Bioinformatics in Genomic and Proteomic Data Analysis

25–27 November 2009
Brno, the Czech Republic

British Embassy
Prague

Bioinformatics in Genomics and Proteomics Data Analysis

Conference Programme

25–27 November 2009
Hotel Voroněž, Brno, Czech Republic

The British Embassy in Prague in co-operation with the Institute of Biostatistics and Analyses of the Masaryk University in Brno

British Embassy Prague

MU IBA

UNIVERSITAS MASARYKIANA BRUNENSIS

Programme

25 November 2009

14.00–15.00	Arrival and registration
15.00–15.10	Official Opening, British Embassy
15.10–15.20	Official Opening, Masaryk University
15.20–16.00	State-of-the-Art Challenges of Genomic and Proteomic data (Torik Ayoubi, VIB MicroArray Facility, K.U. Leuven, Belgium)
16.00–16.20	Coffee break
16.20–17.00	Development of Software Tools for Structural Bioinformatics (Jiří Damborský, Masaryk University, Czech Republic)
17.00–17.30	Education in Analysis of Genomic and Proteomic Data (Eva Budinská, Swiss Institute of Bioinformatics, Switzerland and IBA Masaryk University, Czech Republic)
17.30–17.50	Discussion
19.00–21.00	Welcome Drink

26 November 2009

08.50–09:00	Opening of the second day of the conference
09.00–10:10	On the Causes of Correlations in Affymetrix GeneChip Data (Andrew Harrison, University of Essex, UK)
10.10–10:30	Coffee Break
10.30–11:40	Searching for Differentially Expressed Genes (Eva Budinská, Swiss Institute of Bioinformatics, Switzerland and IBA Masaryk University, Czech Republic)
11.40–12:50	Supervised and Unsupervised Analysis, Building Predictors in Microarrays (Giovanni Montana, Imperial College London, UK)
12.50–13.05	Discussion
13.05–14.20	Lunch
14.20–15.30	Computational Epigenetics (Nuno L. Barbosa-Morais, University of Cambridge , UK)
15.30–16.40	Meta-Analysis for Omics Datasets (Praktyasha Wirapati, SIB, Switzerland)
16.40–17.00	Coffee Break
17.00–18.10	Pathway and Gene Set Analysis of Microarray Data (Claus Dieter Mayer, University of Aberdeen, UK)
18.10–18.25	Discussion and Close
20.30	Conference Dinner

27 November 2009

08.50–09.00	Opening of the last day of the conference
09.00–10.10	Enabling Data Mining: Bioinformatics Resources for Proteomics and Genomics (Rolf Apweiler, European Bioinformatics Institute, UK)
10.10–10.30	Coffee Break
10.30–11.40	Analysis of Mass Spectrometry Protein Data (Jennifer Barrett, University of Leeds, UK)
11.40 – 12.50	Molecular Based Phylogenies: Methods, Theory and Algorithms and the Limit of Their Predictive Power (Gaston H. Gonnet, Swiss Federal Institute of Technology, Zürich, Switzerland)
12.50 – 13.05	Discussion
13.05 – 14.20	Lunch
14.20 – 15.30	The State-of-the-Art in DNA Sequence Analysis (Natália Martínková, IBA, Czech Republic)
15.30 – 16.00	Discussion and Close

Abstracts

The state-of-the-art in genomics and proteomics data analysis and interpretation: challenges and opportunities

Torik Ayoubi, *VIB Microarray Facility, KU Leuven*

All through human history, whether it was the agricultural, the industrial or the digital revolution, major gains in knowledge have always been the driving force for innovation. Such significant gains in knowledge and their application have consistently resulted in increased productivity and wealth. Currently, we are living the genomic revolution which promises to have the most dramatic impact of any previous technological revolution on human civilization, health and maybe even on human nature itself. Genomics research is generating at the moment, on a yearly basis more information than is present in the largest library in the world, The Library of Congress which contains over 140 million books. It should be clear that the trillions of trillions of data points generated by genomics research constitute by far the largest amount of knowledge ever generated by mankind and that it is inevitable that this knowledge will be a major driving force for future, innovation, increase in productivity and wealth. To exploit this knowledge that is about life itself, we will need to develop new data analysis and computational tools, and educate a new generation of scientists to become familiar with the extraction of knowledge from such genomics and proteomics data sets.

