Methods 61 (2013) 219–226

Contents lists available at SciVerse ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Computer aided manual validation of mass spectrometry-based proteomic data

Timothy G. Curran [a,b], Bryan D. Bryson [a,b], Michael Reigelhaupt [b,c], Hannah Johnson [a,b], Forest M. White [a,b,*]

[a] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Available online 13 March 2013

Keywords:
Mass spectrometry
Tandem mass spectrometry
Protein identification
Protein post translational modification
Computational analysis

ABSTRACT

Advances in mass spectrometry-based proteomic technologies have increased the speed of analysis and the depth provided by a single analysis. Computational tools to evaluate the accuracy of peptide identifications from these high-throughput analyses have not kept pace with technological advances; currently the most common quality evaluation methods are based on statistical analysis of the likelihood of false positive identifications in large-scale data sets. While helpful, these calculations do not consider the accuracy of each identification, thus creating a precarious situation for biologists relying on the data to inform experimental design. Manual validation is the gold standard approach to confirm accuracy of database identifications, but is extremely time-intensive. To palliate the increasing time required to manually validate large proteomic datasets, we provide computer aided manual validation software (CAMV) to expedite the process. Relevant spectra are collected, catalogued, and pre-labeled, allowing users to efficiently judge the quality of each identification and summarize applicable quantitative information. CAMV significantly reduces the burden associated with manual validation and will hopefully encourage broader adoption of manual validation in mass spectrometry-based proteomics.

1. Introduction

Recent advances in mass spectrometry technologies have ushered in a new era of high-content, high-resolution proteomic datasets. Acquisition of hundreds of thousands of tandem mass spectra (MS/MS spectra) in a single analysis is now routine, and by coupling high-speed data acquisition to sample pre-fractionation, millions of MS/MS spectra can be generated from the analysis of a single biological sample. Despite the technological advances that have enabled acquisition of these massive datasets, tools to accurately identify the peptides and post-translational modification (PTM) sites defined by these tandem mass spectra have evolved less rapidly.

In a typical workflow, MS/MS spectra are searched, with database search algorithms such as Sequest [1], MASCOT [2], XTandem [3], or Andromeda [4], against protein databases to generate putative peptide and/or PTM identifications for each MS/MS spectrum. Each of these algorithms relies on a scoring system which weights a variety of parameters, including the mass accuracy of the precursor ion m/z and the percentage and/or sequence of fragment ions matching to the theoretical mass and fragmentation pattern of the putative peptide identification. Due to a variety of factors, including the intensity, peptide sequence (including PTMs), complexity of the sample, and fragmentation method, the MS/MS spectra vary greatly in terms of their quality, as defined by their signal-to-noise and complexity. This variation in quality leads to a wide difference in the searching algorithm scores for putative peptide matches. Currently, there are no set ‘thresholds’ or ‘rules’ for determining whether a particular peptide identification is correct, and each database search algorithm weights aspects of the identification differently. With potentially millions of MS/MS spectra per sample, the challenge of sorting through the putative identifications to determine the accuracy of each assignment is monumental, yet is of utmost importance for correct determination of the components within the biological sample.

The difficulty of accurately identifying a peptide defined by a given tandem mass spectrum can be exemplified when one considers how much weight should be given to mass accuracy of the measured precursor ion m/z compared to the theoretical m/z of the putative peptide. As the accuracy of the measured mass improves, the number of potential peptides from a given database matching to that mass decreases significantly. However, mass accuracy alone is typically not sufficient for identification. Fig. 1 illustrates the issue of relying solely on peptide precursor mass to confirm identification. All tryptic fragments from proteins in the Human 2009 proteome database were in silico digested and binned based on different accuracies. At 10 ppm, very few peptides are the sole occupant of their m/z bin. At 1 ppm, roughly 5% of

[*] Corresponding author. Address: 77 Massachusetts Ave., Bldg. 76-353F, Cambridge, MA 02139, USA.
E-mail address: fwhite@mit.edu (F.M. White).

1046-2023/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ymeth.2013.03.004

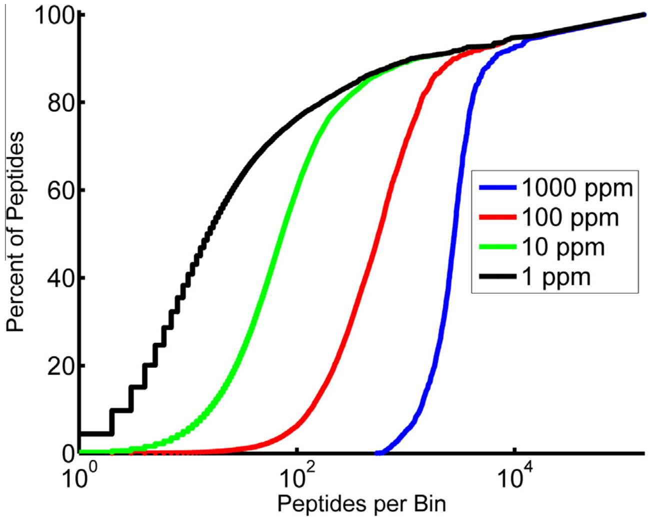

Fig. 1. Sample complexity prevents identification of peptides from complex samples based solely on accurate precursor mass measurements. Precursor masses from all tryptic fragments from the Human 2009 database with charge states +2 to +4 that fall between *m*/*z* 350 and 1500 were considered. Precursor masses were binned into windows generated at four different resolutions. As the number of peptides per bin increases the percentage of total peptides accounted for increases. At 10 ppm a vanishingly small percentage of peptides are the sole occupant of their bin making it nearly impossible to accurately identify peptides based on their precursor mass alone. At 1 ppm this figure is improved to roughly 5%. This problem is exacerbated when post translational modifications and missed or non-tryptic cleavages are included.

