The Effect of Data Analysis Choices on the Number of Proteins Identified in MS/MS Proteomics Experiments

Phillip A. Wilmarth[1] and Brian C. Searle[2]

[1] School of Dentistry, Oregon Health & Sciences University, 3181 SW Sam Jackson Park Road, Portland, OR 97239-3098, United States

[2] 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 numbers of proteins identified in complex mixtures depends on separation strategies and instrumentation choices. The numbers are also dependent on data analysis parameters, such as peptide score thresholds, enzyme search specificity, choice of protein database, use of protein parsimony filters, and minimum peptide count. We analyzed the same complex mixture (human whole saliva) using a variety of criteria and "identified" numbers of proteins that ranged from 138 to 3379. The primary criteria affecting this variation were the use of protein parsimony (i.e., reporting only the minimum set of protein sequences that adequately account for all observed peptides) and minimum peptide count.

One possible way to compare data sets from different experiments is through the use of peptide or protein false positive rates (FPRs) estimated from searches of sequence-reversed databases. However, most literature reports have adopted criteria designed to produce low peptide FPRs and have overlooked the fact that low peptide FPRs do not imply low protein FPRs. In this study, criteria where peptide FPRs were less than 1.7% could have protein FPRs ranging from 0.6% to 45%, depending on the workup criteria.

Introduction

Analysis of complex mixtures using bottom up techniques has become a routine and powerful way to characterize proteomes. Comparison of different experiments is difficult because the results strongly depend on the experimental conditions (sample preparation and separation, instrumentation, and data analysis methods). The increasing use of sequence-reversed databases to estimate peptide false positive rates (FPR) [1] and other probabilistic methods such as Peptide Prophet [2] allows criteria for different database search programs to be standardized. However, many other analysis choices beside search program and valid peptide criteria have a strong influence on the resulting list of identified proteins. Among these choices are which database to search, whether or not to invoke a minimum number of peptides per protein, and how to report an accurate number of proteins when limited sequence coverage makes it impossible to distinguish between multiple loci. Finally, the lack of protein FPR assessment done by most researchers also confounds dataset comparisons. In this study, we explore factors other than search programs and valid peptide criteria to see how strongly the results are affected by these factors.

Analysis Methods

Unstimulated human saliva was digested with trypsin, and peptides separated offline by polysulfoethyl A cation exchange. Fractions were analyzed by reverse phase LC-MS/MS using an LCQ Classic ion trap (ThermoFinnigan, San Jose, CA). Three technical replicates were run, creating 143880 DTA files, which were interpreted using SEQUEST [3] searching three databases assuming alkylated cysteines. Each database was searched twice, once with trypsin cleavage and once with no enzyme specificity. The databases were human subsets of SwissProt (4/18/06, 13,748 entries), IPI (v3.16, 62,322 entries), and NCBI nr (8/1/06, 131,517 entries), where FPRs were derived using concatenated reverse sequences. Results were tallied using DTASelect [4] and Scaffold [5] (Proteome Software, Portland, OR).

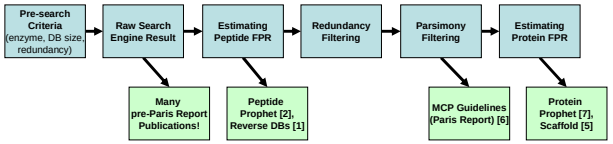
Acknowledgements

We thank Larry David for providing us with the data sets used in this study.

References:

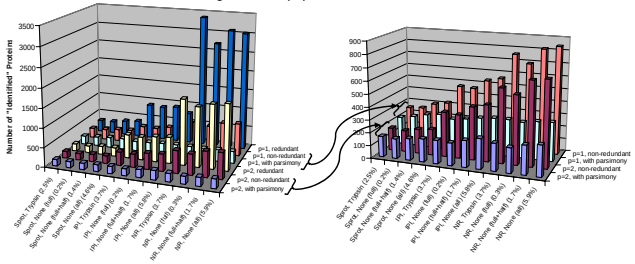
1. Elias, J. E.; et al. *Nature Methods* **2005**, 2, 667-675.
2. Killer, A.; et al. *Anal. Chem.* **2002**, 74, 5383-5392.
3. Eng, J. K.; et al. *J. Am. Soc. Mass Spectrom.* **1994**, 5, 976-989.
4. Tabb, D. L.; et al. J. Proteome Res. **2002**, 1, 21-26.
5. Searle, B. C.; et al. "Scaffold: A Program to Probabilistically Combine Results…", ABRF **2005**.
6. http://www.mcponline.org/misc/ParisReport_Final.shtml
7. Nesvizhskii A. I.; et al. *Anal. Chem.***2003**, 75, 4646-4658.