Development of software tools for structural bioinformatics

Jiří Damborský, *Eva Chovancová and Antonín Pavelka*
Loschmidt Laboratories, Institute of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czech Republic

Bioinformatics is the information technology applied to the management and analysis of biological data. Analysis of the tertiary structures of biomolecules experimentally determined to the atomic resolution provides insight into the function of living organisms, which is useful for example in designing and optimization of biomolecules for industrial applications. A number of computational tools have been developed for storage and analysis of biomolecular structures in recent years. Yet, new tools are needed for prediction of changes in a structure that will result in modification of the function in controlled manner. Predicted changes, called mutations, can be then introduced to the structures by the molecular biology techniques. The lecture will describe the usefulness of bioinformatics for design and construction of proteins with novel properties. Presented will be a general concept of computer-assisted engineering of proteins as well as two in-house programs CAVER1-5 and HOTSPOT WIZARD6 developed for prediction of mutations leading to improved biocatalysts

On the Causes of Correlations in Affymetrix GeneChip Data

Andrew P. Harrisson, *Director of the Centre for the Analysis of Biological Systems, Departments of Mathematical Sciences and Biological Sciences, University of Essex*

Affymetrix GeneChip technology has proven to be an effective way with which to measure the coexpression of tens of thousands of genes. This has resulted in many thousands of publications, that have detailed expression changes across many tissues, developmental stages, phenotypes and diseases, for most of the major model organisms. Much of the data is now deposited in public resources such as the Gene Expression Omnibus, and our research goals are to make best use of this data. Although GeneChips measure many genes simultaneously, most published studies have only utilised a relatively small number of conditions. This has led to many analysts referring to the "curse of dimensionality", since it is not clear how best to extract statistically significant changes which are also detailing interesting biology. However, the availability of large collections of GeneChip experiments now enable us to begin to overcome this problem, since we have more conditions studied than there are genes.
In order to utilise the GeneChip surveys it is crucial that imperfections in the data are identified, and their impact on any subsequent analysis quantified. Our research is focussed on identifying the causes of outliers in Affymetrix GeneChip data. GeneChips use multiple probes to measure the expression of any one gene. It is vital that false signals within each probe-set are understood, so as to be treated appropriately. We are focussed on finding these spurious signals, and establishing how to correct for the effects of this systematic noise. We are screening GeneChips across a range of scales, from whole experiments, to whole arrays, to spatially adjacent fractions of arrays, to single probes.

Searching for differentially expressed genes

Eva Budinská, *Swiss Institute of Bioinformatics, Switzerland & Institute of Bioinformatics and Analyses, Masaryk University, Czech Republic*

The complete human genome contains thousands of genes – units containing necessary information for the production of proteins or functional RNA molecules (tRNA, rRNA, microRNA...) via the process of transcription (expression) into mRNA and consequent translation into sequence of amino acids that create protein. The genes are expressed and proteins created according to individual needs of a cell or an organism. Not all the genes are active and expressed at the same time. The knowledge of gene expression profile of a sample (cell/tissue/organism) is useful information helping to distinguish between different types of samples. Different expression profiles means different behavior of the cell (tissue/organism) in specified conditions. It is the activation of specific genes that helps the cell to response on different stimuli of the environment, to adapt and survive. In medicine, the gene expression comparison helps to understand the epidemiology and causes of diseases, to discover new disease subtypes and search for new therapeutic targets. The gene expression can be easily measured by DNA microarrays, a powerful tool allowing for simultaneous comparison of thousand of genes in one experiment. This lecture will discuss the common approaches in analysis of differential gene expression such as simple effect size calculation, statistical hypothesis testing and selected regression strategies. Real examples together with description of the most common pitfalls in the analysis will be provided.

Computational Epigenomics

Nuno L. Barbosa-Morais, *Computational Biology Group, Department of Oncology, University of Cambridge,United Kingdom*

Research in epigenetics aims to understand heritable changes in phenotype or gene expression that are not directly associated with changes in the underlying DNA sequence. Experimental techniques such as ChIP-chip started to enable the genome-wide mapping and analysis of large amounts of epigenetic information, thereby allowing for the first snapshots of the epigenome. In particular, the recent emergence of high-throughput sequencing technologies has generated a plethora of epigenomic data that demands novel computational solutions for its processing and interpretation. These are made more challenging by the data¹s diversity: ChIP-seq for transcription initiation, bisulfite sequencing for methylation, RNA-seq for alternative splicing, etc. In this lecture, I will present an overview of the dominant technologies in epigenomics and some important aspects of the analysis of the data they generate. I will also cover some of their applications by describing a few illustrative high-impact research projects.

