Journal of Theoretical Biology 362 (2014) 44–52

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Pathway and network analysis in proteomics

Xiaogang Wu [a,b,c], Mohammad Al Hasan [d], Jake Yue Chen [a,b,d,e,*]

[a] Institute of Biopharmaceutical Informatics and Technology, Wenzhou Medical University, Wenzhou, Zhejiang Province, China
[b] School of Informatics and Computing, Indiana University-Purdue University, Indianapolis, IN 46202, USA
[c] Institute for Systems Biology, Seattle, WA 98109, USA
[d] Department of Computer Science and Information Science Purdue University, Indianapolis, IN 46202, USA
[e] Indiana Center for Systems Biology and Personalized Medicine, Indiana University, Indianapolis, IN 46202, USA

ARTICLE INFO

Article history:
Received 1 January 2014
Received in revised form
15 May 2014
Accepted 21 May 2014
Available online 6 June 2014

Keywords:
Pathway analysis
Functional analysis
Hybrid strategy
Network modules
Complex networks

ABSTRACT

Proteomics is inherently a systems science that studies not only measured protein and their expressions in a cell, but also the interplay of proteins, protein complexes, signaling pathways, and network modules. There is a rapid accumulation of Proteomics data in recent years. However, Proteomics data are highly variable, with results sensitive to data preparation methods, sample condition, instrument types, and analytical methods. To address the challenge in Proteomics data analysis, we review current tools being developed to incorporate biological function and network topological information. We categorize these tools into four types: tools with basic functional information and little topological features (e.g., GO category analysis), tools with rich functional information and little topological features (e.g., GSEA), tools with basic functional information and rich topological features (e.g., Cytoscape), and tools with rich functional information and rich topological features (e.g., PathwayExpress). We first review the potential application of these tools to Proteomics; then we review tools that can achieve automated learning of pathway modules and features, and tools that help perform integrated network visual analytics.

1. Introduction

Proteomics, the collective study of all measured proteins in cells of a given condition, is inherently a systems science that requires the understanding of not only the independent parts – protein constituents and their expressions in a cell – but also the interplay of proteins, protein complexes, signaling pathways, and network modules as a whole for achieving biochemical functions. Ideker et al. (2001) introduced an integrated approach to identify metabolic networks and build cellular pathway models, by using measurements from DNA microarrays, protein expressions, and protein interaction knowledge. This work provides systems biology researchers with a practical example how biological networks could be used to perform integrative functional genomics data analysis. By gaining system-wide perspectives of protein functions, Proteomics promises to further study which subsets of proteins are essential in regulating specific biological process. In Proteomics analysis, the incorporating of prior knowledge how groups of proteins work in concert with each other or with other genes and metabolites has made it possible to unravel the complexity inherent in the analysis of cellular functions (MacBeath, 2002). New network biology and systems biology techniques have emerged in recent Proteomics studies (Bensimon et al., 2012; Sabidó et al., 2012) including cancer (Goh and Wong, 2013).

There has been a rapid accumulation of data due to advances in Proteomics technologies (MacBeath, 2002). Proteomics data are often generated from high-throughput experimental platforms, e.g., two-dimensional (2D) gel, liquid chromatography coupled tandem mass spectrometers (LC–MS/MS), multiplexed immunoassays, and protein microarrays (Altelaar et al., 2013; Kingsmore, 2006). These platforms can assay thousands of proteins simultaneously from complex biological samples (Aebersold and Mann, 2003) to measure the relative abundance of proteins or peptides in various biological conditions. More accurate quantitative measure of peptides could also be performed with isotopic labelling of proteins in two different samples (Ong and Mann, 2005). Similar to Genomics, Proteomics studies have been widely used to extract functional and temporal signals identified in biological systems (Blagoev et al., 2004). Popular experimental techniques to measure protein–protein interactions include the yeast two-hybrid (Y2H) system (Ito et al., 2001).

In contract to the recent accelerated application of next-generation sequencing (NGS) in biology, a primary hurdle that slows down Proteomics' applications is the Proteomics data's high

* Corresponding author at: Indiana University School of Informatics & Computing, Indiana Center for Systems Biology and Personalized Medicine, Indiana University - Purdue University Indianapolis, 719 Indiana Avenue, Indianapolis, IN 46202, USA. Tel.: +1 317 278 7604; fax: +1 317 278 9201.
E-mail address: jakechen@iupui.edu (J.Y. Chen).
URL: http://bio.informatics.iupui.edu/ (J.Y. Chen).

http://dx.doi.org/10.1016/j.jtbi.2014.05.031
0022-5193/© 2014 Elsevier Ltd. All rights reserved.

variability, which makes it difficult to interpret Proteomics data analysis results biologically (Colinge and Bennett, 2007). Possible sources of data variations arise from biological sample heterogeneity, sample preparation variance, protein separation variance, detection limits of various proteomics techniques, and pattern-matching peptide/protein identification or quantification inaccuracies from Proteomics data management software. The unusual high level of data noises inherent in Proteomics studies in contrast to those in DNA microarrays or NGS instruments have made Proteomics experiments difficult to repeat, and many statistical methods developed for Genomics applications ineffective. There are plenty of reviews that cover the computational challenges (Vitek, 2009; Noble and MacCoss, 2012; Barla et al., 2008) and solutions to apply statistical machine learning approaches to the problem, e.g., with the use of support vector machines (SVM) (Elias et al., 2004), Markov clustering (Krogan et al., 2006), ant colony optimization (Ressom et al., 2007), and semi-supervised learning (Käll et al., 2007) techniques. The ultimate challenge, however, is how to extract functional and biological information from a long list of proteins identified or discovered from high-throughput Proteomic experiments, in order to provide biological insights into the underlying molecular mechanisms of different conditions (Khatri et al., 2012). Therefore, additional protein functional knowledge, e.g., the abundance of proteins, cellular locations, protein complexes, and gene/protein regulatory pathways, should be incorporated in the second phase of proteomics analysis in order to filter out noisy protein identifications missed in the first statistical analysis phase of Proteomics analysis.

