Peptide Database Search Strategies to Improve Peptide Identifications Using High Resolution Mass Spectrometry

Edward J. Hsieh, Michael R. Hoopmann, Brendan MacLean, Michael J. MacCoss
Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065

Overview

In shotgun proteomics, the analysis of tandem mass spectrometry data from peptides can benefit greatly from high mass accuracy measurements. We have developed a software program, *Bullseye*, that uses high resolution precursor scans to assign precursor monoisotopic masses to tandem spectra more accurately than what is determined by the instrument's onboard computer. We examine two different search strategies to determine the most effective manner to utilize this high mass accuracy information for peptide identifications. The first method constrains the database search to those peptides in a narrow mass window around the precursor mass. The second method expands the search to a wide mass window, but then discards peptide identifications that lie outside a narrow mass tolerance.

Our results indicate:

- *Bullseye* offers improvements in mass accuracy and is able to remove poor quality spectra.
- Peptide identifications are improved when mass accuracy is used as a post-database search filter.
- Database searches with a narrow search window result in fewer peptide identifications, but significant improvements in database search times.

Introduction

High resolution mass spectrometers such as hybrid LTQ-FT or LTQ-Orbitrap combine the strengths of a Fourier transform mass analyzer with a linear ion trap mass spectrometer. These hybrid instruments are capable of recording survey mass spectra in high resolution while simultaneously acquiring MS/MS spectra at a lower resolution in the linear ion trap.

Before using the Fourier transform mass spectrum to improve the analysis of the low resolution MS/MS spectrum several technical hurdles need to be overcome. First, the high resolution Fourier transform mass spectrum must be must be deisotoped to obtain the monoisotopic mass and the charge state information. Secondly, the correct assignment of the resulting monoisotopic mass and charge state must be made to MS/MS spectra. This is complicated by the wide fragmentation window used for MS/MS and because peptides are often isolated prior to the chromatographic apex of their elution.

We developed a software tool, *Bullseye*, which we use to determine the monoisotopic mass of the precursor ion of MS/MS fragmentation spectra. *Bullseye* uses the results from data that is processed with the Hardklör algorithm to identify persistent peptide isotope distributions (PIDs) from high resolution mass spectra. For each MS/MS spectrum all possible matching precursor PIDs are determined.

We present results from the evaluation of two different database search methods to determine the optimal approach for the use of high mass accuracy. The first method limits the precursor mass in the database search to only those that are within a very narrow mass window surrounding precursor monoisotopic mass; as has been shown previously. The second method uses a wide mass tolerance in the database search and then uses the accurate mass as a post-filter to discard all matches that possess a mass difference that is outside a narrow mass window. We have also examined the potential database search speed increases obtained from reducing the number of peptides searched.

Methods

Sample Preparation

- Tryptic digestion of ammonium bicarbonate soluble component of cell lysate from yeast *Saccharomyces cerevisiae* strain S288C.

Instrumentation

- The mass analyzer used was a Thermo LTQ-FT Ultra.
- Scan cycle consisted of one MS scan (400 – 1,400 m/z, 50,000 FWHM at m/z 400, profile mode) in the FTICR followed by five data-dependent MS/MS spectra acquired in the LTQ ion trap.
- The HPLC system used was an Agilent 1100 Quaternary Pump. Buffer flow was split to achieve an approximate flow rate of 200 nl/min through the chromatography column.
- Peptides were separated over a homemade 40 cm fused silica capillary (75 µm inner diameter) column packed with Jupiter Proteo (Phenomenex) reverse phase resin. The mobile phase gradient used was a 200 minute linear gradient from 9% acetonitrile, 0.1% formic acid to 29% acetonitrile, 0.1% formic acid.

Database Searching

- MS/MS spectra were searched with SEQUEST against a *S. cerevisiae* protein database.
- The false discovery rate (FDR) is defined as the ratio of the number of decoy peptide-spectrum matches (PSMs) greater than a given XCorr divided by the target peptide-spectrum matches greater than the same XCorr.

