Introduction

Given the peptides identified by shotgun proteomics of a biological sample, protein inference assigns which proteins are present in the sample. Compared with peptide identification, protein inference is more biologically relevant, and it represents a major challenge in shotgun proteomics. Yet, not enough work has been done to address the protein inference problem. Here, we present a Bayesian approach which incorporates peptide detectability to solve the protein inference problem. Our algorithm achieves high accuracy, and it can effectively assign proteins with degenerate detected peptides.

Protein configuration graph: a connected component from real data

The graphs show the connected component with proteins and tryptic peptides as nodes, and protein-peptide belonging relationship as edges. Peptides with <=6 amino acids are grey colored, the other peptides and proteins are golden colored.

Left: Purple squares – 18 peptides identified by a shotgun proteomics experiment
Right: Purple circles – the 47 candidate proteins, each of which contains at least one identified peptides.

Purple squares – the 204 tryptic peptides of these candidate proteins

Summary of Bayesian model & inference algorithm

Assumptions
- Protein prior probabilities are independent
- Peptide identifications are conditionally independent
- Detectability as the prior probability for peptide detection
- Protein quantities influence the peptide detectabilities

Input
- Basic model: confident peptide identifications (Φ or Ψ)
- Advanced model: all identified peptides with various scores

Output
- Protein identifications: Maxima A Posterior (MAP) solution
- Confidence of identifications: marginal posterior probabilities

BayesInfer 1.0 Algorithm

- Protein configuration graph is constructed from protein database
- Candidate proteins are grouped into connected components
- Protein quantity is estimated by maximal likelihood approach
- Detectability is adjusted by quantity
- Protein prior probability is estimated empirically
- Peptide identification probability is adjusted by detectability
- Gibbs sampling is combined with maximum-product algorithm
- Blocking and random updating schema are employed in Gibbs sampling

Performance comparison: previous methods and variants of BayesInfer

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMP</td>
<td>0.867</td>
<td>0.943</td>
<td>0.892</td>
</tr>
<tr>
<td>ProteinProphet</td>
<td>0.877</td>
<td>0.997</td>
<td>0.955</td>
</tr>
<tr>
<td>Basic Bayesian model</td>
<td>0.878</td>
<td>0.977</td>
<td>0.925</td>
</tr>
<tr>
<td>Advanced Bayesian model</td>
<td>0.880</td>
<td>0.943</td>
<td>0.902</td>
</tr>
</tbody>
</table>

Conclusions

BayesInfer 1.0 achieved a superior performance in both precision and recall (precision = 0.878, recall = 0.977, F-measure = 0.925), in comparison with the MDAP algorithm we previously proposed (precision = 0.867, recall = 0.886, F-measure = 0.892), and with ProteinProphet (precision = 0.847, recall = 0.943, F-measure = 0.892); and the model is flexible for further extensions. We expect potential applications of the method in shotgun proteomics.

References


Acknowledgements

We acknowledge the support of the NCI grant U24 CA134680-01 to F. Regnier, RUA, PR, HT et al. HT and RJA acknowledge the support from NIH/NCRR grant P41RR018942. HT is partially supported by NSF award DBI-0842997. PR is partially supported by NSF award DBI-0644017.