23.1

The PRIDE Database: Giving Researchers Access to Proteomics Data and Providing This Data with a Home

Lennart Martens[1], Henning Hermjakob[2], Chris Taylor[2], Kris Gevaert[1], Joël Vandekerckhove[1], and Rolf Apweiler[2]

[1]Department of Medical Protein Research, Flanders Interuniversity Institute of Biotechnology, Faculty of Medicine, Health Sciences, Ghent University, Ghent, Belgium; and [2]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

The advent of high-throughput proteomics, enabling the ambitious scope of the HUPO projects, has prompted the need for a centralized proteomics data repository. Recently, the completion of the HUPO Plasma Proteome Project pilot phase, the largest proteomics experiment to date, has led to the realization of such a repository. In order to manage and communicate proteomics data from the numerous participating labs across the globe, the European Bioinformatics Institute (EBI) started the PRotein IDEntifications database (PRIDE) project to provide the necessary infrastructure and a suitable set of tools and standards to interface with this infrastructure. The basis of PRIDE is an XML file format that enables standardized communication of results between researchers. XML has the advantage of being readily readable to both humans and machines and, because it is text-based, implicitly delivers cross-platform portability. XML can also be validated against an XML schema for standards compliance testing. From the hierarchical XML schema, a reference relational schema has been created that serves as the queryable data repository. This repository can be accessed via a web application that allows the user to visualize results in either HTML or the PRIDE XML format. The entire PRIDE database is also available for download as a single XML flatfile. Finally an open-source API representing an object model of the PRIDE XML schema is available for manipulating PRIDE data formatted as XML or directly from the relational database schema.

23.2

Phenyx: Combining High-throughput and Pertinence in Protein Identification

Alexandre Masselot[1], Pierre-Alain Binz[1,2], Lorenzo Cambria[1], and Ron D. Appel[1,2,3]

[1]GeneBio SA, Geneva, Switzerland; [2]Swiss Institute of Bioinformatics and Geneva University, Geneva, Switzerland; and [3]Geneva University Hospital, Geneva, Switzerland

For more than ten years, Proteomics has attracted considerable efforts in technology and methodology developments. Dedicated bioinformatics has evolved in parallel to, and wherever possible in interaction with, the tasks of extracting and handling instrumental outputs as well as making possible pertinent biological interpretations. This includes for instance the identification and characterization of protein forms using data generated with mass spectrometry. Even if a number of such software is available today, many end-users are still faced with the problem of handling and comfortably validating huge amounts of data while trying to associate the obtained results with biological pertinence. Phenyx, the result of close collaboration between wet-lab and dry-lab scientists, addresses these issues. The core calculation engine incorporates the true statistical scoring models OLAV developed at GeneProt Inc. that can be fine-tuned for each individual experimental and instrumental MS/MS setup. The specific needs of PMF calculation as implemented are developed in collaboration with the Swiss Institute of Bioinformatics. Both approaches are able to query the detailed annotations of the UniProt-SwissProt database, and therefore search for described alternative splicing events, mutations or post-translational modifications. Using load-balancing technology, Phenyx has shown very-high throughput capabilities while enabling hands-on detailed results management. In particular, a java application has been designed to manage job submission, progress visualization and dynamic results evaluation. End-users can define dedicated profiles, sets of enzymatic cleavage rules or amino acid modifications on one hand, and dynamically navigate through graphical outputs such as spectral interpretation, statistics on jobs or representations of job comparisons. These imbedded functionalities allow therefore users to control and validate the pertinence of their experiments' outputs.

23.3

Computational Heuristics to the Peptide Secuencing Task

J. Alberto Medina, Alberto Paradela, and J. Pablo Albar

Proteomics Facility, Centro Nacional de Biotecnología, CSIC, Madrid, Spain

De novo sequencing is presently one of the main goals pursued by bioinformatic development. This is mainly of sets of peptides which enable the subsequent identification of the proteins they come from. Within this goal, we are developing a new system which enables fast and efficient *de novo* sequencing. The strategy presented ranks and scores the different ions generated by a mass spectrometer by using the top ranked candidates to obtain the peptide sequence.