Data processing with *Bullseye*

- http://proteome.gs.washington.edu/software/bullseye

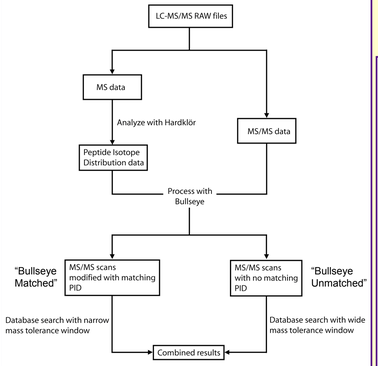


Figure 1 – Data processing scheme for Bullseye. High resolution MS and MS/MS scan data are extracted from RAW data files. MS data is analyzed with Hardklör to identify peptide isotope distributions (PID). *Bullseye* attempts to match a persistent peptide isotope distribution to each MS/MS scan. MS/MS scans are separated into two data files based on whether a matching PID is found. MS/MS scans with matching PIDs can be searched with a narrow mass tolerance window. MS/MS scans without a matching PID are searched with a wide mass tolerance window.

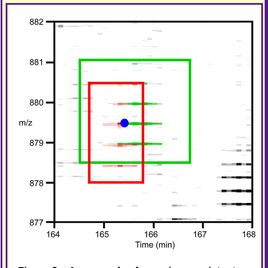


Figure 2 – An example of mapping persistent peptide isotope distributions (PPIDs) to a MS/MS spectrum. The MS full scan data is plotted and the center of a MS/MS scan event is indicated by the blue circle. Two separate PPID are identified by *Bullseye* (red and green highlighted signals) and are assigned to the MS/MS scan event. *Bullseye* generates a target box (red and green boxes) around each PPID and MS/MS scans that lie within the target box have their precursor mass values reassigned to the monoisotopic mass of the PPID.

Results

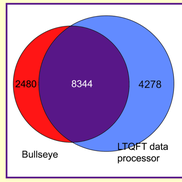


Figure 3 – Sample dataset was processed with *Bullseye*, and spectra that had a precursor mass assigned, were compared with the spectra from a dataset with precursor masses assigned by the LTQFT data processor.

Only 8344 spectra had precursor masses assigned by both methods.

Bullseye assigned a precursor mass to 2480 spectra exclusively and the LTQFT data system assigned a precursor mass to 4278 spectra exclusively.

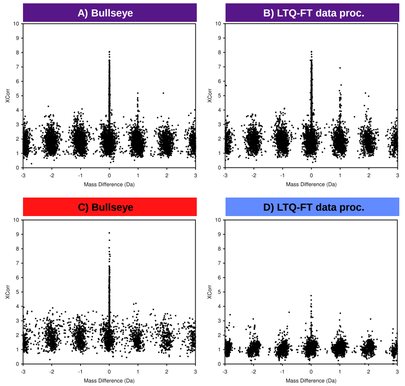
Figure 4 – Data subsets from Figure 3 were searched with SEQUEST, and their XCorr values vs Mass difference (observed – exact mass) plotted.

A & B) For spectra with mass assignments from *Bullseye* and onboard processor, *Bullseye* shows fewer instances of M+1 and M+2 mass assignment errors.

C) Spectra with *Bullseye* assigned masses that were not detected by the instrument.

D) Spectra with instrument assigned masses that were not detected by *Bullseye*.

Bulleye's requirement that PIDs persist chromatographically allows it to avoid spectra of poor quality that may be a result of noise, chemical contamination or very low abundance peptides.



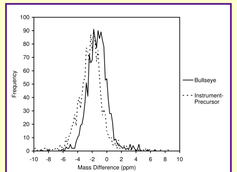


Figure 5 – Histogram of the mass error for precursor mass assignments made by *Bullseye* and the LTQ-FT Ultra firmware.

Bullseye uses the average of the mass measurement taken over the course of the peptide's elution to achieve an overall more accurate monoisotopic mass assignment.

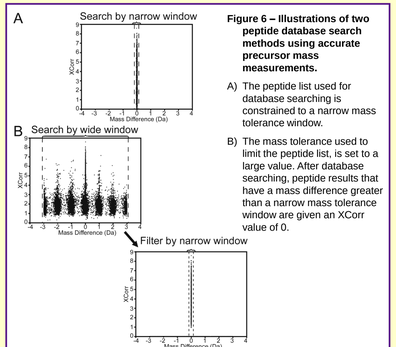


Figure 6 – Illustrations of two peptide database search methods using accurate precursor mass measurements.

