Identifying Peptides the Way Experts Do: A Scientist-Driven Scoring System for Improving MS/MS Peptide Identification Confidence

Brian C. Searle and Mark Turner

Proteome Software Inc., 1336 SW Bertha Blvd, Portland, OR, 97219-2039, United States

E-mail: Brian.Searle@ProteomeSoftware.com, Telephone: (503) 244-6027, Fax: (503) 245-4910

Abstract

The validation of spectra derived from peptides is generally the most time consuming portion of post MS/MS acquisition research. While validating, scientists make decisions about peptide identification accuracy based on several features, such as the quality of match, spectrum noise, and the presence of unidentified peaks. To alleviate this tedious task, we have captured the pattern recognition features that scientists use to judge spectra. To do this, we first surveyed several scientists and found the peptide match features they rely on to validate MS/MS results. From these features, we developed an expert system, called the Sanity Score, and trained it against a controlled truth set. Finally, we have probabilistically combined the Sanity Score with results from the SEQUEST, Mascot and X!Tandem database search engines using the Scaffold computer program. Since the Sanity Score is based on chemical observations, rather than numerical algorithms used by the search engines such as cross correlation or shared peak count, we have found that combining database searching methods with the Sanity Score provides more accurate interpretation of MS/MS spectra. By mimicking the analysis done by scientists, this improved accuracy lessens the need to manually validate SEQUEST, Mascot and X!Tandem results.

Peptide Validation Survey

Since manual validation of peptide identifications is difficult, subjective and time consuming, we sought to create an automated validation scheme that is easy, objective and fast. We started by finding out how the expert mass spectrometrists validate data by having them respond anonymously to a survey. The results from the eleven people those who declared themselves as experts (median: over 5 years of experience) and the nine who said that they are intermediates (median: 2-3 years of experience) are summarized in Figure 1. There was no significant difference between the two groups. The complete survey can be accessed at: http://www.proteomesoftware.com/validation_quiz/.

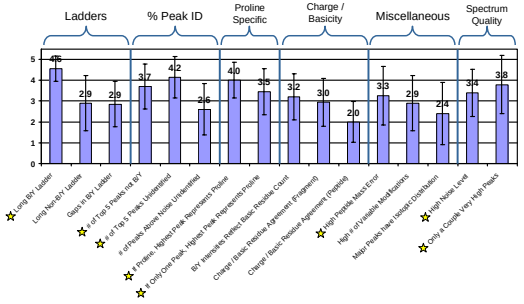

Figure 1: Results from peptide validation survey of 20 anonymous intermediate or expert level scientists. Features were ranked from 1 (least important) to 5 (most important). Error bars represent one standard deviation away from the mean value. Stars are used to signify the 8 highest rated features features.

Important Peptide Validation Features

The 8 highest rated peptide validation features were selected for further analysis. Quantitative representations of these features were approximated using the following criteria:

1) **Long B/Y Ladder**: length of the longest contiguous B or Y ladder
2) **# of Top 5 Peaks not B/Y**: total number of the top 5 peaks not identified as B or Y ions
3) **# of Top 5 Peaks Unidentified**: total number of the top 5 peaks that could not be interpreted. Possible interpretations considered included B/Y/A +1 and +2 ions, $-NH_3$ and $-H_2O$ when supported by the peptide, immonium ions, and internal proline fragments
4) **Highest Peak Represents Proline**: 1 if sequence contains a proline and 1 of the top 3 highest peaks is n-terminal to proline (-1 if not highest, 0 if no proline)
5) **Only Peak Represents Proline**: 1 if highest peak is > 50% of second highest peak and highest peak is n-terminal to proline (-1 if not highest, 0 if no proline)
6) **High Peptide Mass Error**: negative log of the mass difference between recorded and calculated peptide masses (delta mass), normalized by the median delta mass for the dataset
7) **High Noise Level**: intensity of the 15th peak (median signal) divided by the median peak intensity (noise)
8) **Only Very High Peaks**: highest intensity divided by the intensity of the 10th peak

Also included in this study was **Percentage Identified** (feature 9), the percentage of interpreted ion current divided by the total ion current, which is an estimation related to features 2 and 3.

Statistics and Score Generation

Stepwise linear discriminant analysis was used to generate a discriminant score (Sanity Score) from the nine validation features. Peptide identifications were derived using Mascot [1] to analyze the training datasets. Peptides were considered correctly identified if they matched proteins identified at the 99% probability level with at least 2 peptides identified at the 95% level using Scaffold [2]. Across all four training datasets, only four features were determined to be significantly useful in creating the Sanity Score (P-Value <0.05). Coefficients analysis for these features derived by the linear discriminant are listed in Table 1.

