Automated Parameter Setting for Protein Database Searches

Wilfred H. Tang, Yong J. Kil, Christopher Becker, Marshall Bern Protein Metrics Inc., San Carlos, California Contact: wtang@proteinmetrics.com

PROTEIN METRICS

Powerful software for protein characterization

www.proteinmetrics.com

Introduction

Mass-spectrometry-based proteomics has seen dramatic experimental improvements in usability as well as performance. By contrast, informatics for identifying proteins has seen little improvement in usability.

We present an informatics workflow featuring the **Preview™ and Byonic™ software** that improves usability and greatly decreases the rate of inadvertent mistakes by setting database search parameters in a semi-automated, guided fashion. This approach also provides valuable feedback on sample characteristics and laboratory practices.

Informatics Workflow

Simple inputs
- Mass spectra
- Protein database (FASTA)
- Fragmentation type (collisions or electrons)
- Digestion enzyme
- (If applicable) Chemical labelling (e.g., TMT, iTRAQ)

Now available as FREEWARE
www.proteinmetrics.com

Preview™ Protein Metrics Inc. — Quick (gen. < 1.0 min)

Suggested parameters for full database search

Scientist quickly reviews parameters

Byonic™ Protein Metrics Inc. → **Search results**

Also available as part of:

Ease of Use

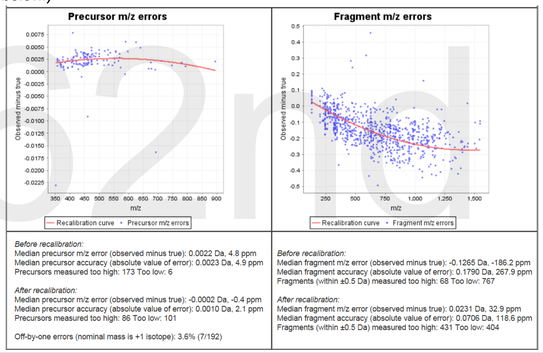
The inputs to this workflow are few in number and easy to set:
- Mass spectra
- Protein database (FASTA)
- Fragmentation type (collisions or electrons)
- Digestion enzyme
- (If applicable) Chemical labelling (e.g., TMT, iTRAQ)

Preview™ runs a quick analysis on a sample to determine the optimal parameters to be used in a full Byonic™ database search. These parameters include:
- Digestion specificity (e.g., fully tryptic or semi-tryptic)
- Precursor mass tolerance
- Fragment mass tolerance
- Cysteine alkylation
- Significant variable modifications (e.g., oxidation, deamidation)

Some of these parameters can be unobvious for the user to set well. Improper parameter setting (often referred to as "user error") can lead to poor results.
Automatic parameter setting greatly reduces such "user error."

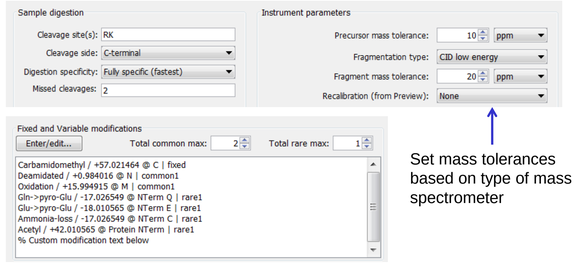
Semi-Automated vs. Fully Automated

We recommend that the user review and potentially edit the **Byonic parameters suggested by Preview**. The user may have knowledge that is challenging for Preview to infer. For example:
- Preview excels at finding sample handling artifacts but may miss biological post-translational modifications if they are only found on low-abundance proteins. (For speed, Preview focuses analysis on the top proteins.)
- Preview cannot distinguish ion trap CID vs. QTOF/HCD vs. TOF-TOF fragmentation.
- Preview does not perform glycopeptide analysis.
- There is an option to recalibrate the precursor and fragment m/z from . If recalibration is chosen, the precursor mass tolerance and fragment mass tolerance should be manually adjusted accordingly – we suggest using 5 times the post-recalibration median precursor & fragment accuracies. (See figure below.)