Pathway and network analysis techniques can help address the challenge in interpreting Proteomics results. Analysis of proteomic data at the pathway level has become increasingly popular (Fig. 1). For pathway analysis, we refer to data analysis that aims to identify activated pathways or pathway modules from functional proteomic data. Biological pathways can be viewed as signaling pathways, gene regulatory pathways, and metabolic pathways, all of which are curated carefully in reputable scientific publications. Pathway analysis can help organize a long list of proteins onto a short list of pathway knowledge maps, making it easy to interpret molecular mechanisms underlying these altered proteins or their expressions (Khatri et al., 2012). For network analysis, we refer to data analysis that build, overlay, visualize, and infer protein interaction networks from functional Proteomics and other systems biology data. Network analysis usually requires the use of graph theory, information theory, or Bayesian theory. Different from pathway analysis, network analysis aims to use comprehensive network wiring diagram derived both from prior experimental sources and new in silico prediction to gain systems-level biological meanings (Wu and Chen, 2009). Many large knowledge bases on biological pathways and protein networks have been published, e.g., BioGRID (Chatr-aryamontri et al., 2013), STRING (Franceschini et al., 2013), KEGG (Kanehisa and Goto, 2000), Reactome (Matthews et al., 2009), BioCarta (Nishimura, 2001), PID (Schaefer et al., 2009), HAPPI (Chen et al., 2009), HPD (Chowbina et al., 2009), and PAGED (Huang et al., 2012) databases.

Compared to pathway and network analysis approaches applied in genomics, the advantages of the related researches in proteomics are listed below: (1) Pathway analysis for proteomic data can be directly interpreted in signaling pathways with signal proteins. (2) Network analysis for proteomic data can have direct evidences supported by protein–protein interaction data validated by in-vitro experiments. (3) Both pathway analysis and network analysis for proteomic data can be visualized in a functional protein network with transcriptional factors labeled, which are all measured indirectly in genomic studies.

2. Pathway and network analysis for proteomics

Many pathway databases and pathway analysis software tools have become available in the last decade (Khatri et al., 2012; Ramanan et al., 2012), with some directly applicable to Proteomics (Goh and Wong, 2013; Goh et al., 2012). In Proteomics, statistically significant proteins identified from high-throughput Proteomic instruments often suffer from high false discovery rate (Vitek, 2009), partly because the inherently high level of variance in Proteomics data can make it difficult to identify true biological signals (Noble and MacCoss, 2012). To assess the biological significance of Proteomics results, additional information such as Gene Ontology (GO) and pathways is needed. While there are numerous approaches to incorporate biological pathway and network data into Proteomics data analysis, we categorize existing approaches into two major characteristics, one focusing on integration of "functional information" and the other focusing on integration of "topological information". For functional information, we refer to functional descriptions that aggregate genes into common protein complexes, biological pathways, network modules, and other genes sets consisting of genes playing similar roles. For topological information, we refer to regulatory relationships that exist among

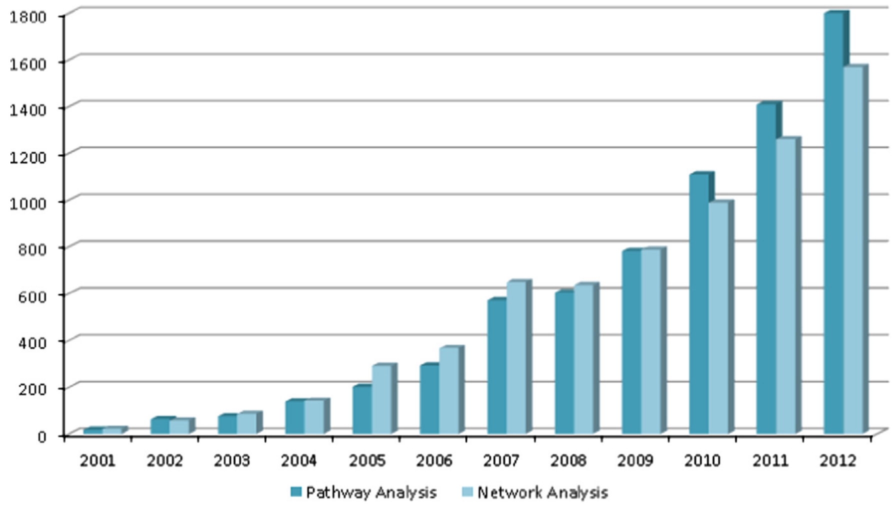

Fig. 1. Trends of pathway and network analysis in Proteomics from decade publications (searched in Google Scholar with terms of ["pathway analysis" AND "Proteomics"], and ["network analysis" AND "Proteomics"]).