Traditionally, the question of *de novo* peptide sequencing via tandem mass spectrometry has been tackled by plotting every ion measured on an acyclic graph; thus every node in the graph corresponds to a singly protonated ion from MS/MS fragmentation spectrum and every arc between two nodes corresponds to the mass difference between two ions, provided that this mass difference matches one or several amino acid residues. Consequently, this solution to the sequencing question in this representation is achieved by covering the longest path between the initial and final nodes [1, 2].

The main problems that face this approach are computing effort and the lack of complete sets of fragment ions. Hence, we are attempting to develop a new *de novo* sequencing tool capable of lessening these drawbacks to some extent and at the same time to improve the final output quality of the results.

A two-step procedure was considered for developing this system. Initially a tool for filtering and labelling measured ions was implemented. This tool relies on an intelligent system trained with know spectral data which weights the quality of every ion.

The second phase was aimed at the development and implementation of new algorithms for *de novo* sequencing. These algorithms take that peak weighted in the previous phase as a starting point for attempting to decrease the search space according to a theory about the occurrence of amino acids in the fragmentation spectrum. The output provides a list of sequence candidates ranked by a score based on the quality of the ions selected, the error between theoretical and experimental masses, the error's typical deviation and signal-noise ratio. Despite its simplicity, the scoring method helps understand the relevance of each weighting previously obtained. This tool has been successfully applied for obtaining peptide sequences from protein digests submitted to on-line nano-LC-Ion Trap mass spectrometry analysis providing a correct assignation between spectra and candidate sequences.

[1] Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., Pevzner, P. A. (1999) *J. Comp. Bio*. **3/4,** 327–342

[2] Chen, T., Kao, M. Y., Tepel, M., Rush, J., Church, G. M. (2001) *J. Comp. Bio*. **8,** 325–337

23.4

The InterPro Database

S. E. Orchard, N. Mulder, R. Apweiler, and the InterPro Consortium

EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

InterPro is an integrated resource that provides functional classification of proteins. It allows the user to assign an unknown protein to a family and also identify domains, motifs, active- or binding-sites and potential sites of post-translational modification within its sequence. Each InterPro entry consists of one or more member database signatures, which may be derived from regular expressions and profiles (ProSite), motifs (PRINTS), Hidden Markov Models (Pfam, Smart, TIGRFAM, PIRSF, SUPERFAMILY and CATHSF) and sequence-clustering (ProDom). Each entry contains a detailed, fully-referenced abstract, database links and is mapped to relevant GO terms. Hierarchical relationships link entries and an interactive taxonomy wheel provides a simple view of the taxonomic range associated with sequences within an entry. Different views enable the user to see protein matches to InterPro entries for all proteins found in UniProt and display the structural domains of proteins, as defined by CATH and SCOP using PDB chains. New features include IDA, a domain architecture tool, and a 3-D viewer for protein structures and their signatures.

23.5

An Automated Method for Identification and Quantification of Protein and Peptide Markers in Mammalian Samples Using LC-MS/MS Data

Harald Pettersen[1], David Fenyö[1], Staffan Lindqvist[1], Leo Bonilla[2], Tori Richmond-Chew[2], and Lennart Björkesten[1]

[1]GE Healthcare Bio-sciences, Uppsala, Sweden; and [2]Thermo Electron Corporation, Boston, MA, USA

Creating protein expression profiles and comparing levels of individual proteins e.g. between healthy/diseased tissue samples is a key to a better understanding of the role proteins play in disease. A fully automated LC-MS/MS data analysis method is presented. Software has been implemented to scan for, and further profile, biologically significant peptides and proteins, including tools for visualization, detection, comparison and statistics to simplify the evaluation of large LC-MS/MS data sets for the relative quantification of peptides and proteins. It supports automatic detection and comparison as well as interactive confirmation of the assignments. Available MS/MS data is used for identification of individual peptides/proteins.