Data Analysis Workflow



Redundancy Inflates Protein Counts

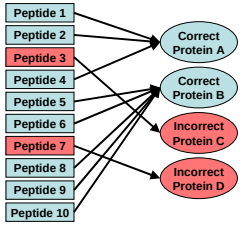
Database redundancy can greatly inflate the number of identified proteins. When redundant proteins were counted, the number of proteins increased 8-fold from the SwissProt database to the NCBI nr database where as many as 3379 proteins were reported. The figures below shows the number of non-redundant proteins (burgundy and salmon series), where we still observe a 3-fold increase with database size. This increase is similar for proteins identified by two or more peptides (burgundy) and for one or more peptide/protein identifications (salmon), which indicates that the increase is a feature of the databases, and not due to random matches. Parsimony filtering (lavender and light blue series) greatly reduces the dependence on database choice. It is important to note that all of the results are from identical data using the same peptide identification criteria.



Parsimony Filtering

Parsimony filtering reduces artifactually inflated numbers of identified proteins that occur when databases have large numbers of related sequences, such as NCBI nr. Since protein sequence coverage is typically low in complex mixtures, peptides may match to many similar protein sequences. Parsimony filtering is of critical importance for large databases such as NCBI nr and is much less important for minimally redundant databases like SwissProt. The above table shows consistent numbers of identified proteins across three protein databases, for tryptic and non-enzymatic searches, and for different peptide criteria when parsimony filtering is used. The small increase in number of proteins with the NCBI nr database are probably sequence variant immunoglobulin entries. All numbers in the table have been corrected by the number of false positive protein matches to sequence-reversed entries concatenated to each of the three databases.

Scaffold Criteria	Sprot tryptic	Sprot none	IPI tryptic	IPI none	nr tryptic	nr none	Ave. (SD)
1.8, 2.5, 3.5, 0.08, p=2	149	154	155	166	153	165	157(7)
Pep=0.95, p=2	161	153	172	167	174	165	165(8)
1.8, 2.5, 3.5, 0.08, p=1	234	235	248	272	275	279	257(21)
Pep=0.95, p=1	243	254	238	257	275	277	257(16)

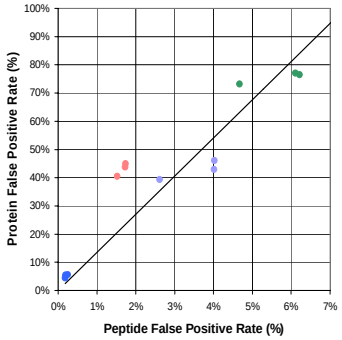


20% Peptide FPR → 50% Protein FPR

Peptide FPR is not Protein FPR!

False positive peptides are almost always assigned to single peptide/protein identifications, whereas true positive peptides are generally assigned in groups. The schematic to the left shows how only a few incorrect peptides can substantially taint the overall protein false positive rate.

Using the same commonly accepted criteria (xCorr 1.8, 2.5, 3.5 for +1, +2, and +3 peptides, respectively, with a 0.08 deltaCn minimum), there is a greater than ten-fold increase in protein FPRs over peptide FPRs in this saliva dataset, as shown in the figure below.



Protein and peptide false positive rates do not increase dramatically with increasing database size, but do vary substantially with digestion enzyme choice. This is likely due to the increase in decoy peptides, which provides more accurate deltaCn statistics. In this figure, the blue dots represent non-enzymatic searches, with a post filter to select only tryptic peptides. Salmon and green colored dots represent the same search filtered for tryptic/semi-tryptic peptides, and unfiltered, respectively. Finally, lavender dots represent tryptic-only searches.

As shown in the table to the right, two peptides per protein as a minimum requirement dramatically reduces the protein FPR. Interestingly, relatively low protein FPRs for <u>single</u> peptide/protein identifications can be achieved by performing a non-enzymatic search and then filtering for fully tryptic peptides (blue dots).

Criteria (1.8, 2.5, 3.5, 0.08)	IPI tryptic	IPI none full	IPI none full or half	IPI none any
Peptide FPR (%)	4.0%	0.2%	1.7%	6.1%
Protein FPR (%), p=2	5.6%	0.6%	2.4%	16.3%
Protein FPR (%), p=1	46.0%	5.4%	44.9%	77.0%

Conclusions

- Use less redundant protein databases and parsimony filtering.
- Low peptide FPRs do not ensure low protein FPRs.
- A minimum of two peptides per protein eliminates most false positive proteins. Alternatively, methods that estimate the probability of single false positive peptide hits versus multiple true positive peptide hits, such as Protein Prophet and Scaffold can be effective.
- Single peptide/protein identifications often have high protein FPRs and require manual validation, but the actual number of spectra to validate is small since there is only one spectrum per protein.
- The visualization and validation features in Scaffold facilitate efficient manual validation