

Tutorial proposal

Title: Computational Methods for Proteomic Analysis

Intended Audience: Introductory. Requires working knowledge, at the advanced undergraduate level, of cell biology (biochemistry and molecular biology), and computation. An introductory level of statistical knowledge is also necessary.

Contact: giddings@unc.edu

Instructors: Michael Giddings, Ph.D. and Michael Wisz, Ph.D.

Length: Half-day

Instructor Backgrounds

Dr. Michael Giddings has been working in the field of computational biology since 1991, and within the proteomics field since 1999. After receiving a Ph.D. in 1997 from the University of Wisconsin, he went to the University of Utah for post-doctoral work under Prof. Ray Gesteland to develop new computational approaches to analyze proteomic data. In January 2002 he joined the departments of Microbiology & Immunology and Biomedical Engineering at the University of North Carolina at Chapel Hill as assistant professor. His group continues work related to proteomics and understanding how protein diversity arises from genes and genomes. His lab has developed several novel tools, including Genome Fingerprint scanning, which is used to identify genes in unannotated genome sequence using protein data (Giddings *et al.*, 2002), and the Protein Cleavage and Modification Engine (PROCLAME) that uses intact-protein mass spectrometry data to identify post-translational modifications (Holmes *et al.*, 2003). He was recently invited to write a chapter on proteome informatics for a popular bioinformatics book by Oulette and Baxevanis, which would be a nice companion to this tutorial, if accepted.

His instructional experience includes designing and teaching a graduate-level bioinformatics class on computational sequence analysis, an introductory programming course ('C' and Fortran), and work as a lab TA for an introductory computer course. He created and organized a bioinformatics seminar from 1999-2001 at the University of Utah, and has given numerous invited talks and lectures, most recently on various aspects of his proteomics work.

Dr. Michael Wisz earned his Ph.D. in Biochemistry from Duke University, where he studied in the laboratory of Dr. Homme Hellinga. While at Duke, Dr. Wisz used computational design techniques to build proteins with novel functions in the laboratory. Part of this work involved the development of a new geometry-dependent electrostatics model that uses empirically-derived parameters. Through these projects, he developed experimental skills in molecular biology and protein characterization as well as computer skills in programming and protein modeling. Dr. Wisz's teaching experience includes a semester as a teaching assistant in a graduate-level Physical Biochemistry course, as well as training others in his lab to use the protein design software he helped develop.

Dr. Wisz is presently working as a post-doctoral scientist in the laboratory of Dr. Michael Giddings at the University of North Carolina at Chapel Hill, where he is developing novel computational methods to analyze complex proteomics data for gene identification in eukaryotes.

Tutorial Motivation

Proteins are nature's machines, responsible for a majority of cellular reactions and processes. In recent years there has been great technological progress in the biochemistry of protein separation and detection. This has led to a concomitant explosion of interest in the field loosely coined as "proteomics". Though differing opinions exist, proteomics is safely described as the comprehensive analysis of proteins in a cell. This kind of analysis may provide valuable new insights into cellular mechanisms, potentially leading to new cures for human disease and increased understanding of cellular processes.

The ability to analyze protein expression on a cell-wide basis grows ever closer with the advances in mass spectrometry, liquid chromatography, protein chips, and gel based separation methods. Computation plays a critical role in proteomic analysis, having the responsibility of linking complex mass spectral data to gene and protein sequence data in a meaningful way. Though useful software tools exist, the informatics lags behind the rapidly advancing chemical and analytical methods in proteomics.

