

Scaffold: A Program to Probabilistically Combine Results from Multiple MS/MS Database Search Engines

Brian C. Searle, James M. Brundage, 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

Database-searching programs generally identify only a fraction of the spectra acquired in a standard LC/MS/MS study of digested proteins. Subtle variations in database-searching algorithms of MS/MS spectra have been known to provide different identification results [1]. To leverage this variation, we developed Scaffold to probabilistically combine the results of multiple search engines, including SEQUEST, Mascot, and X!Tandem. We normally gain 20% to 100% more MS/MS spectrum identifications with each additional search engine, primarily due to increased confidence in low scoring matches. In addition, we use probabilistic clustering to mine information from the remaining spectra. Together, these methods increase the number of spectrum identifications by 50% to 350% in control experiments and in cataractous lens tissue, allowing us to explain a considerably larger fraction of LC/MS/MS studies of digested proteins.

Introduction to the Scaffold Analysis Framework

Scaffold is a computer program that performs analysis and collation of MS/MS sequencing results from SEQUEST [2], Mascot [3], and X!Tandem [4]. Scaffold (Figure 1) is built out of a series of individual components that:

- 1) organize results from various database-searching algorithms,
- 2) estimate peptide identification probabilities from the results,
- 3) merge the probabilities from the database-searching algorithms,
- 4) compute protein identification probabilities,
- 5) identify redundant spectra, and
- 6) filter remaining unidentified spectra.

New Implementations of the PeptideProphet [5] algorithm, and the ProteinProphet [6] algorithm estimate peptide and protein identification probabilities from results of database-search algorithms. Scaffold Merger probabilistically combines the results by increasing the probability if the algorithms agree and decreasing it if they report different solutions. Scaffold Clustering Engine identifies unmatched spectra that resemble previously matched spectra at the 95% probability level and marks them as redundant. Finally, Scaffold Noise Remover uses signal to noise thresholds to mark peptides with poor fragmentation and electronic noise.

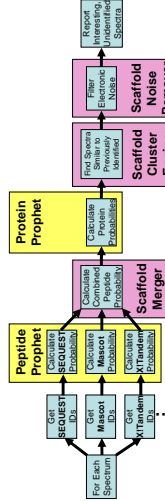


Figure 1: The Scaffold Analysis Framework for collating and analyzing results of multiple database-searching programs.

Experimental Data Sets

In this study, two different datasets were analyzed using quadrupole-time of flight (Q-TOF) and ion trap tandem mass spectrometers [7]: (a) ten purified proteins combined as a known control mixture and (b) soluble human lens separated using two-dimensional liquid chromatography.

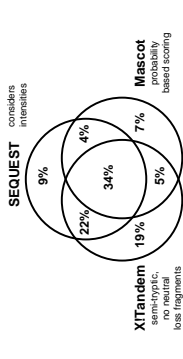


Figure 2: Venn diagram of correct spectrum identifications made by SEQUEST, X!Tandem, and Mascot at the 95% confidence level. The overlap between the three programs is surprisingly small.

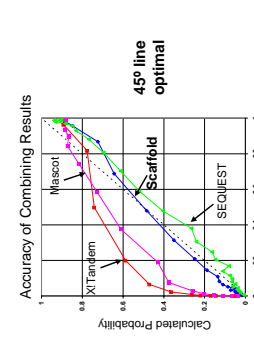


Figure 4: Scaffold Merger probabilities are closer to the ideal 45° line than individual PeptideProphet analyses of SEQUEST, Mascot, and X!Tandem.

Scaffold Merger
 As illustrated in Figure 2, there is little overlap between confident spectrum identifications made by popular database searching programs. Scaffold Merger is designed to probabilistically leverage scoring variation (Figure 3) between these programs. We estimate the agreement between the programs using a number of sibling programs score (NSP):

$$NSP = \sum_{i=1}^n p(+|D_i,+) / p(+|D_i,+)$$

where the probability assigned by each search program is $p(+|D_i)$. The probability of having a specific NSP value and being correct, $p(NSP|+)$, is estimated using an algorithm similar to that used in ProteinProphet [6]. A more accurate estimation of the probabilities is assigned using Bayes' Law:

$$p(+|D_i, NSP) = \frac{p(+|D_i)p(NSP|+) + p(-|D_i)p(NSP|-)}{p(+|D_i, NSP) + p(-|D_i, NSP)}$$

The Scaffold Merger probabilities are more accurate than those calculated by PeptideProphet individually (Figure 4).

