Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data

Oliver Serang
Department of Genome Sciences
University of Washington
Seattle, WA, USA

Michael J. MacCoss
Department of Genome Sciences
University of Washington
Seattle, WA, USA

William Stafford Noble
Department of Genome Sciences
Department of Computer Science and Engineering
University of Washington
Seattle, WA, USA

Derivation of the Model and Algorithms

Here we derive our model and optimizations from our assumptions. Some equations are numbered to facilitate implementation of the model in software; these equations correspond to a function in our C++ implementation.

Basic Notation

Using standard statistical notation, random variables are given capital letters. For a random variable (for instance R) a particular value that can be taken on by that random variable is given with the corresponding lowercase letter (r). The event that the random variable takes the value is noted $R = r$.
For any random variable R we will write $\forall r$ to mean "the set of all possible values for the random variable R." If R represents the set of present proteins, then $\forall r$ will represent all possible sets of present proteins (*i.e.* the power set). If R_ρ indicates whether protein ρ is present or absent, then $\forall r_\rho$ will represent the instance where it is present (we will write r_ρ) and the case when it is absent (we will write $\overline{r_\rho}$).

Definitions

R	array of indicators for presence of proteins in the sample, indexed by ρ
E	array of indicators for presence of peptides in the sample, indexed by ϵ
D	observed data, with paired spectra and total masses indexed by δ
$m_{\rho,\epsilon}$	indicator for protein ρ is expected to produce peptide ϵ using the digest
$G_{\rho,\epsilon}$	the event that peptide ϵ is produced by emission from a protein
H_ϵ	the event that peptide ϵ is produced by an event other than $G_{\rho,\epsilon}$

We denote the size of these arrays as $size(R), size(E), \ldots$.
The numbered equations in this write-up are used in the implementation.

Model

We parameterize the models of G and H in terms of unknown rates α and β.

Assumptions

Assumption $A1$**:** Conditional Independence of E_{ϵ_1} and E_{ϵ_2} given R

$$\Pr(E = e|R = r) = \prod_{\epsilon} \Pr(E_\epsilon|R = r)$$

Assumption $A2$**:** Conditional Independence of D_{δ_1} and D_{δ_2} given E

$$\Pr(D|E = e) = \prod_{\delta} \Pr(D_\delta|E = e)$$

Assumption $A3$**:** Probability of predicted peptide creation

$$\Pr(G_{\rho,\epsilon}|R_\rho) \quad = \quad \begin{cases} \alpha, & m_{\rho,\epsilon} = 1 \\ 0, & else \end{cases}$$

Assumption $A4$**:** Probability of sponteneous peptide emission given the peptide is not created by proteins

$$\Pr(H_\epsilon|\forall\rho,\ \overline{G_{\rho,\epsilon}}) \quad = \quad \beta$$

Assumptions $A5, A6$**:** Prior, independent belief on presence of proteins

$$\forall\rho, \Pr(R_\rho) = \gamma$$

$$\Pr(R = r) = \gamma^{|r|}(1-\gamma)^{size(R)-|r|}$$

Assumption $A7$**:** Conditional Independence of D and R given E

$$\Pr(D|E = e, R = r) = \Pr(D|E = e)$$

Assumption *A*8**:** Dependence of D_δ only on best matching peptide E_ϵ

$$\Pr(D_\delta|E = e) = \Pr(D_\delta|E_{\epsilon(\delta)} = e_{\epsilon(\delta)}),$$

$$\Pr(D_\delta|E_{\epsilon(\delta)} = e_{\epsilon(\delta)}, \alpha, \beta) = \Pr(D_\delta|E_{\epsilon(\delta)} = e_{\epsilon(\delta)})$$

given $\epsilon(\delta)$ is the index of the best matching peptide to spectrum and precursor mass pair δ

$$\epsilon(\delta) = \arg\max_{\epsilon'} \Pr(E_{\epsilon'}|D_\delta)$$

$$\delta(\epsilon) = \arg\max_{\delta'} \Pr(E_\epsilon|D_{\delta'})$$

Furthermore, a unique $\epsilon(\delta)$ is assumed to exist for each δ, and a unique $\delta(\epsilon)$ is assumed to exist for each ϵ.

