# A Combinatorial Perspective of the Protein Inference Problem

Chao Yang, Zengyou He and Weichuan Yu

June 18, 2013

## 1 Conditional Probability Based on A Loose Assumption

Given the conditional probability:

$$\Pr(y = 0|X_{\mathcal{G}}, X_{\bar{\mathcal{G}}}) = 0. \tag{1}$$

Here, $\mathcal{G}$ is not empty. Show that:

$$\Pr(y = 0|S) = \prod_{i=1}^{M} (1 - \Pr(x_i = 1|S)). \tag{2}$$

**Proof**:
The absent probability is calculated as:

$$\begin{aligned} &\Pr(y = 0|S) = \prod_{i=1}^{M} \Pr(x_i = 0|S) + \sum_{X_{\mathcal{G}}, X_{\bar{\mathcal{G}}}, \mathcal{G} \neq \varnothing} \Pr(y = 0|X_{\mathcal{G}}, X_{\bar{\mathcal{G}}}) Pr(X_{\mathcal{G}}|S) Pr(X_{\bar{\mathcal{G}}}|S) \\ &= \prod_{i=1}^{M} \Pr(x_i = 0|S) \\ &= \prod_{i=1}^{M} (1 - \Pr(x_i = 1|S)). \end{aligned} \tag{3}$$

## 2 Conditional Probability Based on A Strict Assumption

Given the conditional probability:

$$\Pr(y = 0|X_{\mathcal{G}}, X_{\bar{\mathcal{G}}}) = \frac{N_a}{N_t} = \prod_{i \in \mathcal{G}} \frac{(n_i - 1)}{n_i}. \tag{4}$$

Here, $\mathcal{G}$ is not empty. Prove that:

$$\begin{aligned}&\Pr(y=0|S)=\sum_{X_{\mathcal{G}},X_{\bar{\mathcal{G}}}}\Pr(y=0|X_{\mathcal{G}},X_{\bar{\mathcal{G}}})\Pr(X_{\mathcal{G}}|S)\Pr(X_{\bar{\mathcal{G}}}|S)\\=&\prod_{i=1}^{M}(1-\frac{1}{n_i}\Pr(x_i=1|S)).\end{aligned}\tag{5}$$

**Proof**:

$$\begin{aligned}&\prod_{i=1}^{M}(1-\frac{1}{n_i}\Pr(x_i=1|S))\\=&\prod_{i=1}^{M}(1-\Pr(x_i=1|S)+\frac{(n_i-1)}{n_i}\Pr(x_i=1|S))\\=&\prod_{i=1}^{M}(\Pr(x_i=0|S)+\frac{(n_i-1)}{n_i}\Pr(x_i=1|S))\\=&\sum_{\mathcal{G},\bar{\mathcal{G}}}\prod_{j\in\bar{\mathcal{G}}}\Pr(x_j|S)\prod_{i\in\mathcal{G}}\left(\frac{(n_i-1)}{n_i}\Pr(x_i|S)\right)\\=&\sum_{X_{\mathcal{G}},X_{\bar{\mathcal{G}}}}\prod_{i\in\mathcal{G}}\frac{(n_i-1)}{n_i}\Pr(X_{\mathcal{G}}|S)\Pr(X_{\bar{\mathcal{G}}}|S)\\=&\sum_{X_{\mathcal{G}},X_{\bar{\mathcal{G}}}}\Pr(y=0|X_{\mathcal{G}},X_{\bar{\mathcal{G}}})\Pr(X_{\mathcal{G}}|S)\Pr(X_{\bar{\mathcal{G}}}|S).\end{aligned}\tag{6}$$

## 3 Conditional Probability Based on A Mild Assumption

Given the conditional probability:

$$\Pr(y=0|X_{\mathcal{G}},X_{\bar{\mathcal{G}}})=\frac{N_a}{N_t}=\prod_{i\in\mathcal{G}}\frac{2^{(n_i-1)}-1}{2^{n_i}-1}.\tag{7}$$

Here, $\mathcal{G}$ is not empty. Prove that:

$$\begin{aligned}&\Pr(y=0|S)=\sum_{X_{\mathcal{G}},X_{\bar{\mathcal{G}}}}\Pr(y=0|X_{\mathcal{G}},X_{\bar{\mathcal{G}}})\Pr(X_{\mathcal{G}}|S)\Pr(X_{\bar{\mathcal{G}}}|S)\\=&\prod_{i=1}^{M}(1-\frac{2^{n_i}}{2(2^{n_i}-1)}\Pr(x_i=1|S)).\end{aligned}\tag{8}$$