	Feature #	Soluble Ion Trap	Insoluble Ion Trap	Soluble Q-ToF	Insoluble Q-ToF	Average
ladder length	1	0.71	0.70	0.58	0.80	0.70
-log(delta mass)	6	0.16	-0.25	0.30	0.35	0.14
signal to noise	7	0.10	0.09	-0.20	-0.10	-0.03
% identified	9	0.46	0.44	0.57	0.36	0.46

Table 1: Normalized coefficients for the four features determined to be significant for discrimination.

The B/Y ion ladder length and the percentage identified scores were determined to be the most critical features. Although high signal to noise is a positive feature, the linear discriminant analysis performed used it only as a correction factor. As a result, it's significance could not be determined by the mean coefficient and was dropped.

Sanity Score Performance

The Sanity Score was tested using control protein datasets and was shown to effectively discriminate between known correct and incorrect peptide identifications. Discrimination using the Mascot Ion-Identity Score is shown for comparison (Figure 2). The sensitivity and specificity values for the Sanity Score and Mascot are very similar at 95% specificity (Table 1).

	Sensitivity	Specificity
Sanity (Q-ToF)	89%	95%
Mascot (Q-ToF)	90%	95%
Sanity (Ion Trap)	62%	95%
Mascot (Ion Trap)	64%	95%

Table 1: Sensitivity and specificity for the Sanity Score and for the Mascot Ion - Identity Score.

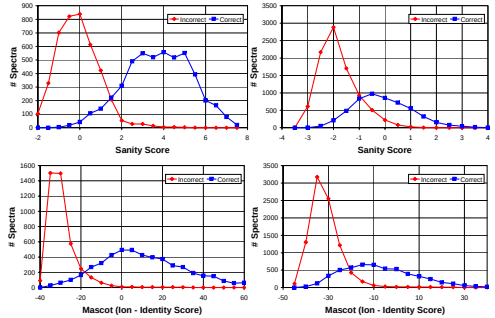

Figure 2: Score distributions for the Sanity Score and the Mascot Ion-Identity Score for Q-ToF (left) and Ion Trap (right) control datasets.

Validating Search Engines with the Sanity Score

Database search programs, such as Mascot [1], SEQUEST [3] and X! Tandem [4] use mathematical scoring methods based on correlation between a spectrum and a peptide model to differentiate between correct and incorrect identifications. The Sanity Score utilizes the same chemical intuition that scientists use to validate database searching engines. As a result, the interpretations made by the Sanity Score are somewhat orthogonal to interpretations made by database searching programs (Figure 3).

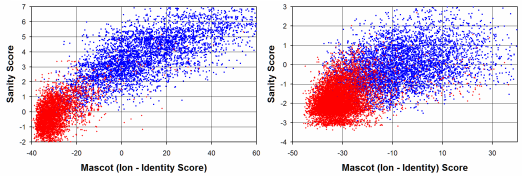

Figure 3: Scatterplots showing the variation between the Sanity Score and the Mascot Ion-Identity Score for Q-ToF (left) and Ion Trap (right) in the control protein testing datasets.

Figure 3 suggests that rescoring Mascot identifications with the Sanity Score will improve the separation between correct and incorrect identifications, particularly with Ion Trap data. As shown in Table 3, we can improve sensitivity by considering both Mascot and Sanity Scores using linear discriminant analysis.

	Sensitivity	Specificity
Both (Q-ToF)	93%	95%
Mascot (Q-ToF)	90%	95%
Both (Ion Trap)	79%	95%
Mascot (Ion Trap)	64%	95%

Table 3: Improved sensitivity and specificity when considering both the Sanity Score and the Mascot Ion - Identity Score.

Using the Sanity Score with Scaffold for Validation

Scaffold [2] is a computer program that allows the user to consider data from multiple sources simultaneously. One of the features of Scaffold is to probabilistically consider multiple search engines simultaneously to derive more accurate peptide identifications. Using this framework, we are currently evaluating using the Sanity Score to automatically objectively validate other database searching programs, including Mascot, SEQUEST, and X! Tandem.

Experimental Data Sets

In this study, two different datasets were analyzed using quadrapole-time of flight (Q-TOF) and ion trap tandem mass spectrometers [5]: (testing) ten purified proteins combined as a known control mixture and (training) soluble and insoluble human lens proteins separated using 2D liquid chromatography.

References

1.Perkins, D. N.; et al *Electrophoresis* **1999**, 20, 3551-3567.
2.Searle, B. C.; et al "Scaffold: A Program to Probabilistically Combine Results…", ABRF **2005**.
3.Eng, J. K.; et al *J. Am. Soc. Mass Spectrom.* **1994**, 5, 976-989.
4.Fenyo, D. and Beavis, R. C. *Anal. Chem.* **2003**, 768-774.
5.Searle, B. C.; et al *J. Proteome Res.* **2005**, 4, 546-554.

Acknowledgements

We thank Srinivasa Nagalla and Larry David for providing us with the data sets used in this study. We especially thank participants of the Peptide Validation Survey.