Some analogy can be drawn between DNA/RNA array analysis and proteomic analysis: both aim to use molecular expression data to understand genome function. On the surface it might seem that the challenge is similar: understanding how observed up- or down-regulation of genes relates to a particular cellular state. However, if nucleic acid array analysis isn't challenging enough, proteomics adds several layers of complexity. Proteins are usually more multi-faceted than nucleic acids. The first challenge in proteomics is identification and characterization: unlike DNA array methods, protein identity in a separation is not usually known *a priori*, and must be determined after the fact. This poses challenges that have been the focus of much research. Beyond identification, the analysis of proteins involves other complexities, including: a) differential production and/or modification of proteins dependent on cellular state; b) polymorphism causing differences in the proteins produced; c) complex protein-protein interactions; and d) understanding protein localization to compartments or organelles. Proteomics technology has promised to address each of these, but significant advancement must occur in the handling, storage, and analysis the copious, complex data that is starting to be generated by efforts to study these issues.

Tutorial Goals

The goals of the tutorial are to:

1. Provide an introduction to proteomics: what are the goals and promises
2. Introduce the commonly used computational methods in proteomics (and associated chemistries)
3. Present perspective on the many informatics challenges remaining in the field
4. Interest researchers with bioinformatics skills to work in proteomics

Tutorial Outline

1. Provide an introduction to proteomics and its sub areas
 - a. Discussion about the expanding interest in proteomics – why people are interested and the promises proteomics makes
 - b. Rapid advancements in mass spec and other high-throughput protein analysis methods

- c. Sub areas of proteomics:
 - i. Protein characterization and genomic annotation
 - ii. Expressional studies (protein quantification)
 - iii. Global protein-protein interactions
 - iv. High-throughput structural assessment
 - d. Computation as a critical bottleneck
 - e. Outline the remainder of the tutorial. It will focus primarily on protein expression – the quantification, identification, and characterization of proteins expressed in a cell. 3 parts – proteomics basics, informatics in-depth, the future.
2. Proteomics: The Basics
 - a. Emphasize differences and challenges compared to nucleic acids analysis
 - b. Discuss the general idea: Purify, separate, quantitate, and identify proteins
 - c. Separations: 2D gel electrophoresis (2DGE), HPLC, chips
 - d. Mass spec.
 - i. Types of MS: TOF, Quadrupole, Fourier Transform
 - ii. MS Sources – their differences and uses: MALDI, ESI, ECD
 - iii. Tandem MS: Quad-TOF, TOF-TOF, Data generation
 - e. Quantitation: 2DGE-labelling methods & detection, ICAT
 - f. Cell culture, growth, processing – tissue specific dissection, growth states, quantity and mass spec detection limits
 - g. Emerging efforts in cell-wide protein cross-linking and analysis
 3. Proteomics: The informatics
 - a. Information handling and storage; LIMS, databases, data integration
 - b. Protein cataloging and identification
 - i. Peptide mass fingerprinting – statistics, accuracy, reliability, existing tools
 - ii. Tandem MS of peptides – 2 differing approaches – de novo sequencing vs. database matching: statistics, accuracy, reliability
 - iii. Difficulty using these methods with existing databases: importance of looking at the genomic level and methods for doing so
 - c. Protein Quantification and Comparative Proteomics
 - i. Application of microarray methods to proteomics: the good, the bad and the ugly – noise, standards, comparisons
 - ii. Examples of comparative proteomics in the literature
 - iii. Challenges: post-translational modifications, polymorphism, etc.
 - d. Analysis of cross-linked peptide data – a very brief literature survey
 4. Proteomics: The future
 - a. Top-down, MSⁿ analysis
 - b. Data-driven proteomics: all about integrating information sources into more complete picture of protein character and function
 - c. Chemistry and informatics working together to understand protein state and localization
 - d. Potential integration with microarray and genomic data; challenges
 - e. Identification of alternative protein expression

References

Giddings, M. C., A. A. Shah, R. F. Gesteland and M. Moore (2002). "Genome-based peptide fingerprint scanning." Proceedings of the National Academy of Sciences of the U.S.A. **100**(1): 20-25.

Holmes, M. and M. C. Giddings (2003). "PROCLAME: Protein Cleavage and Modification Engine." Analytical Chemistry **Accepted for publication.**