**Proof**:

$$
\begin{aligned}
&\prod_{i=1}^{M}(1-\frac{2^{n_i}}{2(2^{n_i}-1)}\Pr(x_i=1|S)) \\
=&\prod_{i=1}^{M}(1-\Pr(x_i=1|S)+\frac{(2^{(n_i-1)}-1)}{(2^{n_i}-1)}\Pr(x_i=1|S)) \\
=&\prod_{i=1}^{M}(\Pr(x_i=0|S)+\frac{(2^{(n_i-1)}-1)}{(2^{n_i}-1)}\Pr(x_i=1|S)) \\
=&\sum_{\mathcal{G},\bar{\mathcal{G}}}\prod_{j\in\bar{\mathcal{G}}}\Pr(x_j|S)\prod_{i\in\mathcal{G}}\left(\frac{(2^{(n_i-1)}-1)}{(2^{n_i}-1)}\Pr(x_i|S)\right) \\
=&\sum_{X_{\mathcal{G}},X_{\bar{\mathcal{G}}}}\prod_{i\in\mathcal{G}}\frac{2^{(n_i-1)}-1}{2^{n_i}-1}\Pr(X_{\mathcal{G}}|S)\Pr(X_{\bar{\mathcal{G}}}|S) \\
=&\sum_{X_{\mathcal{G}},X_{\bar{\mathcal{G}}}}\Pr(y=0|X_{\mathcal{G}},X_{\bar{\mathcal{G}}})\Pr(X_{\mathcal{G}}|S)\Pr(X_{\bar{\mathcal{G}}}|S).
\end{aligned}
\tag{9}
$$

# 4 Marginal Protein Probability

Prove the inequality:

$$
\begin{aligned}
&\Pr_L(y=1|S)=1-\prod_{i=1}^{M}(1-\frac{1}{n_i}\Pr(x_i=1|S)) \\
\leq&\Pr_E(y=1|S)=1-\prod_{i=1}^{M}(1-\frac{2^{n_i}}{2(2^{n_i}-1)}\Pr(x_i=1|S)) \\
\leq&\Pr_U(y=1|S)=1-\prod_{i=1}^{M}(1-\Pr(x_i=1|S)).
\end{aligned}
\tag{10}
$$

Here, $n_i \geq 1$ is the number of times that peptide $i$ is shared.

**Proof**:
To prove the inequality (10), we only have to show that:

$$
\frac{1}{n_i} \leq \frac{2^{n_i}}{2(2^{n_i}-1)} \leq 1. \tag{11}
$$

**Case 1**: When $n_i = 1$, we have:

$$
\frac{1}{n_i} = \frac{2^{n_i}}{2(2^{n_i}-1)} = 1. \tag{12}
$$

**Case 2**: When $n_i \geq 2$, we have:

$$2 \cdot 2^{n_i} - 2 \leq n_i \cdot 2^{n_i} - 2 < n_i \cdot 2^{n_i}. \tag{13}$$

Thus, we have $\frac{1}{n_i} < \frac{2^{n_i}}{2(2^{n_i}-1)}$.
Since

$$\frac{2^{n_i}}{2(2^{n_i}-1)} = \frac{1}{2} + \frac{1}{2(2^{n_i}-1)} \tag{14}$$

is monotonically decreasing with respect to $n_i$, we have:

$$\frac{2^{n_i}}{2(2^{n_i}-1)} \leq \frac{2^{n_i}}{2(2^{n_i}-1)}\Big|_{n_i=2} = \frac{2}{3} < 1. \tag{15}$$

In conclusion:

$$\frac{1}{n_i} \leq \frac{2^{n_i}}{2(2^{n_i}-1)} \leq 1. \tag{16}$$

The equality is obtained when $n_i = 1$.

# 5 Running Time

In Table 1, we show the running time of our method, ProteinProphet, Fido and the greedy method. The comparison is conducted on a computer with 4GB memory and Intel(R) Core(TM) i5-2500 CPU running the 32bit Windows 7 operating system. The final result shown in the table is the average running time of ten runs. The total running time includes loading the peptide identification result, estimating protein probabilities and reporting the result. The comparison of running time indicates the efficiency of our method in calculating protein probabilities.

Table 1: Running time of ProteinInfer, ProteinProphet, Fido and the greedy method.

