signal, the related matrices $\mathbf{K}, \mathbf{G}, \mathbf{H}, \mathbf{y}$ and $\mathbf{\Sigma}_e$ are restricted to the corresponding peptide signal regions. This does no harm to the calculation accuracy while dramatically increases the speed.

- **Sample a_k, the total apex intensity of candidate k**

  The conditional distribution of a_k takes different forms for different values of λ_k.

  When $\lambda_k = 1$ (the kth candidate is inferred to be present), by definition,

$$a_k\,|(c_{k,ij}, \lambda_k = 1) = \sum_{i=1}^{cs}\sum_{j=0}^{iso} c_{k,ij} \cdot I_{c_{k,ij}>0}. \tag{16}$$

When $\lambda_k = 0$ (the kth candidate is inferred to be absent), the distribution of a_k, which is independent of the observation $\mathbf{c}_k$, is modeled by a uniform distribution as below:

$$P(a_k \,|c_{k,ij}, \lambda_k = 0\,) = \text{Unif}(0, u_k), \tag{17}$$

where u_k is the upper bound of a_k.

- **Sample $\eta_k \triangleq [\eta_{k,1}, \eta_{k,2}, \ldots, \eta_{k,cs}]^T$, the charge state distribution of candidate k**

  Unlike the isotopic distribution, the charge state distribution cannot be theoretically predicted even when the peptide sequence is given. Thus η_k needs to be estimated by the Gibbs sampling process. Let $\mathbf{b}_k \triangleq [b_{k,1}, b_{k,2}, \ldots, b_{k,cs}]^T$, where $b_{k,i}$ is the total apex abundance of peptide k at charge state i. Given the charge state distribution and the total apex abundance of peptide k, the likelihood of $\mathbf{b}_k$ is multinomial:

$$P(\mathbf{b}_k|\eta_k, a_k) = \text{MN}(a_k, \eta_k). \tag{18}$$

  As is well known, the conjugate prior to a multinomial likelihood is Dirichlet, which is also a reasonable choice for the prior of η_k. Thus, let the prior of η_k be a Dirichlet distribution with parameter $w\alpha$, where w is a weight parameter that controls the strength of the prior information. A small w is preferable if uncertainty resides in the prior, and vice versa. Then the posterior distribution of η_k is given by

$$\begin{aligned} P(\eta_k\,|\mathbf{b}_k\,) &\propto P(\mathbf{b}_k\,|\eta_k\,)\text{Prior}(\eta_k) & (19)\\ &= \text{Dirichlet}(w\alpha + \mathbf{b}_k). & (20) \end{aligned}$$

- **Sample the peptide existence indicator variable λ_k**

  The conditional posterior distribution of λ_k is given by

$$\begin{aligned} P(\lambda_k\,|\mathbf{y}, \theta_{-\lambda_k}\,) &\propto P(\mathbf{y}\,|\theta\,)\text{Prior}(\lambda_k) \\ &\propto \exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{G}\lambda)^T\Sigma_e^{-1}(\mathbf{y} - \mathbf{G}\lambda)\}\text{Prior}(\lambda_k), & (21) \end{aligned}$$

  where $\mathbf{G}$ is defined in Eq. 6.

  The log-likelihood ratio (LLR) of λ_k can be calculated as below

$$\begin{aligned} LLR_{\lambda_k} &= \ln\frac{P(\lambda_k = 1\,|\mathbf{y}, \theta_{-\lambda_k}\,)}{P(\lambda_k = 0\,|\mathbf{y}, \theta_{-\lambda_k}\,)} \\ &= -\frac{1}{2}\left[(\mathbf{y} - \mathbf{G}\lambda^{(1)})^T\Sigma_e^{-1}(\mathbf{y} - \mathbf{G}\lambda^{(1)}) - (\mathbf{y} - \mathbf{G}\lambda^{(0)})^T\Sigma_e^{-1}(\mathbf{y} - \mathbf{G}\lambda^{(0)})\right] + \ln\frac{P(\lambda_k = 1)}{P(\lambda_k = 0)}, \end{aligned} \tag{22}$$

where $\lambda^{(q)}, q \in \{0,1\}$ is defined by Eq. 4.

If no prior knowledge is available about which peptide candidates are more likely to be present in the sample, then a reasonable choice for the prior of λ_k could be the uniform distribution. But we would like to be a bit conservative about the existence of peptide candidates. The idea is that by adding more candidates, it is possible to reduce the mean squared error (MSE) between the inferred spectra and the observed denoised spectra, but at the same time the chances of overfitting increases as the model becomes more complex. Thus, a prior based on Bayesian information criterion (BIC) [7] is adopted to resolve the problem by introducing a penalty term for the number of parameters of the model. And the above equation can be rewritten as:

$$LLR_{\lambda_k} = -\frac{1}{2}\left[(\mathbf{y} - \mathbf{G}\lambda^{(1)})^T \Sigma_e^{-1} (\mathbf{y} - \mathbf{G}\lambda^{(1)}) - (\mathbf{y} - \mathbf{G}\lambda^{(0)})^T \Sigma_e^{-1} (\mathbf{y} - \mathbf{G}\lambda^{(0)})\right] - \frac{\ln(MT)}{2}\Delta, \tag{23}$$

where $\Delta = Card(\theta) - Card(\theta_{-\lambda_k, -c_k}) = Card(\mathbf{c}_k)$ is the difference between the number of free parameters of the two models – with and without candidate k, respectively.

