

iProphet: Improved Validation of Peptide and Protein IDs in the Trans-Proteomic Pipeline

David Shteynberg¹; Ruedi Aebersold^{1,3,4}; Eric Deutsch¹; Henry Lam¹; Luis Mendoza¹; Joshua Tasman¹; Alexey Nesvizhskii²;

¹Institute for Systems Biology, Seattle, WA; ²Department of Pathology, University of Michigan, Ann Arbor, MI; ³Institute for Molecular Systems Biology (ETH), Zurich, Switzerland; ⁴University of Zurich, Zurich, Switzerland

Introduction

The Trans-Proteomic Pipeline (TPP)⁹ is a freely available, open-source software suite for the analysis of shotgun mass-spectrometry data. The suite includes tools for conversion of raw instrument data to our open mzXML format; import of spectral search engine (Sequest, Mascot, X!Tandem, Phenyx, OMSSA, and Comet) results to our open pepXML format; peptide-level statistical validation of search engine results with PeptideProphet¹, protein-level statistical validation with ProteinProphet³ and export to our open protXML format; and quantitation for many differential labeling techniques including SILAC, ICAT, and iTRAQ. We also provide tools for visualizing and interacting with the data as it is processed through the pipeline. The software is available for Windows and Linux systems. The Windows distribution includes an easy-to-use installer, which installs and configures a webserver for a graphical user interface to the tools.

PeptideProphet

The PeptideProphet classifier is a statistical tool that is an integral part of the TPP. This tool uses the expectation maximization (EM) algorithm to model the most likely distributions of spectral matches, returned by database search algorithms (e.g. Sequest⁴), among correctly and incorrectly assigned peptide ion matches.

Peptide Features Modeled by PeptideProphet:

- Database Search Scores (combined in a Discriminant Score)
- Number of Enzymatically Tolerable Termini (NTT)
- Number of Missed Cleavage Sites (NMC)
- Peptide Theoretical vs. Measured Mass Difference (MassDiff)
- Peptide ICAT-compatibility (ICAT Cysteine enriched data)
- Peptide N-glyco motif (N-glycosylation motif enriched data)
- Peptide pI⁷ (FFE and OGE data)
- Peptide Retention Time (Expected vs. Measured)

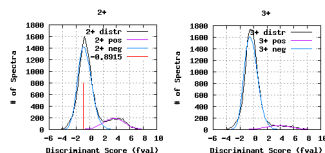
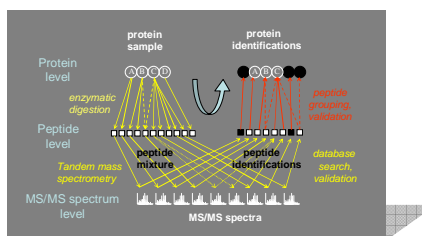


Fig 1: Parametric mixture distribution models estimated by PeptideProphet for 2+ and 3+ spectra from a *Streptococcus pyogenes* sample analyzed on an LTQ.

ProteinProphet

Since MS/MS spectra are produced by peptides, and not proteins, there is a need for an additional statistical model for validation of the identifications at the protein level. ProteinProphet is another vital part of the TPP, it implements a model that has as input the list of peptides assigned to MS/MS spectra and corresponding probabilities that those peptide assignments. Different peptide ion identifications corresponding to the same protein are combined together to estimate the probability that their corresponding protein is present in the sample. This protein grouping information is then employed to adjust the individual peptide probabilities, thus making the approach more accurate. ProteinProphet also address the problem that we call degeneracy, which occurs when one peptide corresponds to several different proteins.



PeptideProphet → iProphet → ProteinProphet

PeptideProphet is a commonly used open-source tool for validating individual MS/MS spectral matches to peptides made by a single database search engine. The proposed novel computational approach, iProphet, allows more precise integration of information supporting the identification of each unique peptide sequence from multiple MS/MS spectra. iProphet takes as input PeptideProphet spectrum-level results from multiple LC-MS/MS runs, and then computes a new probability at the level of unique peptide sequence. The new framework allows combining results from multiple search tools, and also takes into account other supporting factors including: number of sibling experiments identifying same peptide ions, number of replicate ion identifications, sibling ions, and sibling modification states. ProteinProphet is then used in IPROPHET mode to recompute iProphet peptide probabilities by including the number of sibling peptides identified within a protein (NSP) information in the probability calculation and computing final protein level probabilities with improved accuracy utilizing the discriminating power provided by these additional models.

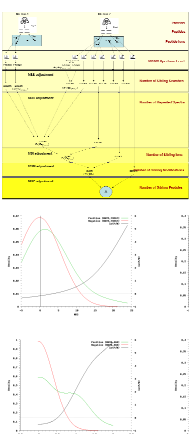


Fig 2: Overview of the strategy to combine the probabilities for spectrum peptide assignments from multiple lines of evidence for the same peptide as implemented by the iProphet program. Probability adjustments are made for the number of sibling searches (NSS), replicate spectra (NRS), sibling ions (NSI, i.e. different charges), sibling modifications (NSM), and sibling peptides (NSP, i.e. within the same protein).

iProphet Algorithm

All models in iProphet are computed using Kernel Density Estimates initially weighted by PeptideProphet probabilities p for the correct models and weighted by $1 - p$ for the incorrect models. The 2 separate EM steps are used during computation:

EM Step I:

1. Iteratively Compute NSS Model Until Convergence using estimates for NSS and probabilities computed according to the equations below:

$$NSS_{i,j,k} = \frac{\sum_{l \in \text{Siblings}} p(P_{i,j,k,l})}{\sum_{l \in \text{Siblings}} p(P_{i,j,k,l}) + \sum_{l \in \text{Incorrect}} p(P_{i,j,k,l})}$$

$$p(NSS_i | +) = \frac{\sum_{j,k} p(P_{i,j,k})}{\sum_{j,k} p(P_{i,j,k}) + \sum_{j,k} p(NSS_i | -)}$$

$$p(NSS_i | -) = \frac{\sum_{j,k} [1 - p(P_{i,j,k})]}{\sum_{j,k} [1 - p(P_{i,j,k})] + \sum_{j,k} p(NSS_i | +)}$$

$$p(P_{i,j,k}) = p(+ | NSS_{i,j,k}) = \frac{p(NSS_{i,j,k} | +) p(P_{i,j,k})}{p(NSS_{i,j,k} | +) p(P_{i,j,k}) + p(NSS_{i,j,k} | -) [1 - p(P_{i,j,k})]}$$

2. For each spectrum take the peptide identification with the highest probability

EM Step II:

- Using new probabilities iteratively compute NRS, NSI, NSM models (assuming independence) until convergence.
- The estimates parameters NSI and NSM are computed for each spectrum identification by summing all probabilities of identifications in the dataset matching the current identification according to the given model.
- For a spectrum match having n replicate spectra in the dataset with probabilities p_1, \dots, p_n , the NRS parameter is estimated according to the formula: $NRS = \sum_{i=1}^n (p_i - 0.5)$

This method prevents iProphet from over-promoting identifications of similar spectra which consistently match the same incorrect peptide

Results

The performance of the new method is demonstrated using a Yeast SILAC-labeled sample analyzed on an LTQ-Orbitrap instrument and searched using SEQUEST, MASCOT, and TANDEM with K-score. The false discovery rates (FDR) were estimated using the target-decoy database approach. The search results were analyzed separately using PeptideProphet and then using the new iProphet approach. The receiver operating characteristics (ROC) curves plotting the estimated number of correct identifications for a given FDR (Fig. 3) demonstrated improved discrimination between correct and incorrect peptide identifications compared to the results of any of the PeptideProphet analyses applied to output from any search engine alone. The improvement in the accuracy and statistical power of computed probabilities has been further demonstrated using a large collection of diverse experiments extracted from PeptideAtlas database. The new tool is fully compatible with ProteinProphet for subsequent protein inference analysis and is integrated in the Trans-Proteomic Pipeline.

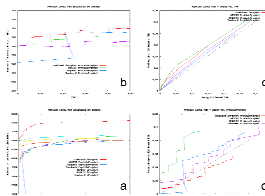


Fig 3: (a) On the peptide level iProphet performs much better at any given FDR than PeptideProphet alone, best results are obtained when all search engines are combined; (b) ProteinProphet using iProphet results performs better than standard ProteinProphet on any single analysis at a set FDR; (c) iProphet estimated probabilities can be used to compute another measure of FDR (separate from Decoy estimated FDR), this measure of peptide FDR is just as accurate as computing FDR when compared to Decoy FDR as FDR computed by using PeptideProphet probabilities; (d) ProteinProphet probabilities generated on iProphet results as a measure of protein FDR is just as accurate as FDR computed by the standard ProteinProphet probabilities.

Conclusions

- Combining multiple sources of information improves validation of spectrum assignments.
- iProphet recomputes peptide level probabilities using additional information from the PeptideProphet analysis.
- ProteinProphet in IPROPHET mode works on iProphet data to generate more accurate protein level probabilities
- iProphet improves sensitivity at a fixed FDR both at the peptide and proteins levels.
- iProphet improves accuracy of statistical modelling with large amounts of data such as PeptideAtlas database.
- iProphet is integrated with the Trans-Proteomic Pipeline (TPP), open-source proteomics data analysis and validation toolset, and can be enabled with the xinteract command
- iProphet has options to selectively disable any of its models.

References

- 1) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal Chem* **2002**, *74*, (20), 5383-92.
- 2) Nesvizhskii, A. I.; Aebersold, R. *Drug Discov Today* **2004**, *9*, (4), 173-81
- 3) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal Chem* **2003**, *75*, 4646-58.
- 4) Eng, J.; McCormack, A. L.; Yates, J. R., Jr. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976-989.
- 5) Choi, H.W.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7* (01), 47-50.
- 6) Choi, H.W.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7* (01), 254-265.
- 7) Malmstrom, J.; Lee, H.; Nesvizhskii, A.; Shteynberg, D.; Mohanty, S.; Brunner, E.; Ye, M.; Weber, G.; Eckerskorn, C.; Aebersold, R. *J. Proteome Res.* **2006**, *Sep 15* (9):2241-2249
- 8) <http://tools.proteomecenter.org>

Acknowledgement

This work is supported in part by NHLBI contract No. N01-HV-28179. This work is supported in part by NIH grant No. CA-126239 (P.A. Nesvizhskii). We would also like to thank all Aebersold lab members who helped with this project.