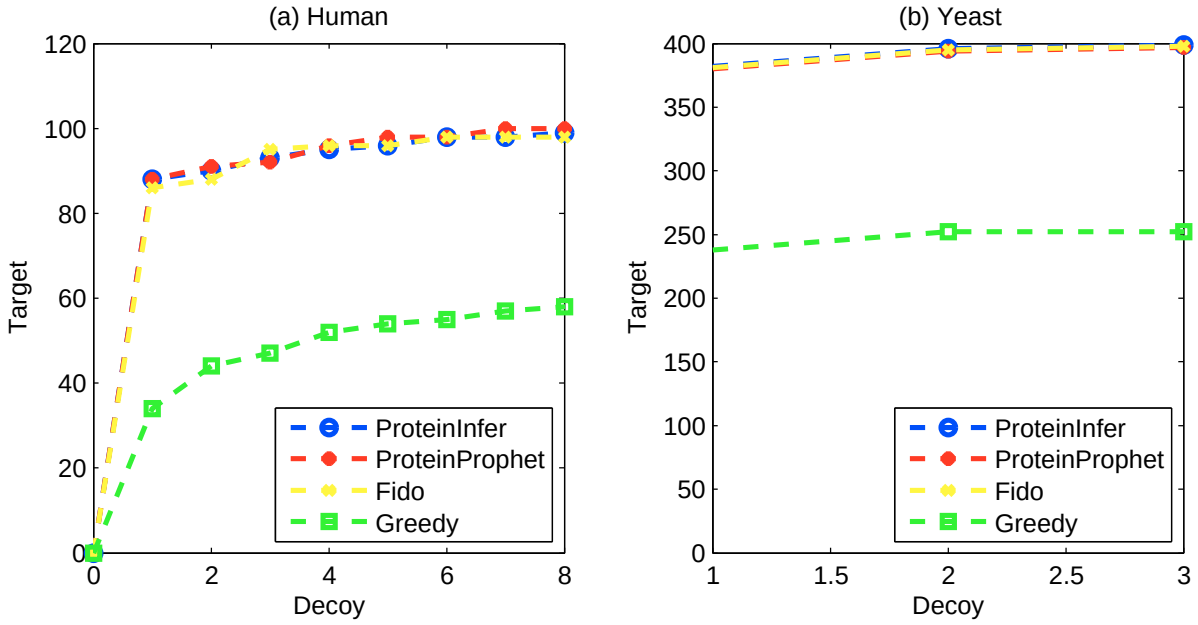
| Program | ISB | Sigma49 | Human | Yeast |
|---|---|---|---|---|
| ProteinInfer | 0.340s | 0.478s | 1.057s | 0.920s |
| ProteinProphet | 14.273s | 15.103s | 16.473s | 14.710s |
| Fido | 4.556s | 4.406s | 1.082s | 7.977s |
| Greedy | 0.650s | 1.671s | 6.845s | 1.631s |

The total running time includes loading the peptide identification results, estimating protein probabilities and reporting the final results.

# 6 The impacts of protein databases on protein identification results

In our experiment, we use the UniProtKB/Swiss-Prot protein database, which is a general purpose database containing sequences of different species. In real applications, we may choose

Figure 1: The performance of ProteinInfer, ProteinProphet, Fido and the greedy method on the Humand dataset and the Yeast dataset. We use organism specific databases drawing from the UniProtKB/Swiss-Prot database in this experiment. The curves of decoy versus target are used to measure the performance.




to identify proteins from organism specific database to improve the efficiency of protein identification. In this section, we study the impact of protein database size on the performance of different protein identification methods.

We conduct experiment on the Humand dataset and the Yeast dataset. We create organism-specific databases by drawing Human proteins and Yeast proteins from the UniProtKB/Swiss-Prot database for two datasets, respectively. Figure 1 shows the performance of ProteinInfer, ProteinProphet, Fido and the greedy method on two datasets. The curves of decoy versus target are used to measure the performance. Based on the experimental results, we can see that ProteinInfer, ProteinProphet and Fido performs similarly when the protein database only contains organism specific proteins.

# 7 Adjustment of Probabilities for Unique Peptides

## 7.1 Method

Protein inference models take peptide identification results as input. If the peptide identification results are perfect, peptide probability adjustment is not essential. However, inferior peptide probabilities always exist.

Unique peptides are important to protein identification. A confident misidentified unique peptide (i.e. $\Pr(x = 1|S) = 0.99$) will result in a high confident protein identification with a high $\text{Pr}_E$ and a low $\text{Pr}_D$. For example, if a tandem mass spectrum is matched to a decoy peptide, the peptide is very likely to be unique. The unique high confident decoy peptide will lead the decoy protein to be identified with a high confidence. This motivates us to make the adjustment of probabilities for unique peptides a preprocessing step of our method.

