Computational Biology and Chemistry 63 (2016) 21–29

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Protein inference: A protein quantification perspective

Zengyou He[a,b,*], Ting Huang[c], Xiaoqing Liu[a], Peijun Zhu[a], Ben Teng[a], Shengchun Deng[d]

[a] School of Software, Dalian University of Technology, Dalian, China
[b] Key Laboratory for Ubiquitous Network and Service Software of Liaoning, Dalian, China
[c] College of Computer and Information Science, Northeastern University, USA
[d] School of Computer Science and Engineering, Harbin Institute of Technology, China

ARTICLE INFO

Article history:
Received 15 January 2016
Accepted 1 February 2016
Available online 13 February 2016

Keywords:
Shotgun proteomics
Protein inference
Protein quantification
Spectral counting
Linear programming

ABSTRACT

In mass spectrometry-based shotgun proteomics, protein quantification and protein identification are two major computational problems. To quantify the protein abundance, a list of proteins must be firstly inferred from the raw data. Then the relative or absolute protein abundance is estimated with quantification methods, such as spectral counting. Until now, most researchers have been dealing with these two processes separately. In fact, the protein inference problem can be regarded as a special protein quantification problem in the sense that truly present proteins are those proteins whose abundance values are not zero. Some recent published papers have conceptually discussed this possibility. However, there is still a lack of rigorous experimental studies to test this hypothesis.

In this paper, we investigate the feasibility of using protein quantification methods to solve the protein inference problem. Protein inference methods aim to determine whether each candidate protein is present in the sample or not. Protein quantification methods estimate the abundance value of each inferred protein. Naturally, the abundance value of an absent protein should be zero. Thus, we argue that the protein inference problem can be viewed as a special protein quantification problem in which one protein is considered to be present if its abundance is not zero. Based on this idea, our paper tries to use three simple protein quantification methods to solve the protein inference problem effectively. The experimental results on six data sets show that these three methods are competitive with previous protein inference algorithms. This demonstrates that it is plausible to model the protein inference problem as a special protein quantification task, which opens the door of devising more effective protein inference algorithms from a quantification perspective. The source codes of our methods are available at: http://code.google.com/p/protein-inference/.

1. Introduction

Mass spectrometry (MS)-based shotgun proteomics is currently the most widely used method for the identification and quantification of proteins (Nesvizhskii et al., 2007). As shown in Fig. 1, it first digests proteins in the sample into a mixture of peptides by enzymes such as trypsin. The resulting peptide mixtures are scanned by tandem mass spectrometry (MS/MS) to generate a set of MS/MS spectra. Then the peptide identification algorithm reports a set of peptide-spectrum matches (PSMs) by searching the MS/MS spectra against a protein database. From these peptide identifications, we infer the existence of proteins with protein inference algorithms and calculate the relative or absolute abundances of proteins with protein quantification approaches.

Until recently, people tackle the identification and quantification of proteins as two individual and subsequent tasks: first select a subset of proteins that are truly present and then determine the quantities of these proteins. For both problems, many elegant approaches have been developed in the past decades. The readers can refer to two recent reviews Huang et al. (2012) and Nikolov et al. (2012) for details.

The starting point of this paper is fact that protein inference can be regarded as a special case of protein quantification. In protein inference, the objective is to generate a binary presence indicator value (1 or 0) for each candidate protein. In this regard, "protein existence inference" is probably more accurate for describing the original protein inference task. In protein quantification or "protein abundance inference", the goal is to determine the abundance of each protein. Clearly, if one protein is not present, its abundance value should be 0. Hence, the protein inference problem can be

* Corresponding author at: School of Software, Dalian University of Technology, Dalian, China.
E-mail address: zyhe@dlut.edu.cn (Z. He).

http://dx.doi.org/10.1016/j.compbiolchem.2016.02.006
1476-9271/

investigated from the perspective of protein quantification: present proteins are those proteins whose abundance values are not zero. In other words, we can adopt available protein quantification methods directly to solve the protein inference problem. This new angle may enable a better understanding of the protein inference problem and help in devising improved or hybrid protein inference methods by borrowing the power from protein quantification.

The possibility of exploiting protein quantification methods to solve the protein inference problem has been conceptually discussed in several papers (Dost et al., 2012; Li and Radivojac, 2012). Dost et al. (2012) used a simple example to show that it is feasible to obtain more accurate protein identifications with protein quantification methods than traditional parsimonious approaches. Li and Radivojac (2012) also pointed out that the protein inference problem can be regarded as a special protein quantification problem. However, they argued that existing protein quantification methods have not yet reached the accuracy needed for the wide dynamic range of quantities observed in cellular proteomics. As a result, solving the more general and difficult quantification problem may not provide a more accurate solution for the protein inference problem.

Although people have realized the potential of solving the protein inference problem from a quantification perspective, there are still no rigorous and extensive experimental studies to test this hypothesis. To fulfill this void, we empirically demonstrate the feasibility of solving the protein inference problem with existing protein quantification methods in the context of label-free proteomics. In the label-free quantitative proteomics studies, quantification methods which are based on peak ion intensities (from MS data) (Neilson et al., 2011) and spectral counting (from MS/MS data) (Lundgren et al., 2010; Choi et al., 2008) have been widely used.

