

Incorporating Theoretical Peptide pI Information in PeptideProphet to Improve Validation of MS/MS Fractions Separated by Isoelectric Focusing

David Shteynberg¹, Alexey Nesvizhskii³, Johan Malmstroem², and Ruedi Aebersold²

¹Institute for Systems Biology, Seattle, WA;

²Institute for Molecular Systems Biology (ETH), Zurich, Switzerland;

³Department of Pathology, University of Michigan, Ann Arbor, MI

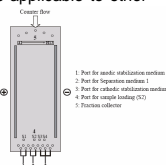


Introduction

Traditional multidimensional LC-MS/MS experiments are generally carried out by separating the peptides by an ion-exchange chromatography in the first dimension, followed by reversed-phase liquid chromatography. Isoelectric focusing (IEF) has recently proven itself a viable alternative to ion-exchange separation in the first dimension by allowing separation of peptides based on their relative pI. Free-flow electrophoresis (FFE) is a technology for IEF that can currently be directly interfaced to an LC-MS/MS analysis system (Fig 1). Peptide pI can easily be calculated from the peptide sequence. Here we present a novel addition to the PeptideProphet¹ model, which uses this information to improve the confidence of database search identified peptides sequenced in FFE-LC-MS/MS experiments. This model is applicable to other means of peptide separation by IEF.

Fig 1: Schematic representation of the continuous FFE apparatus^{5,6}.

Separation chamber: 500x100x4 mm; 11-17 ports for stabilization and separation media; S1-S4 ports for sample delivery (S2 was used here). Counter flow medium: 8 M urea and 250 mM mannitol; anodic stabilization medium: 100 mM α -hydroxyisobutyric acid (HIBA), 150 mM DL-2-aminobutyric acid, 100 mM nicotinamide and 15 mM glycol-glycine, 8 M Urea and 250 mM mannitol. Separation media 1: 23% protyle mixture 4-7, 8 M urea and 250 mM mannitol and the cathodic stabilization media: 75 mM ethanolamine, 75 mM AMPPO, 150 mM TAPS and 30mM HEPES, 8 M urea and 250 mM mannitol. Flow rate for separation media: 57 mL/h; flow rate for sample delivery: 1mL/h. Applied voltage 350 V (19 mA).



Method

PeptideProphet is a commonly accepted open-source tool for validating MS/MS spectra assignments to peptides made by database search engines such as SEQUEST³ or MASCOT⁴. This program already includes a number of peptide attributes when calculating the probability of peptide assignments to MS/MS spectra. Peptide pI is an attribute which is easy to compute reliably, but until now has not been utilized by PeptideProphet to improve classification of peptide identifications in FFE experiments. The presented work provides a novel approach for using this information to improve the PeptideProphet classifier. Although, we tried several variations of the algorithm, the basic theme involves calculating the distribution of peptide pI values among high confidence peptides from every fraction, and then computing a pI z-score for every peptide relative to the fraction in which it was identified.

Basic Algorithm:

1. Run PeptideProphet until the model converges on a solution
2. For each MS run:
 - a. Calculate pI distribution probability-weighted mean among high probability peptides (probability at least 90%)
 - b. Calculate pI distribution standard deviation among all peptide IDs (weighted by probability) using mean calculated in step 2a.
3. Calculate peptide pI z-scores among all peptide IDs
4. Using PeptideProphet probabilities as the weighting factor, calculate the distribution of pI z-scores among "correct" and "incorrect" identifications.
5. Calculate new PeptideProphet probabilities using the calculated pI model.

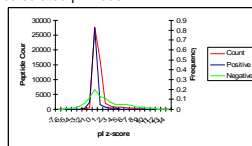


Fig 2: Learned and Actual pI z-score Distributions - The red curve shows the distribution of peptide pI z-scores among all peptides with probability at least 5%, the blue curve shows the distribution of pI z-score among correct IDs, the green curve shows the distribution of pI z-scores among incorrect IDs including peptides with probability below 5%.

Results

In the FFE workflow, peptides are separated into multiple fractions with different pI ranges. Peptides that are observed in the same fraction are more likely to have pI values that are within a narrow range. This information can be used to further increase the discrimination between correct and incorrect peptide identifications. The work we present here allows PeptideProphet to model pI values from all fractions of an FFE experiment and is fully compatible with this workflow. The way peptide pI model gets computed is unlike any other model in PeptideProphet; in that a single pI z-score model (Fig 2) is computed for all charge states, since pI distributions are not expected to vary depending on the charge. We tested our approach on a set of 40 fractions generated from a single run of the FFE instrument on a *Drosophila melanogaster* sample. These fractions were analyzed on a ThermoFinnigan LTQ ion trap mass-spectrometer and generated a total of 409,131 MS/MS spectra which were all searched with SEQUEST against the drosophila NCI database (Dec. 18, 2004). This data was then analyzed by PeptideProphet with and without applying the novel pI model for calculating the probabilities. ProteinProphet² was used to infer proteins and calculate protein probabilities from both runs.