Suppose a protein has $m$ unique peptides. The adjusted unique peptide probability can be calculated as:

$$\Pr(x_i = 1|m, S) = \frac{\Pr(m|x_i = 1, S)\Pr(x_i = 1|S)}{\Pr(m|x_i = 1, S)\Pr(x_i = 1|S) + \Pr(m|x_i = 0, S)\Pr(x_i = 0|S)}. \tag{17}$$

Here, peptide $i$ is a unique peptide; $\Pr(x_i = 1|S)$ is the probability that the unique peptide is true. The terms $\Pr(m|x_i = 1, S)$ and $\Pr(m|x_i = 0, S)$ describe the probabilities of observing $m$ unique peptides of the protein when the unique peptide $i$ is a true and a false identification, respectively. We model $\Pr(m|x_i = 1, S)$ and $\Pr(m|x_i = 0, S)$ as Poisson distributions with different expected numbers of unique peptides (i.e. $\lambda_1$ and $\lambda_2$):

$$\begin{aligned} \Pr(m|x_i = 1, S) &= \frac{\lambda_1^m e^{-\lambda_1}}{m!} \\ \Pr(m|x_i = 0, S) &= \frac{\lambda_2^m e^{-\lambda_2}}{m!} \end{aligned}. \tag{18}$$

Generally, a true unique peptide tends to have more sibling unique peptides than a false unique peptide on average. Thus, we have $\lambda_1 > \lambda_2$. In our program, these two parameters can be manually specified. Alternatively, these two parameters can be obtained empirically.

Suppose there are $N$ candidate proteins and the number of unique peptides of protein $j (j \in \{1, 2, ..., N\})$ is $m_j$. The empirical value of the expected value $\lambda_1$ is estimated as:

$$\lambda_1 = \frac{\sum_{j=1}^{N} I(m_j \geq 2) m_j}{\sum_{j=1}^{N} I(m_j \geq 2)}. \tag{19}$$

Here, $I(\cdot)$ is an indicator function with value being either 0 or 1.

Empirically, $\lambda_2$ can be 1. It is common to observe that false proteins such as decoy proteins to have a single unique peptide.

The adjusted unique peptide probability $\Pr(x_i = 1|m, S)$ is used as the probability of peptide $i$ in our model (10).

### 7.2 The Parameter Issue

In the preprocessing step, there are two parameters $\lambda_1$ and $\lambda_2$ corresponding to the expected number of unique peptides of true proteins and false proteins, respectively. These two parameters are estimated empirically from data. Alternatively, these two parameters can be set manually.

Here, the performances of our method with different parameter settings are compared to show whether our method is sensitive to the parameter setting.

In this section, we conduct our experiment on the ISB dataset and the Sigma49 dataset. These two datasets have groundtruth, which can reflect the impacts of parameter settings accurately.

The empirical estimations of $\lambda_1$ for the ISB dataset and the Sigma49 dataset are 12 and 7, respectively. The parameter settings we consider are shown in Table 2. Under each parameter setting, we use the curve of false positives versus true positives to measure the corresponding performance of our method. The curve obtained by using the empirical parameters is taken as a reference. We calculate the correlation of other curves with the reference to illustrate the performance variation in different conditions. Figure 3 shows the results.

Table 2: Parameter settings

| Index | ISB | Sigma49 |
|---|---|---|
| 1 | $\lambda_1 = 5$ | $\lambda_1 = 2$ |
| | $\lambda_2 = 1$ | $\lambda_2 = 1$ |
| 2 | $\lambda_1 = 15$ | $\lambda_1 = 9$ |
| | $\lambda_2 = 1$ | $\lambda_2 = 1$ |
| 3 | $\lambda_1 = 10$ | $\lambda_1 = 5$ |
| | $\lambda_2 = 1$ | $\lambda_2 = 1$ |
| 4 | $\lambda_1 = 12$ | $\lambda_1 = 7$ |
| | $\lambda_2 = 5$ | $\lambda_2 = 3$ |
| 5 | $\lambda_1 = 12$ | $\lambda_1 = 7$ |
| | $\lambda_2 = 10$ | $\lambda_2 = 5$ |

In the experiment on each dataset, the performance based on the default parameter setting is shown for reference.

In our model, we require that $\lambda_1 > \lambda_2$. We also conduct experiments to show what the result is when parameters are misspecified (i.e. $\lambda_1 < \lambda_2$). The result is shown in Figure 2.