Assumption *A*9**:** Independence of predicted peptide creation

$$\Pr(G_{\rho_1,\epsilon}, G_{\rho_2,\epsilon}) = \Pr(G_{\rho_1,\epsilon}) \Pr(G_{\rho_2,\epsilon})$$

Method

Method Given Known α, β

Let the proteins and peptides be partitioned so that all proteins in each partition have no connections to peptides in another partition, and all peptides in each partition have no connections to proteins in another partition. This can be trivially accomplished by tracing the graph with depth first search. Denote the protein set corresponding to partition i as $r^{(i)}$ and the peptides in the partition $e^{(i)}$.

$$\Pr(R^{(i)} = r^{(i)}|D) = \frac{\Pr(D|R^{(i)} = r^{(i)}) \Pr(R^{(i)} = r^{(i)})}{\sum_{\forall r^{(i)\prime}} \Pr(D|R^{(i)} = r^{(i)\prime}) \Pr(R^{(i)} = r^{(i)\prime})}$$

This probability can be defined in terms of a likelihood function:

$$\Pr(R^{(i)} = r^{(i)}|D) = \frac{L(R^{(i)} = r^{(i)}|D) \Pr(R^{(i)} = r^{(i)})}{\sum_{\forall r^{(i)\prime}} L(R^{(i)} = r^{(i)\prime}|D) \Pr(R^{(i)} = r^{(i)\prime})}$$

$$\begin{aligned}
\Pr(D|R^{(i)} = r^{(i)}) &= \sum_{\forall e^{(i)}} \Pr(D|E^{(i)} = e^{(i)}) \Pr(E^{(i)} = e^{(i)}|R^{(i)} = r^{(i)}) \\
&= \sum_{\forall e^{(i)}} \prod_{\delta} \Pr(D_\delta|E^{(i)} = e^{(i)}) \Pr(E^{(i)} = e^{(i)}|R^{(i)} = r^{(i)}) \\
&= \sum_{\forall e^{(i)}} \prod_{j \neq i} \prod_{\delta} \Pr(D_\delta^{(j)}) \prod_{\epsilon} \Pr(D_{\delta(\epsilon)}^{(i)}|E_\epsilon^{(i)} = e_\epsilon^{(i)}) \Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}|R^{(i)} = r^{(i)})
\end{aligned}$$

$$\Pr(D_{\epsilon(\delta)}^{(i)}|E_\epsilon^{(i)} = e_\epsilon^{(i)}) = \Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}|D_{\epsilon(\delta)}^{(i)}, Q) \frac{\Pr(D_{\delta(\epsilon)}^{(i)}|Q)}{\Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}|Q)}$$

$$\Pr(D|R^{(i)}=r^{(i)}) =$$
$$\prod_j \left[\prod_\delta \Pr(D_\delta^{(j)}|Q)\right] \sum_{\forall e^{(i)}} \prod_\epsilon \frac{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|D_{\delta(\epsilon)}^{(i)},Q)}{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|Q)} \Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|R^{(i)}=r^{(i)})$$

$$L(R^{(i)}=r^{(i)}|D) =$$
$$\sum_{\forall e^{(i)}} \prod_\epsilon \frac{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|D_{\delta(\epsilon)}^{(i)},Q)}{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)},Q)} \Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|R^{(i)}=r^{(i)})$$

$$= \frac{\Pr(E_1^{(i)}|D_{\delta(\epsilon)}^{(i)},Q)}{\Pr(E_1^{(i)},Q)} \Pr(E_1^{(i)}|R^{(i)}=r^{(i)})$$
$$\sum_{\forall e^{(i)}:e_1^{(i)}} \prod_{\epsilon\neq 1} \frac{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|D_{\delta(\epsilon)}^{(i)},Q)}{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)},Q)} \Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|R^{(i)}=r^{(i)})$$
$$+ \frac{\Pr(\overline{E_1^{(i)}}|D_{\delta(\epsilon)}^{(i)},Q)}{\overline{\Pr(E_1^{(i)},Q)}} \Pr(\overline{E_1^{(i)}}|R^{(i)}=r^{(i)})$$
$$\sum_{\forall e^{(i)}:\overline{e_1^{(i)}}} \prod_{\epsilon\neq 1} \frac{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|D_{\delta(\epsilon)}^{(i)},Q)}{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)},Q)} \Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|R^{(i)}=r^{(i)})$$

