Extracting Biological Information from System-scale Protein Interactome Data

(Half-day Tutorial Proposal for ISMB 2003)

Sudhir Sahasrabudhe, Ph.D.
Chief Scientific Officer
Myriad Proteomics, Inc.

Jake Chen, Ph.D.
Head of Computational Proteomics
Myriad Proteomics Inc.

Motivation

The human genome project left us with an enormous amount of data, but little additional knowledge of human health and disease states. The term "**Proteomics**", or the PROTEein complement of the genOME as coined by Marc Wilkins at the 1994 Conference on Genome and Protein Maps, is used to describe the next quantum leap in understanding the participation of proteins in health and disease. Proteomics encompasses more than just identification and quantification of proteins. It also includes their localization, modifications, interactions, activities, and ultimately, their biological function. The Human Genome Project has presented biological researchers with genetic sequence information. The function of a significant portion of such human genetic sequences and genes, however, still remains unclear and therefore their protein products have not been evaluated as therapeutic targets.

A comprehensive human protein interaction map will facilitate identification of proteins that can be targeted for therapeutic and diagnostic applications. By ascribing the role of these proteins into biochemical pathways and networks, and applying appropriate curation techniques, insight into disease processes will be gained. Newly characterized proteins will provide a wealth of pharmacological targets for disease intervention. Understanding biological pathways for normal and disease states will revolutionize medicine by:

- creating opportunities for novel therapies for the treatment and prevention of disease,
- providing tailored therapies for individual patients, and
- accelerating the drug discovery process.

Binary protein interactions are discovered using an industrialized random yeast two hybrid screening workflow. In parallel, the multiprotein complexes (MPC) are isolated and purified by a "pull-down" methodology and proteins are identified by mass spectrometry. The complementary nature of this integrated platform is enormously valuable in providing validation for protein interactions and subsequently mapping proteins to relevant biological pathways and networks. A detailed protein interaction map will result in discovery of new biochemical networks and identification of novel domains.

In addition, these technologies will elucidate the role of proteins in normal and pathological processes and expedite the discovery of novel drug targets and development of new therapeutics. These two high-throughput experimental methods have already produced an influx of tens of thousands of unique protein interactions at Myriad Proteomics, Inc. We want to share our experience with processing, organizing, and analyzing these protein interaction data at system-scale (human interactome data).

Audience

- Computer scientists with some background in genomics research & development, who are interested in proteomics;
- Biologists who have exposure to system-scale data analysis;
- Bioinformatician who are planning to develop processes, syastems, and computational methods in proteomics;
- Biologists and Bioinformaticians interested in extracting disease biology hypotheses from the protein interactome data.

Objectives

We will help audience understand the following topics related to proteomics:

- Experimental methods to create comprehensive human protein interaction data;
- Organization and Discovery of protein interactome data in the context of both system-scale biology and pharmaceutical drug development
- Organization of interacation data into biochemical networks
- Computational infrastructure, processes, and techniques to process, represent, organize, analyze, and visualize high-throughput protein-protein interaction data.
- The nature of “noises” in high-throughput protein-protein interaction data and bioinformatics methods to reduce them.
- Computational models of biological signaling and regulatory pathways by integrating protein-protein interaction data with genome, gene expression and protein expression data.

Tutorial Outline

A. Introduction (Sudhir Sahasrabudhe)

- Introduction to Proteomics
- Protein interaction sub-domain within the proteomics space
- Protein Interaction discovery- biological approaches

B. Biology of Protein-protein interactions (Sudhir Sahasrabudhe)

- Random Yeast 2-hybrid process- Discovery of binary interactions at systems scale
- Tandem-purification mass spectrometry- Discovery of multiprotein complexes at systems scale
- Challenges and opportunities in interactome data
- Case studies: how pharmaceutical target selection could benefit from interactome data

C. Computational Challenges and Solutions (Jake Chen)

- High-throughput Data Processing and Cleaning Concerns
- Representation and organization of protein interaction data
- Protein fragment identity assignment and annotation
- Visualization of protein interaction pairs, clusters, and networks
- Sources of false positives
- Measures of “noises”
- Prioritization methods

D. Protein Interaction networks (Jake Chen)

- Motifs and biological significance
- Computational/statistical network models

E. Application: network overlay methods (Jake Chen)

- Metabolic pathway and protein interaction network
- Gene expressions and protein interaction network
- Gene regulation and protein interaction network
- Extracting disease biology knowledge

Presenter’s Qualifications

Sudhir Sahasrabudhe, Ph.D. Dr. Sahasrabudhe is Chief Scientific Officer and a Director of Myriad Proteomics Inc. Most recently, he was Executive Vice President for Research and Development at Myriad Genetics. He joined Myriad Genetics in August, 2000 from Aventis Pharmaceuticals, where he was senior director of U.S. Biotechnology responsible for all biotechnology activities within the United States. Dr. Sahasrabudhe represented Aventis on the board of the SNP Consortium, Ltd. The SNP Consortium was founded in April 1999 with the aim of discovering Single Nucleotide Polymorphisms throughout the human genome. Dr. Sahasrabudhe also oversaw the Cambridge Genomics Center and other genomics facilities for Aventis, as well as programs in transgenics, gene expression and disease genomics. In his seven years with Hoechst AG, prior to the company's merger with Rhone-Poulenc S.A. to become Aventis, Dr Sahasrabudhe held several important positions, including most recently, Head of the CNS Molecular Biology department and coordinator of the genomics programs. Dr. Sahasrabudhe earned his B.S., M.S., and doctorate degree in microbiology (1987) from University of Baroda, in Baroda, India. He also completed a post-doctoral fellowship in Molecular Biology and Molecular Genetics at the University of Medicine and Dentistry of New Jersey. Dr. Sahasrabudhe serves on the Board of Associates of MIT Whitehead and serves as a member of the Board of trustees of Utah State University Research Foundation.

Jake Y. Chen, Ph.D. holds a doctorate degree in computer science from the University of Minnesota at the Twin Cities and a bachelor degree in biochemistry and molecular biology from Peking University of China. Chen has worked on large-scale high-throughput bioinformatics knowledge discovery research and development projects for

the past six years. His Bioinformatics research interests include biological data modeling, complex database query modeling, and bioinformatics data mining system development. Currently at Myriad Proteomics as the head of computational proteomics group, Chen leads an interdisciplinary team to process, organize, and mine protein-protein interaction data from both random yeast-2-hybrid and tandem-purified mass-spectrometry systems. Prior to joining Myriad Proteomics, Chen worked as a bioinformatics computer scientist at Affymetrix, Inc., Santa Clara, California, where he help developed several Affymetrix commercial cDNA microarrays for the human, mouse, and rat genomes.

Jake Chen has served several non-profit professional organizations to promote public understanding of bioinformatics. He founded Bay Area Young Scientists Forum, a monthly public research lecture series held at Stanford University, and currently serves on the board of Association of Chinese Bioinformatics. While in San Jose, California, He taught several introductory bioinformatics classes at Mos Institute. He has presented talks at many university lectures, professional seminars, and international conferences.