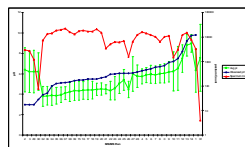


Fig 3: Theoretical Peptide pI and Observed Fraction pH - The observed pH of the FFE Fraction is shown in blue, the theoretical average pI of each FFE Fraction is shown in green (with error bars that correspond to 1 standard deviation), the red curve shows number potential peptide IDs in each fraction

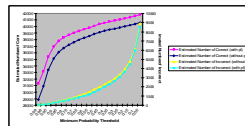


Fig 4: PeptideProphet Estimated Number of Correct and Incorrect Peptide IDs, with and without the pI Model - The estimated number of correct IDs with the pI model is shown in magenta, the blue curve shows the estimated number of correct IDs without the pI model. The cyan and yellow curves shows the estimated number of incorrect IDs, with and without the pI model, respectively.

To evaluate the accuracy of the PeptideProphet probabilities we devised a test dataset where a sample of *Streptococcus pyogenes* was run through the FFE-LC-MS/MS platform, and the spectral data was then searched against a database of the proteome of the organism appended with a reversed human database. The results were then analyzed with PeptideProphet twice, with and without the pI model. We then assumed that all spectral hits to sequences in the reversed human dataset are incorrect, and all others are correct. The spectrum hits were then ordered by PeptideProphet probability, and a sliding window of length 200 was used to get an estimate of the actual probability of the spectrum at the center of this window. The results are shown in Fig 5 and Fig 6.

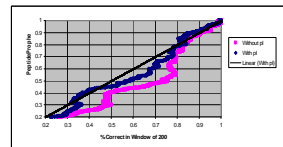


Fig 5: Actual Probability v. PeptideProphet Probability - The blue curve shows the results with pI model, the magenta curve shows the results without using the pI information.

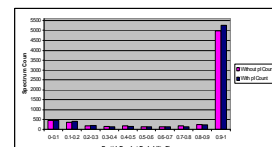


Fig 6: Distributions of Spectrum Matches - The blue bars show the distribution of spectra across PeptideProphet probability bins with pI information, the magenta bars show the distribution without pI.

Discussion

From Fig 3 we can see that given sufficient data, the learned pI distribution closely follows the observed pH for the corresponding FFE fraction, suggesting that this information can contribute some discriminating power for classification of correct and incorrect peptide identifications. Fig 4 further supports this by showing that when the pI model is used on FFE data, the estimated number of correct peptide IDs is higher (by roughly 2,000 high-probability peptides) than when the model is not used, and the estimated number of incorrect peptide IDs is unchanged if not slightly lower. The same minimum probability thresholds at the peptide level applied to both runs showed a greater sensitivity and lower error rate in the run that utilized pI information (Fig 7), suggesting that using the pI model allows PeptideProphet to correctly classify more peptide assignments. On the protein level, the use of the pI information resulted in an increase from 1761(327 single hits) to 1824(356 single hits) in the number of proteins identified with probability of at least 0.9. It also resulted in an increase from 9071 to 9479 in the number of unique peptides identified having ProteinProphet adjusted peptide probability of at least 0.9. The overall relatively small increase in the number of identified peptides and proteins compared to other studies^{7,8} indicates that PeptideProphet performs well even without the pI information.

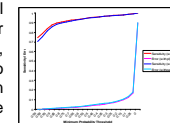


Fig 7: Estimated Sensitivity and Error Rates with and without the pI model - The estimated sensitivity with and without the pI model is shown in the red and blue curves, respectively. The magenta and cyan curves shows the estimated error, with and without the pI model, respectively

Conclusions

- The pI distributions learned by PeptideProphet on FFE data correlate with the observed pH of corresponding the FFE fractions.
- The pI information carries additional power to discriminate between correct and incorrect peptide identifications in FFE experiments.
- PeptideProphet running on FFE data is more accurate when pI model is used than when it is not used and the results have more high probability peptide assignments.
 - When the pI model is used, PeptideProphet probabilities are closer to the actual probabilities, where the best fit is represented by the $y=x$ line (Fig 5).
 - With the pI model the tool finds more high probability peptides (Fig 6).
- Using the pI model on FFE data allows PeptideProphet to correctly classify a greater number of peptide assignments to CID spectra.
 - This subsequently leads to a greater number of correct protein IDs.
 - Allows for deeper mining of the proteome.
- The pI model is fully integrated in the Trans-Proteomic Pipeline (TPP)⁹, open-source proteomics data analysis and validation toolset.

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) Eng, J.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom* **1994**, *5*, 976-989.
- 4) Perkins, D.N.; Pappin, D.J.C.; Creasy, D.M.; Cottrell, J.S. *Electrophoresis* **1999**, *20*, 3551-3567
- 5) Kivankova, L.; Book, P. *Electrophoresis* **1998**, *19*, (7), 1064.
- 6) Hoffmann, P.; Ji, H.; Moritz, R. L.; Connelly, L. M.; Frecklington, D. F.; Layton, M. J.; Eddes, J. S.; Simpson, R. J. *Proteomics* **2001**, *1*, (7), 807-18.
- 7) Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L., Jr. *J Proteome Res* **2004**, *3*, (1), 112-9.
- 8) Xia, H.; Banshakavi, S.; Griffin, T. J. *J. Anal Chem* **2005**, *77*, (10), 3198-207.
- 9) <http://tools.proteomecenter.org>

Acknowledgement

This work was supported in part by NHLBI contract No. N01-HV-28179. We would like to thank Andy Keller, Jimmy Eng and Eric Deutsch for their valuable discussions that helped develop this work. We would also like to thank all current and former Aebersold lab members who helped with this project.