$$= \frac{\Pr(E_1^{(i)}|D_{\delta(\epsilon)}^{(i)},Q)}{\Pr(E_1^{(i)},Q)} \Pr(E_1^{(i)}|R^{(i)}=r^{(i)})$$
$$\sum_{\forall e^{(i)}:e_1^{(i)}} \prod_{\epsilon\neq 1} \frac{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|D_{\delta(\epsilon)}^{(i)},Q)}{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)},Q)} \Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|R^{(i)}=r^{(i)})$$
$$+ \frac{\Pr(\overline{E_1^{(i)}}|D_{\delta(\epsilon)}^{(i)},Q)}{\overline{\Pr(E_1^{(i)},Q)}} \Pr(\overline{E_1^{(i)}}|R^{(i)}=r^{(i)})$$
$$\sum_{\forall e^{(i)}:e_1^{(i)}} \prod_{\epsilon\neq 1} \frac{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|D_{\delta(\epsilon)}^{(i)},Q)}{\Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)},Q)} \Pr(E_\epsilon^{(i)}=e_\epsilon^{(i)}|R^{(i)}=r^{(i)})$$

$$= \left(\frac{\Pr(E_1^{(i)}|D_{\delta(\epsilon)}^{(i)}, Q)}{\Pr(E_1^{(i)}, Q)} \Pr(E_1^{(i)}|R^{(i)} = r^{(i)}) + \frac{\Pr(\overline{E_1^{(i)}}|D_{\delta(\epsilon)}^{(i)}, Q)}{\Pr(\overline{E_1^{(i)}}, Q)} \Pr(\overline{E_1^{(i)}}|R^{(i)} = r^{(i)}) \right)$$
$$\sum_{\forall e^{(i)}: e_1^{(i)}} \prod_{\epsilon \neq 1} \frac{\Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}|D_{\delta(\epsilon)}^{(i)}, Q)}{\Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}, Q)} \Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}|R^{(i)} = r^{(i)})$$

$$\cdots$$

$$= \prod_\epsilon \sum_{\forall e_\epsilon^{(i)}} \frac{\Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}|D_{\delta(\epsilon)}^{(i)}, Q)}{\Pr(E_\epsilon^{(i)} = e^{(i)}, Q)} \Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}|R^{(i)} = r^{(i)})$$

where the likelihood constant does not depend on R, α, β. This final transformation uses the conditional independence assumption to convert the sum over all peptide sets into a product over peptide indices. Thus the likelihood can be computed for each protein set without computing the sum over all peptide sets.

$$\Pr(E_\epsilon^{(i)}) = \sum_{\forall r^{(i)\prime}} \Pr(E_\epsilon^{(i)}|R^{(i)} = r^{(i)\prime}) \Pr(R^{(i)} = r^{(i)\prime})$$

$$\begin{aligned}
\Pr(E_\epsilon^{(i)}|R^{(i)} = r^{(i)}) &= 1 - \Pr(\overline{H_\epsilon}, \bigcap_{\rho: r_\rho^{(i)}} \Pr(\overline{G_{\rho,\epsilon}})) \\
&= 1 - \Pr(\overline{H_\epsilon}| \bigcap_{\rho: r_\rho^{(i)}} \Pr(\overline{G_{\rho,\epsilon}})) \Pr(\bigcap_{\rho: r_\rho^{(i)}} \Pr(\overline{G_{\rho,\epsilon}})) \\
&= 1 - (1-\beta) \prod_{\rho: r_\rho^{(i)}, m_{\rho,\epsilon}} (1-\alpha) \\
&= 1 - (1-\beta)(1-\alpha)^{|\{r_\rho^{(i)}: m_{\rho,\epsilon}\}|} \\
&= \Pr(E_\epsilon^{(i)}| \ \ |\{R_\rho^{(i)} : m_{\rho,\epsilon}\}| = |\{r_\rho^{(i)} : m_{\rho,\epsilon}\}|)
\end{aligned}$$

$$\Pr(E_\epsilon) = \sum_{k=0}^{size(\{R^{(i)}: m_{\rho,\epsilon}\})} \Pr(E_\epsilon^{(i)}| \ \ |\{R_\rho^{(i)} : m_{\rho,\epsilon}\}| = k) \Pr(|\{R_\rho^{(i)} : m_{\rho,\epsilon}\}| = k) \quad (1)$$