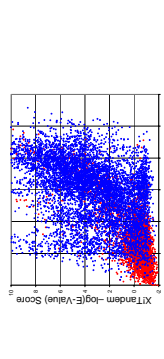


Figure 3: Scoring system discrepancy between SEQUEST and X!Tandem in the Q-TOF 10. Identified peptides. Correct identifications are labeled in blue while incorrect are labeled in red.

Scaffold Cluster Engine

The Scaffold Cluster Engine considers the distribution of Spearman correlation scores between all spectra with parent ion masses less than 5 AMU apart. Scaffold uses expectation maximization to create a mixture model of scores, which is applied to estimate the probability that two spectra were derived from the same peptide. If only one of the spectra is identified (Figure 5), Scaffold Cluster Engine marks the other as identified by association.

Scaffold Noise Remover

Scaffold Noise Remover discards unidentified spectra with signal to noise levels less than five-fold, which are often electrical noise. The Noise Remover also removes spectra with few peaks inside a dynamic range of twenty-fold, which are suggested to have poor fragmentation.

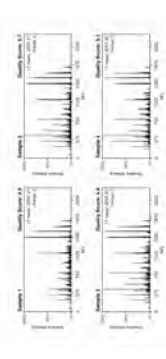


Figure 5: Similar MS/MS spectra identified by the Scaffold Cluster Engine. Confident identification of one will identify the rest.

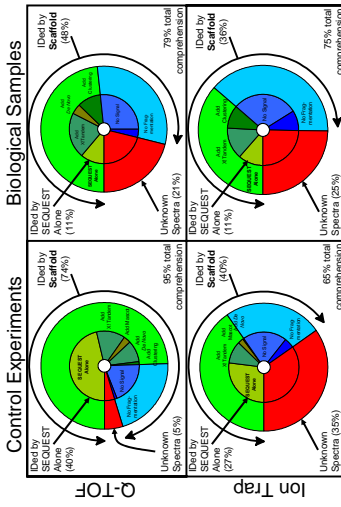


Figure 6: Percentage of MS/MS spectra identified by the Scaffold Analysis Framework, as compared to SEQUEST alone. Green marks identified spectra, blue marks spectra identified as electronic noise or with poor fragmentation, and red marks unidentified spectra. In all cases, Scaffold identified 1.5X to 4.5X more spectra than SEQUEST alone.

Results and Discussion

Scaffold was tested using control and biological samples analyzed on both Q-TOF and ion trap mass spectrometers. Matches identified as either control or lens proteins with greater than 95% confidence were accepted. We found that Scaffold identified 1.5X to 4.5X more spectra over any single database-searching program alone. Overall dataset comprehension rates increased from 11-40% using SEQUEST to 65-95% using Scaffold. Spectrum identification rate increases of 1.5X to 3X were normal when combining just SEQUEST and X!Tandem results with Scaffold. Identification rate increases were similar when comparing Scaffold to X!Tandem and Mascot individually. Scaffold gives scientists piece of mind in knowing that all their data is being considered. Scaffold's new automated probabilistic analysis technology, also allows scientists to confidently identify low abundance proteins by confirming "one-hit wonder" proteins with additional spectral evidence.

References

- (1) Reising, K. A., et al. *Anal. Chem.* 2004, 76, 3565-3569.
- (2) Eng, B., et al. *Anal. Chem.* 1994, 66, 876-880.
- (3) Searle, B. C., et al. *Anal. Chem.* 2002, 74, 3583-3592.
- (4) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* 1999, 20, 3501-3507.
- (5) Field, H. I.; Fungo, D.; Davalos, R. C. *Proteomics* 2002, 3, 64-77.
- (6) Keller, A.; Neuhoff, A.; Rober, E.; Aebersold, R. *Anal. Chem.* 2002, 74, 3583-3592.
- (7) Searle, B. C., et al. *J. Proteome Res.* 2005, in press.

Acknowledgements

We thank Ashley McCormack for helpful discussions and both Shrinvasa Nagalla and Larry DQF for providing us with the QTOF and IonTrap MS/MS data sets used in this study.

Call for Testers

We are currently seeking beta testers for our June 2005 release of Scaffold. Anyone interested in trying Scaffold in their lab should contact: Brian.Searle@ProteomeSoftware.com