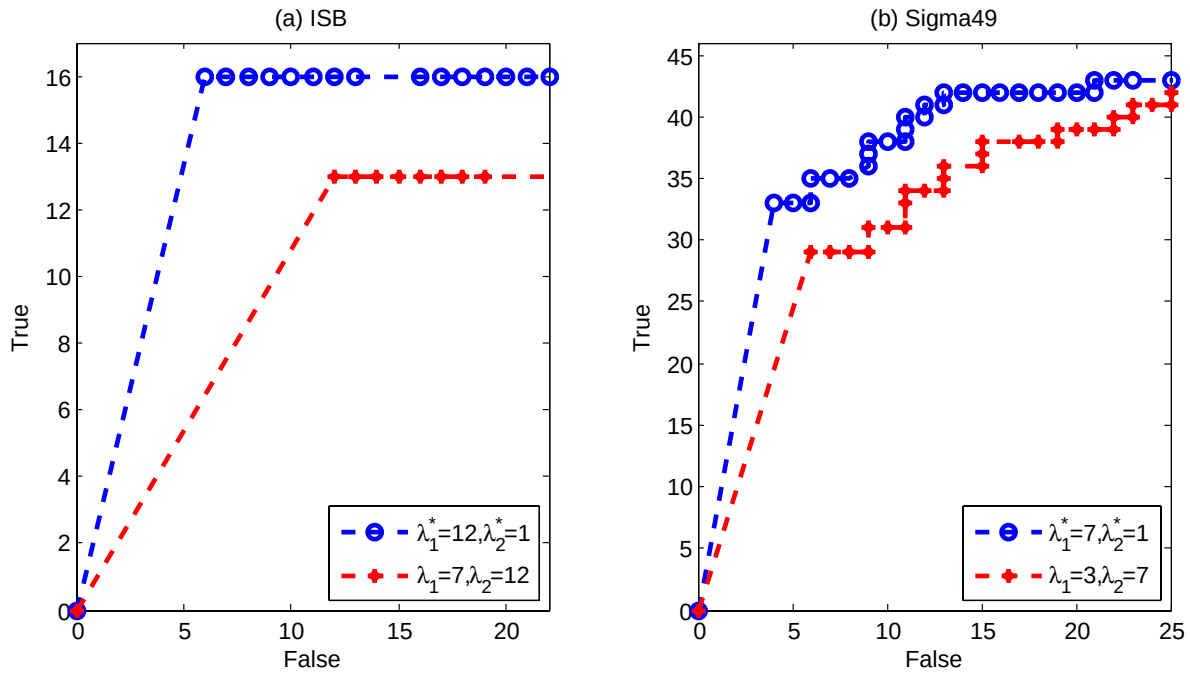



Figure 2: The performances of our method when parameters are set as $\lambda_1 < \lambda_2$. Default empirical parameter settings are marked with "*".

According to the results in two figures, we can see that our method is not sensitive to the parameter setting. The wrong parameter setting has a great impact on the protein identification result. Thus, we do not allow $\lambda_1 < \lambda_2$ in our program.

### 7.3 More on Peptide Probability Adjustment

From the previous experiment, we can see that our method is not sensitive to the parameter setting. The only requirement is that parameters $\lambda_1 > \lambda_2$. According to Figure 3, the performances under different parameters are close. Readers may be interested in the benefit of peptide probability adjustment.