2-dimensional representations of individual LC-MS runs are scanned for consistent peptide patterns, which are quantified and matched between the runs. Statistical methods are used to extract peptides showing significant variation among elution profiles and embedded MS/MS data is used for identification of these peptides/proteins.

Here, we apply the algorithms to data obtained by LC-MS/MS analysis of mammalian samples from control/treated experiments with protein digests and native peptides. The samples were analyzed by various LC-MS/MS techniques including linear ion trap. The results show the unique capabilities of the software to extract relevant information from complex biological samples using uni- and multivariate statistics such as t-test, ANOVA, PCA and cluster analysis, and that the data can be used for quantitative analysis.

23.6

Processing of Large LC MS/MS Datasets: Problems and Solutions

Alexandre Podtelejnikov

MDS, Odense, Denmark

With the advent of modern LC MS/MS technology proteomics faces a new challenge—how to transform the huge quantities of produced data (quantity) into valuable biological information (quality). In our hands single biological sample undergo multiple fractionation steps, a series of LC MS/MS analysis and several steps of database searches in order to obtain maximum information from the sample. As a result investigator ends up with dozens of LC MS/MS analysis with several thousands MS/MS spectra each, dozens of thousands of matched peptides and thousands of identified proteins. To deal with so large datasets we have developed several software solutions:

EPIR (Experimental Peptide Identification Repository) is a generic platform based on relational database that allows parsing, storage and mining of LC MS/MS derived peptide evidence. It includes several modules: automatic validation of peptide assignments; grouping proteins with shared peptide evidence; generic extraction of quantitative data; extracting statistics; and a bioinformatics module.

PepSea Inspector—a data validation package for manual inspection of raw spectra based on two independent algorithms: probability scoring and sequence tag approaches.

Iterative database searches—a software solution to enhance the functional information that can be extracted from MS/MS data by applying multiple searches in static and dynamic databases, including alternative splicing, SNPs, ESTs databases and searches with non-limited number of protein modifications.

We will present the described software solutions along with on-going projects in neuroproteomics (analysis of membrane fractions of mouse brain tissues) and subcellular proteomics (analysis of mitochondria fractions) in which the software has been applied.

23.7

UniProt—The Universal Protein Resource

M. Pruess[1], M.-J. Martin[1], C. O'Donovan[1], A. Bairoch[2], C. Wu[3], and R. Apweiler[1]

[1]EMBL Outstation–The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom; [2]Swiss Institute of Bioinformatics (SIB), CMU, Genève, Switzerland; and [3]Protein Information Resource (PIR), National Biomedical Research Foundation, Georgetown University Medical Center, Washington, DC, USA

The Universal Protein Resource, UniProt (www.uniprot.org), which went online in December 2003, is the world's most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

UniProt is comprised of three components, each optimised for different uses. One component, the UniProt Knowledgebase (UniProt), is the central access point for extensive curated protein information, including function, classification, and cross-reference. It consists of Swiss-Prot, a curated, non-redundant protein sequence database, which provides a high level of annotation and integration with other databases, and TrEMBL, its computer-annotated supplement, which contains translations of all coding sequences in the EMBL Nucleotide Sequence Database not yet included in Swiss-Prot. Another component, the UniProt Non-redundant Reference (UniRef) databases, combines closely related sequences into a single record to speed searches. Three different UniRefs are available, based on 100% (UniRef100), 90% (UniRef90) or 50% (UniRef50) sequence identity, respectively. The third component, the UniProt Archive (UniParc), is a comprehensive repository, reflecting the history of all protein sequences. The sequences and information in UniProt are accessible via text search, BLAST similarity search, and FTP.