A) The peptide list used for database searching is constrained to a narrow mass tolerance window.

B) The mass tolerance used to limit the peptide list, is set to a large value. After database searching, peptide results that have a mass difference greater than a narrow mass tolerance window are given an XCorr value of 0.

Figure 7 – Results of different search methods on peptide spectrum matches.
A) Results from narrowing the mass tolerance for peptides used in database searching.
B) Results for the database search of the *Bullseye* dataset with wide mass tolerances followed by filtering the results with a narrow mass tolerance window.

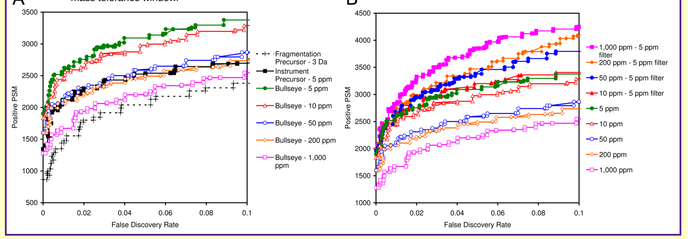


Table 1 –Database search results

Spectra Set	Database Search Mass Tolerance (precursor mass type)	Positive PSM (0.001 FDR)	Positive PSM (0.01 FDR)	Positive PSM (0.05 FDR)	Number of Spectra Searched[a]	**Search Time (min)[ab]**
Isolation m/z	± 3 Da (Average)	864	1440	2040	32003	**154.62**
LTQFT data proc.	± 5 ppm (Monoisotopic)	1403	2121	2525	21928	**8.92**
Bullseye Unmatched	± 3 Da (Average)	5	5	5	10800	**39.45**
Bullseye Matched	± 5 ppm (Monoisotopic)	1963	2543	3093	14774	**7.16**
Bullseye Matched	± 10 ppm, (Monoisotopic)	1898	2463	2978	14774	**8.61**
Bullseye Matched	± 50 ppm, (Monoisotopic)	1665	2111	2587	14774	**15.01**
Bullseye Matched	± 200 ppm, (Monoisotopic)	1600	2057	2492	14774	**22.25**
Bullseye Matched	± 1000 ppm, (Monoisotopic)	1278	1671	2202	14774	**74.94**
Bullseye Matched	± 10 ppm, (Monoisotopic). 5 ppm filtered	1965	2545	3201	14774	**8.61**
Bullseye Matched	± 50 ppm, (Monoisotopic). 5 ppm filtered	2019	2680	3380	14774	**15.01**
Bullseye Matched	± 200 ppm, (Monoisotopic) 5 ppm filtered	2054	2712	3471	14774	**22.25**
Bullseye Matched	± 1000 ppm, (Monoisotopic) 5 ppm filtered	2048	2825	3845	14774	**74.94**

[a] Results from target search.
[b] Average of three replicate searches. SEQUEST processes was distributed to 4 cores of an AMD 64-bit 2.5 Ghz processor.

Conclusions

- *Bullseye* offers improvements in assigning high resolution precursor masses compared to the onboard data processor of the LTQ-FT Ultra.
- By filtering spectra without persistent peptide isotope distributions, *Bullseye* is able to remove poor quality spectra which are placed in the Unmatched dataset.
- Database searching with a wide mass tolerance window, followed by a post-search mass accuracy filter increases the sensitivity of the positive peptide-spectrum-matches made.
- Database searching with a narrow mass constrained search window offers the advantage of greatly reducing the computational time needed to perform the search.
- Mass accuracy represents only one discriminating element that allows for distinguishing correct and incorrect PSMs. Other elements, such as enzyme cleavage specificity and predicted retention times, can be used to discriminate incorrect PSMs and when used in conjunction with mass accuracy should result in the best peptide identification results.

Acknowledgements

Support for this work was provided in part by National Institutes of Health grants P41 RR011823, S10 RR021026, and R01 DK069386. Edward J. Hsieh was supported by National Institutes of Health grant T32 HG00035.