Spectral counting measures the abundance of each protein based on the number of MS/MS spectra that match its constituent peptides. Given the peptide identification result, we can directly obtain spectral counting information since we just need to count the number of MS/MS spectra. In this paper, we use spectral counting as the quantification approach for solving the protein inference problem.

We first try two simple spectral counting methods in the literature. In both methods, the protein abundance is calculated as the sum of peptide abundance values. Their difference lies in how to handle the shared peptide. If the abundance of one shared peptide is b and it has k parent proteins, then b is used as its abundance value in the first method while b/k is used as its abundance value in the second method. These two methods assume that all the candidate proteins are present in the sample. As a result, the abundance value of each candidate protein will not be zero. However, this assumption contradicts the objective of protein inference: distinguishing present proteins (abundance $\neq$ 0) from absent proteins (abundance = 0). Thus, we extend the second linear programming model in Dost et al. (2012) to distribute the abundance values of shared peptides automatically in order to shrink the abundance values of absent proteins to zero.

To our knowledge, our paper is the first rigorous study with extensive experiments to demonstrate the feasibility of using protein quantification methods for solving the protein inference problem. Such an attempt connects two important computational problems that have long been investigated separately. The experimental results show that we can obtain better performance in most data sets even when the most simple version of spectral counting is utilized. Hence, the advance in protein quantification studies will promote the development of more effective protein inference algorithms.

In Section 2, we describe the details of three methods. Section 3 shows the experimental results on six data sets. Section 4 presents some discussions and Section 5 concludes the paper.

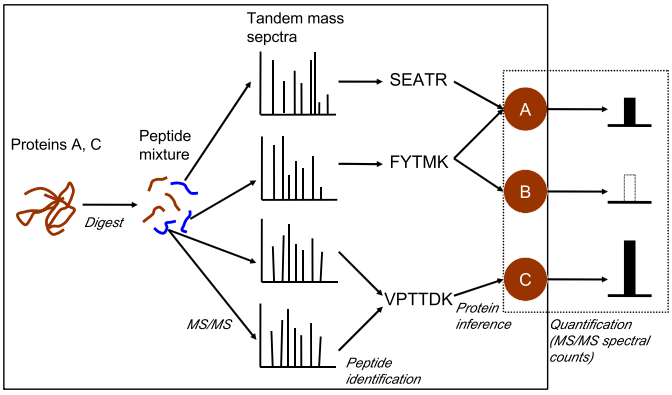

Fig. 1. Protein identification and quantification using mass spectrometry in shotgun proteomics. There are three major computational problems: peptide identification, protein inference and protein quantification.

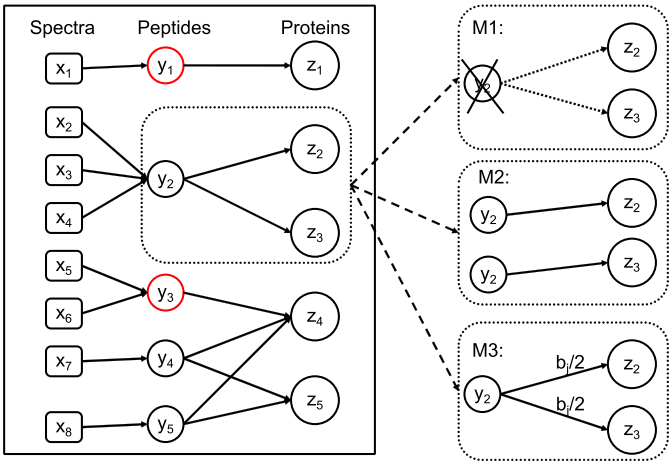

Fig. 2. Three approaches for solving the shared peptide problem. y_1 and y_3 are unique peptides while y_2, y_4 and y_5 are shared peptides. The abundance of peptide y_j is represented by b_j. We use the peptide y_2 as an example to explain how these three approaches work.

2. Methods

As shown in the left side of Fig. 2, the input of the protein inference problem can be represented as a tripartite graph $G = (X \cup Y \cup Z, E_1 \cup E_2)$, where X, Y and Z are the set of l MS/MS experimental spectra, m identified peptides and n candidate proteins, respectively. For all $x_i \in X$, $y_j \in Y$, there is an edge $(x_i, y_j) \in E_1$ if and only if the spectrum x_i matches the peptide y_i in the peptide identification results. Similarly, $(y_j, z_k) \in E_2$ means that the peptide y_j is one part of the protein z_k. Each MS/MS spectrum corresponds to one and only one identified peptide whereas some peptides may have more than one matching spectrum, such as the peptides y_2 and y_3 in Fig. 2. The relationship between peptides and proteins is more complex: one candidate protein may have several identified peptides and each peptide can be shared by multiple proteins. How to correctly distribute these shared peptides is one of the most challenging problem in protein inference.

We first formulate the protein inference problem as a special protein quantification problem. The objective of protein inference is to determine whether each candidate protein is present in the sample. The aim of protein quantification is to estimate the abundance value of each identified protein. Clearly, if one protein is not present in the sample, its abundance value should be 0. In this paper, the protein inference problem is re-visited from the