Meta-Analysis for Omics Datasets

Pratyaksha Wirapati, *Bioinformatics Core Facility Group, Swiss Bioinformatics Institute, Switzerland*

Omics datasets, such as from gene expression and SNP microarrays, are becoming common in biomedical sciences, and many of them are publicly available, potentially allowing more solid conclusions based on larger sample size and consistency of relationships across many contexts. As in classical meta-analysis, there are challenges due to heterogeneities in study designs, methodologies, and so on. Additionally, in omics data such as expression microarrays, there are new issues such as multiple testing, commensurability of measured variables and applications to complex analysis (clustering and prediction). I will outline the basic principles and ranges of methodological approaches in classical meta-analysis and their extension to omics data analysis. Examples will be made using a large collection of expression-clinical studies in breast cancer.

Pathway and Gene Set Analysis of Microarray Data

Claus Dieter Mayer, *Biomathematics & Statistics Scotland, University of Aberdeen, Scotland, UK*

The first step in any microarray analysis tends to be the selection of differentially expressed genes (e.g. between a treatment and a control group), which has initiated a substantial amount of statistical research into various modifications of the t-test and the multiple testing problem arising in this context. During this kind of analysis any knowledge about the annotation or function of the genes is neglected, but only used afterwards, when interpreting the "top list" of most significant genes.
Pathway analysis is an alternative approach that tries to incorporate biological knowledge into the analysis. It utilizes the knowledge that groups of genes are involved in the same pathway or have a similar function and gives a score or p-value to the whole pathway. A group of genes might not necessarily define a biological pathway, so an often use alternative name for this approach is gene set analysis.
In this presentation we will discuss different collections of such gene sets that can be found in internet database like the Kyoto Encyclopedia of Genes and Genomes (KEGG) or the Gene Ontology (GO) database.
We will then introduce various available programs to perform a Pathway analysis, both stand-alone software like GenMAPP or PathVisio but also R/Bioconductor tools like the Globaltest or topGO and particular Gene Set Enrichment Analysis (GSEA) which is available in both forms.
We will particularly discuss the different philosophies behind the various methods using the distinction between "self-contained" (the values of the gene set are compared between treatment and control, neglecting the data of other gene sets) and "competitive" methods (the difference observed in a gene set is compared to the differences in all other gene sets).
We will illustrate methods with real data examples and also outline topics of current and future statistical research in this area.

Enabling Data Mining: Bioinformatics Resources for Proteomics and Genomics

Rolf Apweiler, *European Bioinformatics Institute, United Kingdom*

The completion of the human genome and many other genomes has shifted the attention from deciphering the sequence to the identification and characterization of the encoded components. The identification and functional annotation of the transcriptome, proteome and metabolome is here of special interest and reaches from the identification of genes and transcripts to functional information on many cellular components. Public domain databases like Ensembl, UniProt and many other resources are required to manage and collate this information and present it to the user community in both a human and machine readable manner to enable data mining. My talk will concentrate on the current status of annotating the cellular components, achievements and shortcomings, and future prospects towards a more complete characterization of cellular products, especially in the light of large-scale projects.

Analysis of MASS Spectrometry Protein Data

Jenny Barrett, *Epidemiology Section of Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, University of Leeds, UK*

Mass spectrometry (MS) is a commonly-used tool in proteomics for simultaneously measuring the abundance of many peptides or proteins. The focus of this talk is on quantitative proteomics, where the emphasis is on comparing the proteome across different classes of samples. Issues in experimental design, pre-processing of MS profiles and feature (peak) detection will be discussed. Statistical methods of identifying spectral features that differ across sample groups will be presented and compared, as will methods of classifying a sample on the basis of its entire MS profile. Clinically-motivated examples will be used for illustration.

Molecular based phylogenies: methods, theory and algorithms and the limit of their predictive power

Gaston H. Gonnet, *Swiss Federal Institute of Technology, Zürich, Switzerland*

We discuss the problem of building phylogenetic trees, looking at all the links in the long chain from the data, models, algorithms, validation up to the point that we obtain a final tree. An example of a failure in this process is used to show that our model is not completely accurate, but also possible remedies to this flaw can be identified.

State-of-the-art in DNA sequence analysis

Natália Martínková, *Institute of Biostatistics and Analyses, Masaryk University, Czech Republic*

Amount of sequence data rapidly increases on weekly bases fuelled by intensive genome sequencing projects as well as research of individual laboratories. In this presentation, I will address a principle of analysis of original image files originated from sequencing facilities and their interpretation that includes establishing nucleotide sequence, resolving possible polymorphism and subsequently sequence annotation. Sequence annotation includes identification of functional parts of the genome such as different RNA molecules and especially protein prediction. I will demonstrate that an annotated sequence submitted to public databases can be subsequently utilised for analysis of unknown sequences. Possible origin of such sequences can be identified, and protein function can be estimated based on similarity of query sequence to other available data. Sequence comparison can lead to identification of mutations and their possible links to genetically associated diseases and to reconstruction of evolution of function.