ISMB2006 Tutorial Proposal

Title: Introduction to Computational Proteomics – Open Problems

Topic Area:

- Proteomics (80%)
- Machine Learning and Artificial Intelligence (10%)
- Signal Processing (10%)

Main Presenter:

- Prof(FH)
- Jacques Colinge
- Upper Austria University of Applied Sciences at Hagenberg
- Hauptstraße 117, A-4232 Hagenberg
- Jacques.colinge@fh-hageneberg.at
- +43 (0)7236 3888 - 2720
- +43 (0)7236 3888 – 2499
- http://webster.fh-hagenberg.at/staff/jcolinge/

- Teaching experience: Computer Programming Teacher (IEPIGE, Geneva), Teaching Assisitant in Maths and Computer Science (University of Geneva), Professor of Bioinformatics (UAS Hagenberg, Austria); mathematics, computer programming, introduction to bioinformatics algorithms, computational proteomics, statistical methods in bioinformatics.
- Computational proteomics, positive feedback of proteomics professionals and interested "newcomers", ECCB'05, Madrid, 2005.

50-word abstract: Proteomics has become an important approach to analyze biological samples. This tutorial will introduce the central problem of searching mass spectrometry data against a database. Quantitative proteomics and peptide de novo sequencing will be covered as well. This presentation should stimulate the interest of bioinformatics researchers in other fields and provide a concise introduction to life scientists.

Tutorial level: Introductory

Prior knowledge required: Elementary knowledge of statistics and bioinformatics.

Suitability of this tutorial for ISMB:

Proteomics data are really diverse dpending on used technologies and asked questions, as is proteomics itself. Therefore the methods employed for data analysis range from algorithms to statistics and, interestingly, they are applied on very specific problems such as signal processing, image analysis, quantitation, MS/MS database search and de novo sequencing, but they also employ more usual bioinformatics methods such as the ones for gene chip data analysis. Sequence database must sometimes be prepared with care and dara redundancy and sequence annotation also play a role in proteomics. For all these reasons I believe that multi-disciplinarity is definitely part of proteomics and my presentation should highlight this fact.

Besides the simple exposition of de facto multi-disciplinarity, my intention is more to provide (1) a comprehensive introduction to (most of) the problems in proteomics data handling, (2) present the main successful approaches to these problems, and (3) to indicate problems where obviously there are nice opportunities for further research.

To summary, I would target an audience of people who want a compact and accessible introduction to the topic and who, eventually, plan to start research in this field.

Profile of Presenter 1

- Interests: proteomics data analysis (identification, quantitation, characterization), statistical methods, integration of proteomics and systems biology, differential equations, parallel computing, teaching.
- Experience:
    - Basic training in maths and computer science, PhD in maths (numerical analysis of strongly nonlinear PDEs, parallel computing).
    - Two years experience as a bioinformatician at Serono dealing with gene expression profiles and data integration.
    - Four years and a half experience as head of mass spectrometry bioinformatics and parallel computing at GeneProt Inc. I worked and supervised many projects; in particular: novel PMF and MS/MS database search engine development, semi- and absolute quantitative methods for differential proteomics with statistical analysis and clustering, data visualization, high-throughput MS identification and data handling, peak detection, 2D gel image analysis.
    - One year experience as a professor where I teach computational proteomics among other topics. Colaboration withGeneBio S.A. that now commercializes the MS/MS search engine developed at GeneProt.

Tutorial Outline:

Part 1 (20 min): Introduction to proteomics. We start by introducing the main problems in proteomics: identify proteins in a sample, characterize modified proteins, compare samples and quantify proteins. We point out the difficulty caused by excessively complex samples with high dynamic range of protein concentrations. We then rapidly introduce the concept of mass spectrometry as an analytical method.

Part 2 (30 min): Peptide mass fingerprinting (PMF) and MALDI instruments. On the basis of the general context presented in Part 1, we introduce and detail a first proteomics method. Show a first example with a spectrum and a database search result. Explain a basic algorithm for searching PMF data against a database of protein sequences. Introduce the notion of scoring function and present classical examples, e.g. MOWSE, ProFound, MSA and OLAV-PMF. State the importance of statistical modeling.

Part 3 (20 min): Peak detection. Raw spectrum processing is rapidly covered to actually link the somewhat abstract mass lists used for searching databases with the signal generated by the MS instruments.

Part 4 (60 min): Complex samples and tandem mass spectrometry. Database sizes and sample complexity may limit the usage of PMF. Tandem mass spectrometry is a manner to obtain additional information via fragmentation. Explain the principle of fragmentation.

Present a schematic abstract mass spectrometer with ion source, fragmentation cell and mass analyzer. Present different technologies (collision induced fragmentation, post-/in-source decay). Explain on-line mass spectrometry.

Several peptide scoring functions are reviewed: MASCOT, SEQUEST, post-processing of SEQUEST, OLAV. The problem of scoring protein identification is then discussed.

Part5 (40 min): Other problems, other approaches. We cover several problems which are of great importance in proteomics today: eukaryote genome searches, peptide de novo sequencing, differential proteomics via quantitative and semi-quantitative methods, protein characterization by top-down techniques.

Discussion (30 min).