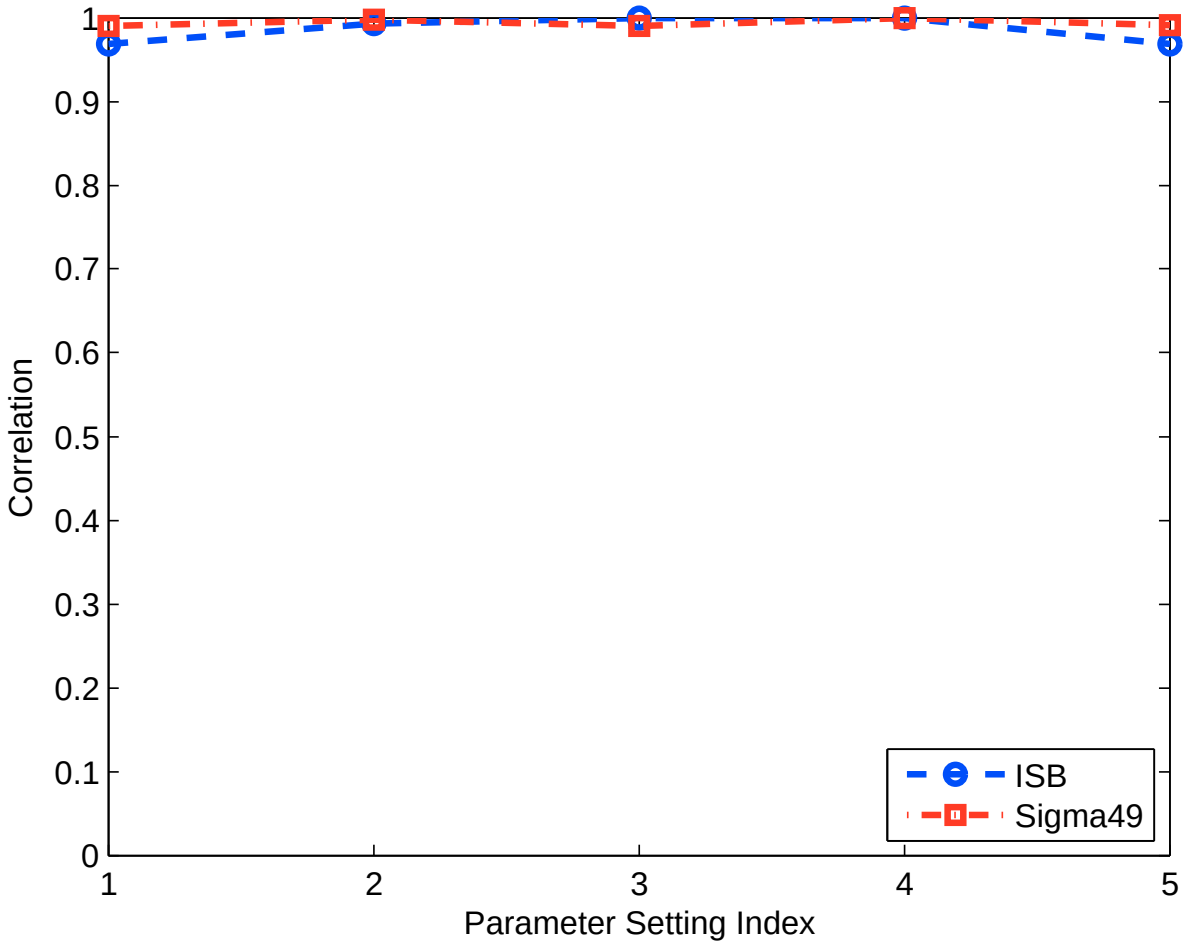



Figure 3: The performances of our method on the ISB dataset and the Sigma49 dataset under different parameter settings shown in Table 2.

Let us consider the top three decoy proteins in the experiments on the ISB dataset and the Sigma49 dataset for the illustration purpose. Table 3 shows the protein probabilities of decoy proteins before and after peptide probability adjustment. According to the result, the protein probabilities of decoy proteins are decreased and they are ordered behind more target proteins after the probability adjustment. Decoy proteins are representative of a kind of error in protein identification. It is not desired to detect a decoy protein with a high confidence (e.g. decoy_499748 is detected with $\Pr_E = 0.9971$ and $\Pr_D = 0.0000$). In this sense, the protein identification result becomes more meaningful after unique peptide adjustment. High confident decoy proteins are detected because its corresponding decoy peptides are detected with high confidence. Since peptide probability calculation is not perfect, we need to adjust peptide probabilities to achieve a more meaningful protein identification result.

Table 3: The protein probabilities of decoy proteins before and after peptide probability adjustment

| Probabilities Without Adjustment | | | | | |
|---|---|---|---|---|---|
| Index | Dataset | Protein | $\Pr_E$ | $\Pr_D$ | Rank |
| 1 | ISB | decoy_499748 | 0.9971 | 0.0000 | 46 |
| 2 | ISB | decoy_237394 | 0.9243 | 0.0000 | 54 |
| 3 | ISB | decoy_224201 | 0.8864 | 0.0000 | 55 |
| 4 | Sigma49 | decoy_519997 | 0.9517 | 0.0000 | 62 |
| 5 | Sigma49 | decoy_170817 | 0.9417 | 0.0000 | 64 |
| 6 | Sigma49 | decoy_271930 | 0.8248 | 0.0000 | 74 |
| Probabilities With Adjustment | | | | | |
| Index | Dataset | Protein | $\Pr_E$ | $\Pr_D$ | Rank |
| 1 | ISB | decoy_499748 | 0.0650 | 0.0000 | 50 |
| 2 | ISB | decoy_237394 | 0.0024 | 0.0000 | 54 |
| 3 | ISB | decoy_224201 | 0.0016 | 0.0000 | 55 |
| 4 | Sigma49 | decoy_519997 | 0.2548 | 0.0000 | 72 |
| 5 | Sigma49 | decoy_170817 | 0.2190 | 0.0000 | 73 |
| 6 | Sigma49 | decoy_271930 | 0.0755 | 0.0000 | 77 |

The last column is the rank position of the protein in the corresponding result. In the table, $\Pr_D$ are 0.0000 because peptides of decoy proteins tend to be unique. According to our inequality (20) in the paper, $\Pr_L = \Pr_U$ when all peptides are unique.