$$\Pr(E_\epsilon^{(i)}| \ \ |\{R_\rho^{(i)} : m_{\rho,\epsilon}\}| = k) = 1 - (1-\beta)(1-\alpha)^k \quad (2)$$

$$\Pr(|\{R_\rho^{(i)} : m_{\rho,\epsilon}\}| = k) = \binom{size(|\{R^{(i)} : m_{\rho,\epsilon}\}|)}{k} \gamma^k (1-\gamma)^{size(|\{R^{(i)}: m_{\rho,\epsilon}\}|)-k} \quad (3)$$

Let $c^{(i)}$ be a collection of sets of proteins from partition i that have idential connectivity in the graph. Let $m_{\nu,\epsilon}$ indicate $m_{\rho,\epsilon}$ for all $\rho \in c_\nu^{(i)}$. Also, use $\nu(\rho)$ as shorthand for the index of the collection containing protein ρ.

$$\Pr(E_\epsilon^{(i)}|\ \ |\{R_\rho^{(i)} : m_{\rho,\epsilon}\}| = k) \quad = \quad \Pr(E_\epsilon^{(i)}|\ \ \sum_{\nu:m_{\nu,\epsilon}} |\{R_\rho^{(i)} : \rho \in m_{\nu,\epsilon}\}| = k)$$

Define a new random variable N as the sum of the present proteins in each collection, and write the previous probability in terms of this variable:

$$\begin{aligned} N_\nu^{(i)} &= |\{R_\rho^{(i)} : \rho \in c_\nu^{(i)}\}| \\ \Pr(E_\epsilon^{(i)}|\ \ |\{R_\rho^{(i)} : m_{\rho,\epsilon}\}| = k) &= \Pr(E_\epsilon^{(i)}|\ \ \sum_{\nu:m_{\nu,\epsilon}} N_\nu^{(i)} = k) \end{aligned}$$

$$\Pr(E_\epsilon^{(i)}|N^{(i)} = n^{(i)}) = Pr(E_\epsilon^{(i)}|\ \ |\{R_\rho^{(i)} : m_{\rho,\epsilon}\}| = \sum_{\nu:m_{\nu,\epsilon}} N_\nu^{(i)}) \tag{4}$$

Summing over all sets of proteins $R^{(i)}$ is now equivalent to summing over the indistinguishable possible values of $N^{(i)}$ and multiplying each term by the number of possible ways it could be made using protein sets.

$$\begin{aligned} \sum_{\forall r^{(i)}} \Pr(D|R^{(i)} = r^{(i)}) \Pr(R^{(i)} = r^{(i)}) &= \sum_{\forall n^{(i)}} \Pr(D|N^{(i)} = n^{(i)}) \Pr(N^{(i)} = n^{(i)}) \\ \sum_{\forall r^{(i)}} L(R^{(i)} = r^{(i)}|D) \Pr(R^{(i)} = r^{(i)}) &= \sum_{\forall n^{(i)}} L(N^{(i)} = n^{(i)}|D) \Pr(N^{(i)} = n^{(i)}) \end{aligned}$$

$$L(N^{(i)} = n^{(i)}|D) = \prod_\epsilon \sum_{\forall e_\epsilon^{(i)}} \frac{\Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}|D_{\delta(\epsilon)})}{\Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}|Q)} \Pr(E_\epsilon^{(i)} = e_\epsilon^{(i)}|N^{(i)} = n^{(i)}) \tag{5}$$

$$\Pr(N^{(i)} = n^{(i)}) = \prod_\nu \Pr(N_\nu^{(i)} = n_\nu^{(i)}) \tag{6}$$

$$\Pr(N_\nu^{(i)} = n_\nu^{(i)}) = \binom{|c_\nu^{(i)}|}{n_\nu^{(i)}} \gamma^{n_\nu^{(i)}} (1-\gamma)^{|c_\nu^{(i)}| - n_\nu^{(i)}} \tag{7}$$