peptides are uniquely identifiable from their precursor *m*/*z*. The problem becomes more daunting when the database increases in complexity to reflect the complexity found in biological samples: the database should contain missed cleavages, non-tryptic cleavages, and at least the most common dozen of the several hundred potential post-translational or chemical modifications, as all of these are realistic possibilities for any given peptide. Searches performed against a database of this size and complexity would require massive computational resources. Searching algorithms fight an uphill battle against combinatorial explosion as they seek to balance runtime considerations against erroneous exclusion of relevant peptides. Unfortunately, searching against an incomplete database can lead to false positive identifications, simply because the true assignments are not contained within the search space, and therefore the next best match will automatically be reported. Distinguishing between high scoring false positives associated with the 'next best match' and true positives is critical, especially given the ultimate goal of utilizing these peptide and PTM assignments to inform biological experimental design [5].

1.1. Statistical approaches to assessing quality

Currently, the most common approaches to assessing the validity of a given set of peptide assignments are based on statistical analyses of the likelihood of incorrect assignments, calculated as either a false-positive or false-discovery rate (FDR). In this approach, the MS/MS spectra are searched against a forward database and also against a decoy, reversed or scrambled, protein database [6]. The score thresholds that separate correct from incorrect peptide assignments are then altered to achieve a pre-determined false discovery rate, defined as the quotient of the number of matches in the randomized decoy database to the number of matches in the target database. While this approach can be used to rapidly assess the quality of the overall set of peptide assignments, there are several factors that need to be considered to accurately calculate the FDR. For instance, construction of an appropriate decoy database is crucial, as the distribution of peptide lengths, amino acid composition, and motif prevalence must be tuned to match that of the species database and the specific enrichment experiment performed. In addition, replicate identifications can artificially deflate the FDR if they are considered as independent tests. Several dozen MS/MS spectra might be generated for an abundant species eluting from the chromatography column (these replicate spectra are the basis for the label free spectral counting quantitative approach); it is expected that each of these spectra will match to the same peptide sequence in the forward database. Since these MS/MS spectra are effectively replicates of the same MS/MS spectrum, if one of the spectra does not match to the decoy database, then it is likely that all of the spectra will not match to the decoy database. Instead of considering these as independent tests with several dozen hits with no decoy hits, corresponding to a low FDR, this set of data should be considered as one hit with no decoy hit. Considering these factors should significantly improve the accuracy of the FDR-based statistical estimate of global data quality.

It is worth noting that the FDR-based statistical approach only provides a global quality metric and fails to identify which assignments are true vs. false-positives, leaving doubt about the validity of any given peptide assignment in the dataset. In fact, it is only on further manual inspection that the quality of each peptide assignment becomes evident, even for MS/MS spectra with similar scores for given assignments. For instance, two peptide identifications with similar MASCOT scores are presented in Fig. 2. The confidence in the accuracy of the assignment is much higher for the spectrum in Fig. 2A, as almost all fragment ions match to the expected theoretical fragment ions from the assigned peptide. By comparison, in Fig. 2B there are multiple intense ions that do not correspond to any of the typically theoretical fragment ions for the given peptide assignment, and therefore this MS/MS spectrum is likely to represent either an incorrect assignment or potentially a 'contaminated' spectrum resulting from the simultaneous isolation and fragmentation of multiple ions. In most cases, the database score reflects the number of matched fragment ions, but does not consider the number of unmatched abundant ions, as can be seen by the varied percentage of unmatched ions for similar database scores in Fig. 2C. Based on this analysis, it appears that any global score threshold will automatically include low-confidence identifications, defined as spectra with a fair number of unassigned abundant fragment ions.

An alternative to the FDR approach is to use machine learning-based techniques to automate the validation process. A decision tree validation scheme has been shown to reduce the FDR, yet still relies on searches against a general decoy database [7]. More recently, a hybrid Support Vector Machine (SVM)/Dynamic Bayes Network (DBN) approach was used to classify MS/MS data, and was shown to increase positive identifications in 1% FDR search results [8]. To circumvent the need for large amounts of training data for classification methods, another approach is to create a rule-based framework where prominent fragments are predicted based on expert criteria [9], although codifying experiential human knowledge still limits results to those peptides that match prescribed criteria, therefore hindering generalization to peptides with different PTM's. The "expect" score in MASCOT provides another, peptide sequence specific, alternative to the FDR. This score reflects the probability that a peptide assignment with a given MASCOT score would occur by chance, taking into account the length of the peptide along with the sequences of other peptides in the database to judge the likelihood of proper assignment.

A number of algorithmic approaches have been proposed to automate the proper localization of PTM's, one of the key factors in the quality of an assignment. Among the most widely used of these algorithms, ASCORE defines the PTM site(s) by assignment