In conclusion, unique peptide probability adjustment can improve the protein identification result (i.e. decoys proteins are ranked behind more target proteins) and make the result more meaningful than the result before adjustment. Thus, keeping the adjustment procedure in our program is essential.

# 8 The Protein Identification Result

There are three different kinds of relationships between two proteins:

- Indistinguishable: If two proteins contain exactly the same set of identified peptides, they are indistinguishable. Indistinguishable proteins can be treated as a group.
- Subset: Identified peptides of a protein form a peptide set. If a peptide set of one protein is the subset of another protein, the former protein is a subset protein of the latter.
- Differentiable: Two proteins are differentiable if they both contain different peptides.

In the literature, subset proteins are generally discarded. However, it is not reasonable to regard all subset proteins as being absent from a sample. Thus, we also calculate the protein

Table 4: Subset and Non-subset Proteins. Proteins 1 and 3 are indistinguishable proteins and protein 2 is a subset protein of protein 1. $\Pr_E$, $\Pr_L$ and $\Pr_U$ denote the empirical estimation, the lower bound and the upper bound of protein probabilities, respectively. $\Pr_D = \Pr_U - \Pr_L$ is the probability confidence interval.

| Non-subset Proteins | | | | | | |
|---|---|---|---|---|---|---|
| Index | Protein | $\Pr_E$ | $\Pr_L$ | $\Pr_U$ | $\Pr_D$ | Other Proteins |
| 1 | 1 | $\Pr_E(y_1\|S)$ | $\Pr_L(y_1\|S)$ | $\Pr_U(y_1\|S)$ | $\Pr_D(y_1\|S)$ | 3 |
| 2 | 4 | $\Pr_E(y_4\|S)$ | $\Pr_L(y_4\|S)$ | $\Pr_U(y_4\|S)$ | $\Pr_D(y_4\|S)$ | - |
| ..... | | | | | | |
| Subset Proteins | | | | | | |
| Index | Protein | $\Pr_E$ | $\Pr_L$ | $\Pr_U$ | $\Pr_D$ | Subset of Protein |
| 1 | 2 | $\Pr_E(y_2\|S)$ | $\Pr_L(y_2\|S)$ | $\Pr_U(y_2\|S)$ | $\Pr_D(y_2\|S)$ | 1 |
| ..... | | | | | | |

probability of subset proteins and organize our result in two separate files as shown in Table 4. In the table, proteins 1 and 3 are indistinguishable proteins and protein 2 is a subset protein of protein 1.

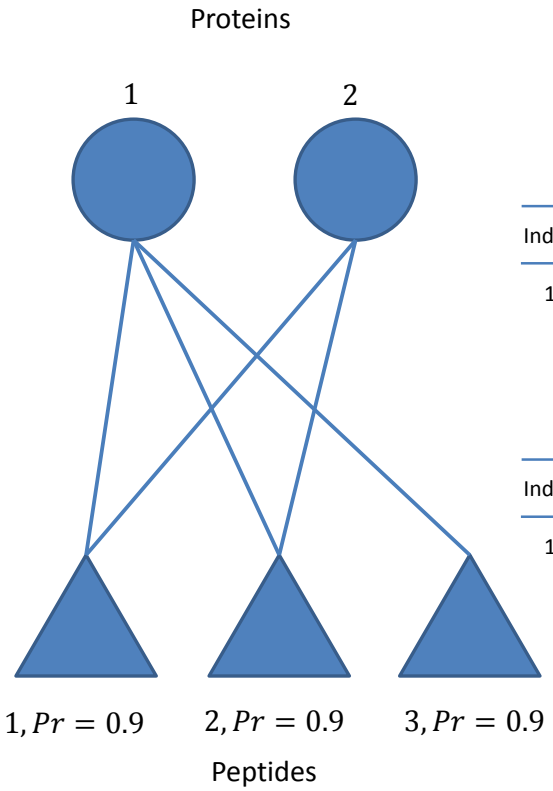
The empirical probability $\Pr_E$ and the bound $\Pr_D$ quantitatively describe the confidence of a protein. Generally, a high $\Pr_E$ and a low $\Pr_D$ mean a confident protein. The identification result is mainly sorted by the empirical protein probability $\Pr_E$. In case two proteins have the same $\Pr_E$, the order is then determined by $\Pr_D$.