$$
\begin{aligned}
\Pr(R_\rho^{(i)}|D) &= \sum_{\forall r^{(i)}:r_\rho^{(i)}} \Pr(R^{(i)}=r^{(i)}|D) \\
&= \sum_{\forall n^{(i)}:n_{\nu(\rho)}^{(i)}>0} \Pr(N^{(i)}=n^{(i)},R_\rho^{(i)}|D) \\
\Pr(N^{(i)}=n^{(i)},R_\rho^{(i)}|D) &= \frac{\Pr(D|N^{(i)}=n^{(i)},R_\rho^{(i)})\Pr(N^{(i)}=n^{(i)},R_\rho^{(i)})}{\Pr(D)} \\
&= \frac{\Pr(D|N^{(i)}=n^{(i)})\Pr(N^{(i)}=n^{(i)},R_\rho^{(i)})}{\Pr(D)} \\
&= \frac{\Pr(D|N^{(i)}=n^{(i)})\Pr(R_\rho^{(i)}|N^{(i)}=n^{(i)})\Pr(N^{(i)}=n^{(i)})}{\Pr(D)} \\
&= \Pr(N^{(i)}=n^{(i)}|D)\Pr(R_\rho^{(i)}|N^{(i)}=n^{(i)}) \\
\Pr(N^{(i)}=n^{(i)}|D) &= \frac{\Pr(D|N^{(i)}=n^{(i)})\Pr(N^{(i)}=n^{(i)})}{\sum_{\forall n^{(i)\prime}}\Pr(D|N^{(i)}=n^{(i)\prime})\Pr(N^{(i)}=n^{(i)\prime})}
\end{aligned}
$$

$$
\Pr(N^{(i)}=n^{(i)}|D) = \frac{L(N^{(i)}=n^{(i)}|D)\Pr(N^{(i)}=n^{(i)})}{\sum_{\forall n^{(i)\prime}} L(N^{(i)}=n^{(i)\prime}|D)\Pr(N^{(i)}=n^{(i)\prime})} \tag{8}
$$

The vector of these posterior probabilities can be defined to further save computation by computing all of them in one pass over the set $\{\forall n^{(i)\prime}\}$:

$$
\Pr(R^{(i)}|D) = \sum_{\forall n^{(i)}} \Pr(N^{(i)}=n^{(i)}|D)\Pr(R^{(i)}|N^{(i)}=n^{(i)}) \tag{9}
$$

$$
\tag{10}
$$

Where $\Pr(R^{(i)}|N^{(i)}=n^{(i)})$ is a vector defined as follows:

$$
\Pr(R^{(i)}|N^{(i)}=n^{(i)})_\rho = \frac{n^{(i)}}{|c_{\nu(\rho)}^{(i)}|} \tag{11}
$$

$$
\tag{12}
$$

Note on Using Peptide Probabilities from PeptideProphet

It is important to distinguish peptide probabilities that use protein information from peptide probabilities that do not use protein information (as well as their corresponding priors). PeptideProphet estimates a posterior probabilitie for each peptide without utilizing information in the associations between proteins and peptides. Our model can make use of the PeptideProphetestimates by using them to compute likelihoods that a peptide emitted it's paired spectrum.

By default, this document assumes that protein information is being used, but when protein information is not being used (and instead we are using the PeptideProphet values), then we condition on Q to indicate that protein-level information is not available. For instance $\Pr(E_\epsilon)$ is the prior probability that peptide E_ϵ is present using protein information, whereas $\Pr(E_\epsilon|Q)$ is the prior probability of E_ϵ given by PeptideProphet. This prior probability should take into account the assumed charge state for this PSM. The correct prior values for different charge states are found in the pepXML output of PeptideProphet.
In this document, conversion to probabilities that do not depend on the proteins will be accomplished using the fact that

$$\Pr(D_{\delta(\epsilon)}|E_\epsilon) = \Pr(D_{\delta(\epsilon)}|E_\epsilon, Q)$$

using assumption $A8$.