The conditional posterior distribution of λ_k is then obtained based on the log-likelihood ratio as follows:

$$P(\lambda_k = 1 \,|\mathbf{y}, \theta_{-\lambda_k}) = \frac{1}{1 + e^{-LLR_{\lambda_k}}}, \tag{24}$$

$$P(\lambda_k = 0 \,|\mathbf{y}, \theta_{-\lambda_k}) = 1 - P(\lambda_k = 1 \,|\mathbf{y}, \theta_{-\lambda_k}). \tag{25}$$

The pseudocode of the Gibbs sampling process is given in Table 1.

Table 1: The Gibbs sampling process

```
Cluster candidates into $S$ clusters.
Sort clusters by their importance in descending order.
For iteration $r = 1$ to $R$
  For cluster $s = 1$ to $S$
    For peptide candidate $k = i_1^s$ to $i_{N_s}^s$
      Draw $\mathbf{c}_k^r$ based on its conditional posterior distribution.
    end of k loop
    Draw $\lambda_k^r, k = 1 \ldots, i_{N_s}^s$ for the cluster according to the
    joint conditional posterior distribution.
  end of $s$ loop
end of $r$ loop
```

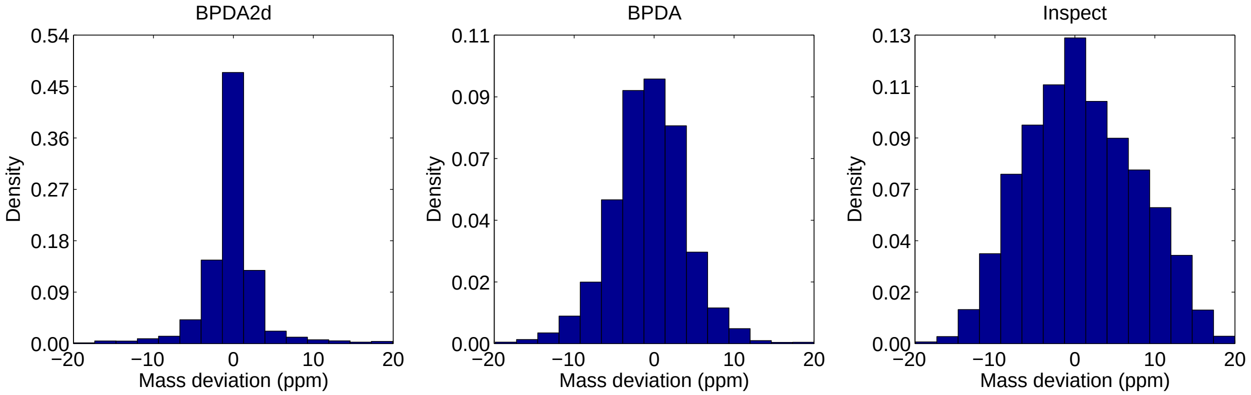

Figure 1: Mass deviation of reported features that can be matched to the ground truth peptide list using a 20 ppm mass window (along with other criteria imposed on the retention time as mentioned in the paper). Each panel represents a detection algorithm as suggested by the subtitle. The plot was obtained by normalizing the mass deviation histogram by the total number of true peptides. It can be seen that BPDA2d has a much higher mass accuracy than the other two algorithms: the density around 0 ppm given by BPDA2d increased by around 4 times compared to BPDA and msInspect; and the SD of mass deviation is 3.7, 4.6, and 6.9 ppm for BPDA2d, BPDA and msInspect, respectively.

Table 2: Statistics of detection results

	TP	FP	TN	FN	ACC^1	SPC^2
BPDA2d	16	16	102	0	0.88	0.86
BPDA	15	58	207	1	0.79	0.78
msInspect	12	4	1	4	0.62	0.2

[1] Accuracy: $ACC \triangleq \frac{TP+TN}{P+N}$

[2] Specificity: $SPC \triangleq \frac{TN}{FP+TN}$

Supplementary results

Figure 1: Mass accuracy of different algorithms in the 100-mix LC-MS data sets

Table 2: Synthetic LC-MS data set with 8 pairs of overlapping peptides

Table 3: Running time on test datasets

References

1. Geman S, Geman D: **Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images**. *IEEE Trans. Pattern Anal. Mach. Intell.* 1984, **6**:721–741.

2. Kaur P, O'Connor PB: **Use of statistical methods for estimation of total number of charges in a mass spectrometry experiment**. *Analytical Chemistry* 2004, **76**:2756–2762.

3. Senko MW, Beu SC, McLafferty FW: **Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions**. *J Am Soc Mass Spectrom* 1995, **6**:229–233.

Table 3: Running time

	100-mix	16-mix	QTOF LC-MS/MS
BPDA2d	35 min	4 hr	6 hr
BPDA	25 min	20 min	2.25 hr
msInspect	0.5 min	2 min	4 min

4. Horn DM, Zubarev RA, McLafferty FW: **Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules**. *Journal of the American Society for Mass Spectrometry* 2000, **11**(4):320–332.

5. Anderson BDO, Moore JB: *Optimal filtering*. Englewood Cliffs, NJ, USA: Prentice-Hall 1979.

6. Burgers G, Leeuwen PJ, Evensen G: **Analysis scheme in the ensemble Kalman filter**. *Monthly Weather Review* 1998, **126**:1719–1724.

7. Schwarz G: **Estimating the dimension of a model**. *Ann. Stat.* 1978, **6**:461–464.