One purpose of protein identification is to select proteins to explain observed peptides. Subset proteins do not increase the peptide explanation power. In general, we can perform downstream analysis by only using non-subset proteins. However, according to the probability theorem, the probability of a subset protein is not necessarily smaller than that of a non-subset protein. The data explanation power and protein probability are totally two different kinds of things. Without any prior knowledge, choosing proteins from data explanation's viewpoint is safe. However, in some cases, we may consider high confident subset proteins according to the prior knowledge we have. For instance, the sample contains homogeneous proteins and it is possible that subset proteins are present.

An example is shown in Figure 4 for the illustration purpose. Protein 1 is more confident than protein 2. From data explanation's viewpoint, protein 1 is present whereas protein 2 is absent. This is because including protein 2 in the final protein list will not improve the data explanation power. When we know that homogeneous proteins are present and desire more proteins, we can merge subset and non-subset proteins to obtain the final result by filtering proteins with thresholds on $\Pr_E$ and $\Pr_D$.

In ProteinProphet, the probabilities of subset proteins are zero. In our experiment, only non-subset proteins are considered to perform a fair comparison.

Figure 4: An example of the protein identification result. In the figure, there are two proteins and three peptides with corresponding probabilities all being 0.9. Protein 2 is a subset protein of protein 1. The probability $\mathrm{Pr}_E$ and the bound $\mathrm{Pr}_D$ are two quantitative measurements of the confidence of a protein. The higher the value of $\mathrm{Pr}_E$ and the smaller the value of $\mathrm{Pr}_D$, the more confident the protein. Protein 1 is a confident protein with a high empirical probability $\mathrm{Pr}_E = 0.984$ and a tight bound $\mathrm{Pr}_D = 0.029$. For protein 2, there are two peptides present. Without any prior knowledge, we cannot determine the presence of protein 2 mathematically. From the data explanation's aspect, we can report protein 1 only. Protein 2 can be considered if the protein coverage is a concern and homogeneous proteins are known to be present.




Non-subset proteins

| Index | Protein | $Pr_E$ | $Pr_L$ | $Pr_U$ | $Pr_D$ | Other Proteins |
|---|---|---|---|---|---|---|
| 1 | 1 | 0.984 | 0.970 | 0.999 | 0.029 | - |

Subset proteins

| Index | Protein | $Pr_E$ | $Pr_L$ | $Pr_U$ | $Pr_D$ | Subset of Protein |
|---|---|---|---|---|---|---|
| 1 | 2 | 0.840 | 0.698 | 0.990 | 0.292 | 1 |

## 9 Protein Probability Interval

Protein probability interval $\Pr_D$ can be used to improve the distinction of protein identification results as well as to filter protein identification results.

Table 5: The number of indistinguishable proteins based on probabilities without and with $\Pr_D$

| Dataset | Without $\Pr_D$ | With $\Pr_D$ |
|---|---|---|
| ISB | 27 | 21 |
| Sigma49 | 38 | 36 |
| Human | 60 | 58 |
| Yeast | 181 | 178 |

Table 5 shows the numbers of indistinguishable proteins (based on the protein probability) without and with $\Pr_D$ on four datasets. From the result, we can see that $\Pr_D$ decreases the number of indistinguishable proteins.

Table 6: The number of subset proteins without and with $\Pr_D$

| Dataset | $\Pr_E \geq 0.9$ | $\Pr_E \geq 0.9 \&\& \Pr_D \leq 0.02$ |
|---|---|---|
| ISB | 325 | 15 |
| Sigma49 | 110 | 4 |
| Human | 12 | 4 |
| Yeast | 126 | 0 |

In the table, "&&" means logical "AND".

More importantly, $\Pr_D$ can be used as an extra filtering standard when $\Pr_E$ alone does not work effectively. This is very useful in the case when subset proteins are considered (e.g. protein identification rate is not satisfactory and homogeneous proteins are known to be present). Table 6 shows the numbers of subset proteins when using $\Pr_E \geq 0.9$ and $\Pr_E \geq 0.9 \&\& \Pr_D \leq 0.02$ as filters, respectively. Here, "&&" means logical "AND". A great number of subset proteins can be filtered out by using $\Pr_D$. Thus, $\Pr_D$ and $\Pr_E$ form an effective filter to pick confident proteins.