Performance Evaluation

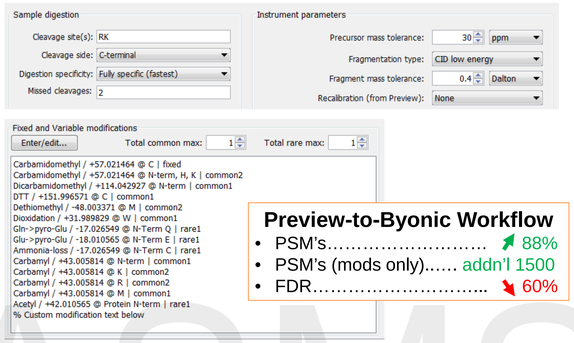
We evaluate the performance of the Preview-to-Byonic workflow by comparing to a typical database search below (referred to as the control or standard search):



Set mass tolerances based on type of mass spectrometer

Case Study #1: HeLa Cell Tryptic Digest

The Byonic search was run as suggested by Preview:

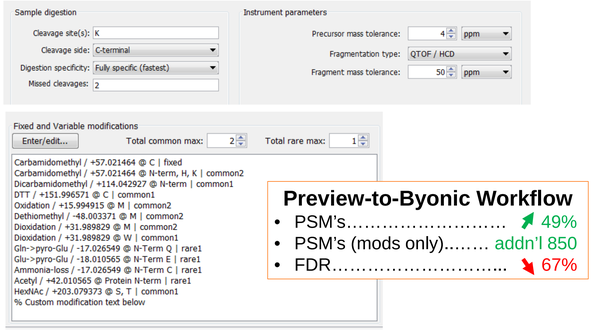


Preview-to-Byonic Workflow
- PSM's…… ↑ 88%
- PSM's (mods only)…… addn'l 1500
- FDR…… ↓ 60%

The Preview-to-Byonic workflow identifies 3172 spectra at 0.4% FDR. (By comparison, the standard search identifies 1688 spectra at 1.0% FDR.) This is an increase of 88%, largely due to searching for side products resulting from **cysteine over-alkylation** and **urea treatment (for protein denaturation)**.

Case Study #2: HCD O-GlcNAc Data

O-GlcNAc modified proteins were enriched from HEK293T cells using O-GlcNAc-specific IgG monoclonal antibodies. We modify the Byonic search parameters suggested by Preview:
- Add O-GlcNAc as a variable modification.
- Specify the fragmentation type as QTOF/HCD rather than CID



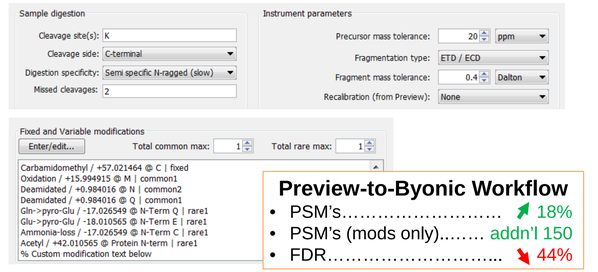
Preview-to-Byonic Workflow
- PSM's…… ↑ 49%
- PSM's (mods only)…… addn'l 850
- FDR…… ↓ 67%

The Preview-to-Byonic workflow identifies 1986 spectra at 0.1% FDR. (By comparison, the standard search identifies 1336 spectra at 0.3% FDR.) This is an increase of 49%, largely due to searching for side products resulting from **cysteine over-alkylation**.

Case Study #3: 2011 ABRF iPRG ETD Data

This is the data set used in the 2011 study conducted by the ABRF Proteome Informatics Research Group (iPRG): Identification of Electron Transfer Dissociation (ETD) Mass Spectra.

The Byonic search was run as suggested by Preview:



Preview-to-Byonic Workflow
- PSM's…… ↑ 18%
- PSM's (mods only)…… addn'l 150
- FDR…… ↓ 44%

The Preview-to-Byonic workflow identifies 3521 spectra at 0.5% FDR. (By comparison, the standard search identifies 2975 spectra at 1.1% FDR.) This is an increase of 18%, largely due to searching for **semi-tryptic peptides (N-ragged, C-term specific)**.

Discussion and Conclusions

Sample characteristics as well as handling artifacts and instrument calibration vary. This inherent variability makes data analysis challenging, as there is no standard database search that works well for all data. Optimizing search results often requires user expertise and trial-and-error iterating of different parameters.

The **Preview-to-Byonic workflow** described here enables users, expert and non-expert alike, to consistently get good search results in a simple and robust manner.
- More total PSMs
- More, relevant modifications identified
- More information about sample and sample processing.

References

Y. Kil, C. Becker, W. Sandoval, D. Goldberg, M. Bern, "Preview: A Program for Surveying Shotgun Proteomics Tandem Mass Spectrometry Data," *Anal. Chem.* **83**, 5259-5267 (2011).

P. Zhao, R. Viner, C. F. Teo, G. J. Boons, D. Horn, L. Wells, "Combining high-energy C-trap dissociation and electron transfer dissociation for protein O-GlcNAc modification site assignment," *J. Proteome Res.* **10**, 4088–4104 (2011).