Increasing the Sparsity of the Graph (Pruning)

In practice, the number of proteins in each partition is often small enough that the algorithm is efficient. However, this is sometimes not the case. A single protein partition can be split by assuming that the peptides that join two subpartitions, $E^{(0)}$, are not present. This assumption introduces no error when these peptides receive a zero score from PeptideProphet. Note that normally ϵ iterates over every index where it is used, however some statements here need to limit to $E^{(0)}$, and so are explicitly referred to as $\epsilon^{(0)}$.
First we demonstrate the error will be zero in this case:

$$\begin{aligned}
|\Pr(R_\rho|D) - \Pr(R_\rho, E^{(0)} = \{\}|D)| &= |\sum_{\forall e^{(0)}} \Pr(R_\rho, E^{(0)} = e^{(0)}|D) - \Pr(R_\rho, E^{(0)} = \{\}|D)| \\
&= \sum_{\forall e^{(0)} \neq \{\}} \Pr(R_\rho, E^{(0)} = e^{(0)}|D) \\
&\leq \sum_{\forall e^{(0)} \neq \{\}} \Pr(E^{(0)} = e^{(0)}|D) \\
&\leq \sum_{\forall \epsilon} \Pr(\overline{E_\epsilon^{(0)}}|D)
\end{aligned}$$

Because, writing $E^{(0)'}$ as the elements other than $E_1^{(0)}$,

$$\sum_{\forall e^{(0)} \neq \{\}} \Pr(E^{(0)} = e^{(0)}|D) = \sum_{\forall e^{(0)'} \neq \{\} \wedge \overline{e_1^{(0)}} \vee e_1^{(0)}} \Pr(E^{(0)'} = e^{(0)'}|D)$$

$$= \Pr(E_1^{(0)}|D) + \sum_{\forall e^{(0)'} \neq \{\}} \Pr(E^{(0)'} = e^{(0)'}, \overline{E_1^{(0)}}|D)$$

$$\leq \Pr(E_1^{(0)}|D) + \sum_{\forall e^{(0)\prime} \neq \{\}} \Pr(E^{(0)\prime} = e^{(0)\prime}|D)$$

and continuing inductively, then

$$\leq \cdots \leq \sum_{\epsilon} \Pr(E_\epsilon^{(0)}|D)$$

Since $\Pr(E_\epsilon^{(0)}|D) = 0$ when $\Pr(E_\epsilon^{(0)}|D_{\delta(\epsilon)}, Q) = 0$ (as long as $\Pr(D) > 0$ which can be shown since $\forall r$, $\Pr(E = e|R = r) > 0$ as long as $\alpha, \beta \in (0, 1)$, meaning that $\Pr(E = e) > 0$, and $\Pr(D|E = e) > 0$ trivially for some $E = e$, and so $\Pr(D) > 0$) then any peptide given a zero score by PeptideProphet can be assumed to be absent. When a peptide is given a very small score by PeptideProphet, then the problem can be approximated more efficiently by assuming the score was zero. In the future, it would be useful to develop a bound on the error introduced by making this approximation, since it would suggest a strategy for which peptides should be changed to zero scored.
Second, we show that assuming $E^{(0)} = \{\}$ gives allows the proteins to be divided into two partitions, $R^{(1)}$ and $R^{(2)}$ with peptides $E^{(1)}$ that associate only with $R^{(1)}$, $E^{(2)}$ that associate only with $R^{(2)}$, and $E^{(0)}$ that associate with both $R^{(1)}$ and $R^{(2)}$, then we have:

$$\Pr(D, E^{(0)} = \{\}|R^{(1)} = r^{(1)}, R^{(2)} = r^{(2)}) =$$
$$\sum_{\forall e^{(1)}, e^{(2)}} \Pr(D^{(1)}|E^{(1)} = e^{(1)}) \Pr(D^{(2)}|E^{(2)} = e^{(2)}) \Pr(D^{(0)}|E^{(0)} = \{\})$$
$$\Pr(E^{(1)} = e^{(1)}|R^{(1)} = r^{(1)}) \Pr(E^{(2)} = e^{(2)}|R^{(2)} = r^{(2)})$$
$$\Pr(E^{(0)} = \{\}|R^{(1)} = r^{(1)}, R^{(2)} = r^{(2)})$$

$$\Pr(E^{(0)} = \{\}|R^{(1)} = r^{(1)}, R^{(2)} = r^{(2)}) =$$
$$\frac{\Pr(E^{(0)} = \{\}|R^{(1)} = r^{(1)}, R^{(2)} = \{\}) \Pr(E^{(0)} = \{\}|R^{(2)} = r^{(2)}, R^{(1)} = \{\})}{\prod_{\epsilon^{(0)}} \Pr(\overline{H_{\epsilon^{(0)}}}| \bigcap_{\rho: r_\rho^{(i)}} \overline{G_{\rho, \epsilon^{(0)}}})}$$

