Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Article

Whole-Cell Protein Identification Using the Concept of Unique Peptides

Yupeng Zhao and Yen-Han Lin*

Department of Chemical Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada.

Genomics Proteomics Bioinformatics 2010 Mar; 8(1): 33-41. DOI: 10.1016/S1672-0229(10)60004-6

Abstract

A concept of unique peptides (CUP) was proposed and implemented to identify whole-cell proteins from tandem mass spectrometry (MS/MS) ion spectra. A unique peptide is defined as a peptide, irrespective of its length, that exists only in one protein of a proteome of interest, despite the fact that this peptide may appear more than once in the same protein. Integrating CUP, a two-step whole-cell protein identification strategy was developed to further increase the confidence of identified proteins. A dataset containing 40,243 MS/MS ion spectra of *Saccharomyces cerevisiae* and protein identification tools including Mascot and SEQUEST were used to illustrate the proposed concept and strategy. Without implementing CUP, the proteins identified by SEQUEST are 2.26 fold of those identified by Mascot. When CUP was applied, the proteins bearing unique peptides identified by SEQUEST are 3.89 fold of those identified by Mascot. By cross-comparing two sets of identified proteins, only 89 common proteins derived from CUP were found. The key discrepancy between identified proteins was resulted from the filtering criteria employed by each protein identification tool. According to the origin of peptides classified by CUP and the commonality of proteins recognized by protein identification tools, all identified proteins were cross-compared, resulting in four groups of proteins possessing different levels of assigned confidence.

Key words: protein identification, unique peptide, tandem mass spectrometry

Introduction

Mass spectrometry (MS) based protein identification experiments have been the major resource for large-scale proteomic studies of a cell or an organism (*1-5*). Presently, there are numerous protein identification packages available such as MS-Tag (*6*), Mascot (*7*) and SEQUEST (*8, 9*). Reviews on these various protein identification tools were reported recently (*10, 11*).

The critical complexity in protein identification lies in the need to provide confidence levels for the results obtained using the above mentioned tools. A set of positive protein results can help derive accurate conclusions and develop an appropriate plan for further study. However, the practice of using a specific set of MS data to predict several peptides necessitates the separation of the "real" proteins by showing their high confidence. This protracted step is one of the most complicated in protein identification. The major difficulties in using these protein identification tools include multi-identification (*i.e.*, a series of identified peptides may be used to identify two or more pro-

* Corresponding author. E-mail: yenhan.lin@usask.ca

teins), low-confidence identification (*i.e.*, the Mowse score of each peptide is lower than the threshold Mowse score, though the total Mowse score may be greater than the threshold value), and pre-set threshold values used to determine the "true" peptide (*e.g.*, X_{corr} in SEQUEST).

An apparent downside in protein identification using SEQUEST is the determination of the X_{corr} value. Under diverse X_{corr} settings, the searched results, based on the same MS/MS data, may show great variation leading to ambiguity among biological researchers. For instance, from the MS/MS data of *Saccharomyces cerevisiae* (*12*), 1,227 proteins were recognized for X_{corr} value set to 2.0 or greater while only 347 proteins were identified for X_{corr} value greater than or equal to 2.5. These two sets of "identified" proteins were derivatives of the same MS/MS spectral dataset using the same protein identification tool. Consequently, these deviant protein results convey confounding messages to scientists when applied to interpreting phenotypic observations.

Comparisons among various protein identification tools were also reported (*13, 14*). For example, Chamrad *et al* (*13*) applied different protein identification tools to the same set of MS and MS/MS spectral data and observed that only 30%-50% of the results were consistent. This underscores the fact that searched proteins from each protein identification tool generate different confidences, and only those proteins with high confidences can be recognized by these tools. Accordingly, a strategy to analyze the confidence of searched proteins is required.

Based on the concept of unique peptides (CUP) and the cross-comparison among identified proteins, a two-step strategy to study the confidence of whole-cell protein identification was developed in this study. The CUP filters first classify peptides into unique and non-unique clusters, and the step of cross-comparison adds the levels of assigned confidence to proteins identified by means of different protein identification tools. Depending on the accessibility of additional protein identification tools, the proposed dual step approach can be applied independently or in a combined mode. To demonstrate effect of the strategy, two extensively used protein identification packages, namely SEQUEST and Mascot, were employed to identify proteins from publicly available MS data, and the recognized proteins from these tools were investigated using the proposed two-step protein identification strategy.

Results

Concept of unique peptides

A unique peptide is defined as a peptide, irrespective of its length, that exists only in one protein of a proteome of interest, despite the fact that this peptide may appear more than once in the same protein. For example, for Proteins 1 and 2 digested by trypsin, the expected peptides with zero missed cleavage are illustrated in **Figure 1**.

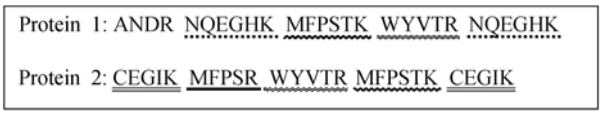

Figure 1 Illustration of the concept of unique peptides.

According to the definition, the peptide ANDR shown in Figure 1 is regarded as unique since it appears once in Protein 1 but not in Protein 2. The peptide NQEGHK is also considered unique based on the same standard. Neither MFPSTK nor WYVTR are unique peptides as they appear in both Proteins 1 and 2. Other unique peptides include MFPSR and CEGIK, found only in Protein 2. The definition of unique peptide is essential in protein identification. It is intuitive to identify Protein 1, if ANDR, NQEGHK, or both are identified. On the contrary, it becomes challenging to conclude whether Protein 1 or 2 exists if only MFPSTK is identified from the MS/MS data. Therefore, a unique peptide can act as a "protein tag" in protein identification.

Whole-cell protein identification

The general procedure implemented in a protein identification tool contains three steps: peptide ranking, peptide filtering and protein identification, which are equivalent to Steps 1, 3 and 4 shown in **Figure 2** (the leftmost column). The MS and MS/MS ion spectra are combined to reconstruct the amino acid sequence of peptides. It is typical that not all experimentally obtained mass spectral data are used during protein