$$\Pr(D, E^{(0)} = \{\}|R^{(1)} = r^{(1)}, R^{(2)} = r^{(2)}) =$$
$$\frac{1}{\prod_{\epsilon^{(0)}} \Pr(\overline{H_{\epsilon^{(0)}}}| \bigcap_{\rho: r_\rho^{(i)}} \overline{G_{\rho, \epsilon^{(0)}}})}$$
$$\sum_{\forall e^{(1)}} \Pr(D^{(1)}|E^{(1)} = e^{(1)}) \Pr(D^{(0)}|E^{(0)} = \{\})$$
$$\Pr(E^{(2)} = e^{(2)}|R^{(2)} = r^{(2)}) \Pr(E^{(0)} = \{\}|R^{(2)} = r^{(2)}, R^{(2)} = \{\})$$
$$\sum_{\forall e^{(2)}} \Pr(D^{(2)}|E^{(2)} = e^{(2)})$$
$$\Pr(E^{(2)} = e^{(2)}|R^{(2)} = r^{(2)}) \Pr(E^{(0)} = \{\}|R^{(2)} = r^{(2)}, R^{(1)} = \{\})$$

$$=$$

$$\frac{1}{\Pr(D^{(0)}|E^{(0)}=\{\})\prod_{\epsilon^{(0)}}\Pr(\overline{H_{\epsilon^{(0)}}}|\bigcap_{\rho:r_\rho^{(i)}}\overline{G_{\rho,\epsilon^{(0)}}})}$$

$$\sum_{\forall e^{(1)}}\Pr(D^{(1)}|E^{(1)}=e^{(1)})\Pr(D^{(0)}|E^{(0)}=\{\})$$

$$\Pr(E^{(2)}=e^{(2)}|R^{(2)}=r^{(2)})\Pr(E^{(0)}=\{\}|R^{(2)}=r^{(2)},R^{(2)}=\{\})$$

$$\sum_{\forall e^{(2)}}\Pr(D^{(2)}|E^{(2)}=e^{(2)})\Pr(D^{(0)}|E^{(0)}=\{\})$$

$$\Pr(E^{(2)}=e^{(2)}|R^{(2)}=r^{(2)})\Pr(E^{(0)}=\{\}|R^{(2)}=r^{(2)},R^{(1)}=\{\})$$

$$=$$

$$\frac{1}{\Pr(D^{(0)}|E^{(0)}=\{\})\prod_{\epsilon^{(0)}}\Pr(\overline{H_{\epsilon^{(0)}}}|\bigcap_{\rho:r_\rho^{(i)}}\overline{G_{\rho,\epsilon^{(0)}}})}$$

$$\Pr(D^{(1)},D^{(0)},E^{(0)}=\{\}|R^{(1)}=r^{(1)},R^{(2)}=\{\})$$

$$\Pr(D^{(2)},D^{(0)},E^{(0)}=\{\}|R^{(2)}=r^{(2)},R^{(1)}=\{\})$$

Finally, we can define the likelihood

$$L(R^{(1)},R^{(2)}|E^{(0)}=\{\},D)=$$

$$\frac{1}{\prod_{\epsilon^{(0)}}\Pr(\overline{H_{\epsilon^{(0)}}}|\bigcap_{\rho:r_\rho^{(i)}}\overline{G_{\rho,\epsilon^{(0)}}})}$$

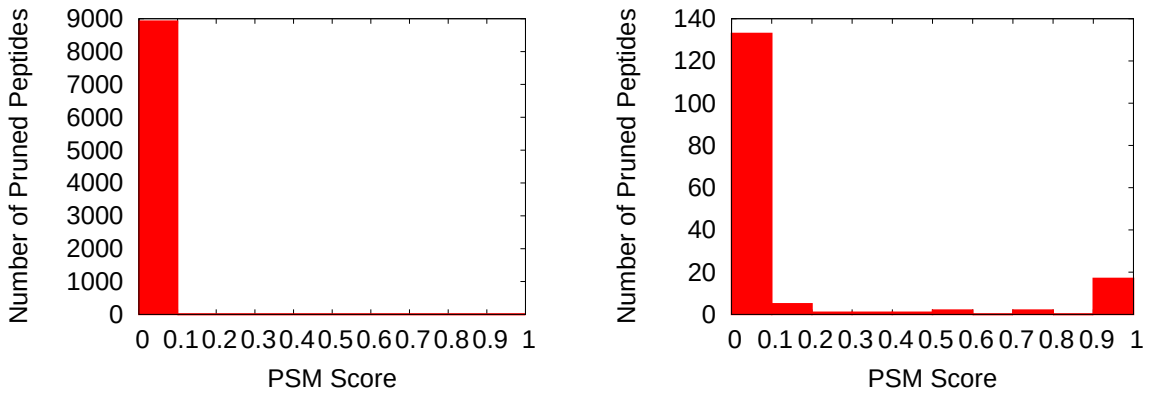
$$L(R^{(1)}=r^{(1)},R^{(2)}=\{\}|E^{(0)}=\{\},D^{(1)},D^{(0)})$$

$$L(R^{(2)}=r^{(2)},R^{(1)}=\{\}|E^{(0)}=\{\},D^{(2)},D^{(0)}) \tag{13}$$

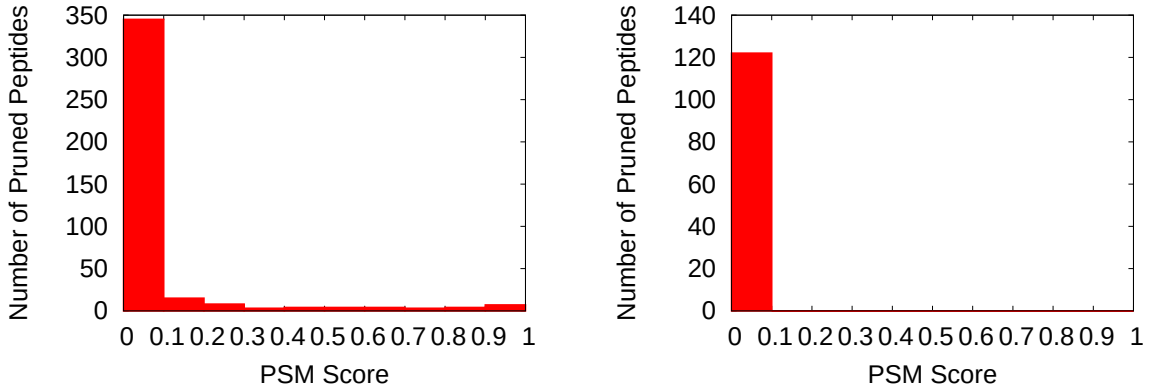
where the likelihood constant does not depend on R, α, β.
The resulting probability and likelihood is equivalent (aside from the leading correction in the previous formula) to those computed after disconnecting $R^{(1)}$ from $R^{(2)}$ by duplicating the peptides that they share so that each $R^{(1)}$ and $R^{(2)}$ associate with their own copy of $E^{(0)}$ and $D^{(0)}$.

Analysis of Approximation Errors from Pruning

The calculation of posterior probabilities is exact when only zero-scoring PSMs are used for pruning. When PSMs with small nonzero scores are used, then a small approximation error is introduced. Occassionaly, it may be necessary to prune a PSM with a large score in order to achieve the desired separability. Even in this case, the error introduced may be small and is confined to the proteins in the same connected subgraph; therefore, the overall approximation error depends not only on the highest-scoring PSM pruned, but on



(A) *H. influenzae* (B) Yeast



(A) ISB 18 (B) Sigma 49

Figure 1: **Distribution of pruned PSM scores.** For each data set, we plot a histogram of the scores of pruned PSMs. The figure shows that almost all pruned PSMs have very low scores. The *C.* elegans data set is not shown because it does not require pruning to achieve this much separability.

the quantity of PSMs pruned and their scores. In Figure 1, we plot the distribution of pruned PSM scores necessary to require no more than 2^{18} marginalization steps for any subgraph. We demonstrate that there are not many high-scoring PSMs that need to be pruned, even to achieve this strict level of separability. The yeast data requires a couple of high-scoring PSMs to be pruned, but the proteins associated with these PSMs are also suppported by other PSMs; therefore, the error is still minimal and only influences these few proteins.