23.8

The HUPO Brain Proteome Project Pilot Study—Status and Outlook, the Bioinformatics Point of View

K. A. Reidegeld[1], C. Stephan[1], G. Körting[2], C. Scheer[2], R. Reinhardt[2], M. Hamacher[1], H. Thiele[3], R. Apweiler[4], M. Blüggel[2], and H. E. Meyer[1]

[1]MPC, Medical Proteom-Center, Ruhr-University of Bochum, Germany; [2]Protagen AG, Dortmund, Germany; [3]Bruker Daltonik GmbH, Bremen, Germany; and [4]EBI, European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom

The HUPO Brain Proteome Project (BPP) pilot study has started to differentially compare both human and mouse brain samples which are analyzed in 18 laboratories worldwide. The participating labs investigate biopsy and autopsy human brain material as well as three different age stages of mouse brain. To reveal the power of different state-of-the-art protein analysis techniques every lab has the opportunity to use its technique of choice. Besides the differential gel and mass spectrometry analysis also complementary methods as mRNA, DNA and Peptidomic analysis will be applied to give a broader insight into the brain constitution. To handle the immense volume of information from all labs with their multifarious hardware and software equipment the BPP bioinformatics committee has elaborated a two layer client/server architecture based on the proteomics project management software ProteinScape™, a development of Bruker Daltonik and Protagen. In every participating lab (i.e. data producers; DPs) the data of the proteomics workflow (i.e. sample descriptions, MS spectra, gel images, . . .) is collected via the workplace client software and sent to the local ProteinScape server, where a first processing will be performed. After approval project data is sent to the BPP data collection center (DCC) located at the Medical Proteom-Center (MPC). Here all data from the participating labs is collected and in depth analyses (e.g. cross lab comparison) as well as a reprocessing will be executed on PAULA, the high performance Linux cluster of the MPC. The common underlying database scheme of ProteinScape for both DPs and DCC ensures highest grade of data compatibility and excludes operational dependencies. The estimated data submission deadline will be January '05. The global data analysis will be performed by several additionally invited teams with different scientific focus for getting the maximum increase in knowledge from the gathered data. The BPP DCC will be in close contact with the other HUPO data collection centers and the central repository at the EBI. The HUPO BPP will be one of the first major projects to support the upcoming standards MIAPE (Minimum Information About A Proteomics Experiment) and mzData of the HUPO Proteomics Standards Initiative (PSI). Thus the data submission to the DCC and data retrieval from the DCC will be in an open and standardized way allowing upcoming software tools to rapidly get access to the data gathered by efforts of the HUPO Brain Proteome Project.

23.9

Proliferating High Throughput Data Analysis by Using Service Oriented Clusters in Proteomics

Ralf Reinhardt[2], Michael Kuhn[1], Jens Decker[1], Gerhard Körting[2], Martin Blüggel[2], Helmut Meyer[2], and Herbert Thiele[1]

[1]Bruker Daltonik GmbH, Leipzig, Germany; and [2]Protagen AG, Dortmund, Germany

Proteomics data acquisition is rapidly growing, because of advanced MS instrumentation, improved and automated sample preparation (Gel digestion robots, multi-Dim. HPLC) and joined projects like the Proteomics efforts within the HUPO. Despite that the Proteomics community has been hesitant to adopt cluster computing for more complicated analysis. Only in specialized areas like protein identification and multiple alignment cluster computing is used. The reason behind the rejection lies in the kind of clusters available. Some specialize to single applications, leading to efficient but inflexible systems, others facilitate Grid type systems which are powerful but neither easy to learn, nor easy to administrate. The BioClust project proposes a third way called service oriented cluster (SOC). SOCs deliver predefined services through a standardized interface like specialized solutions but can run different applications in parallel and are easily extendable. Administration is separated from application.

For the work on proteomics specific problems a set of services were developed. The tasks include front end data processing like peak detection on newly acquired MS data, single and combined protein searches with several popular engines and protein identification by combining de novo analysis and MS-BLAST. Command line oriented applications like Blast or EMBOSS tools can be added very easily. The services were used both standalone and integrated into Proteinscape™, a database system for proteomics data management (see "The HUPO Brain Proteome Project Pilot Study –Status and Outlook, the bioinformatics point of view," K. Reidegeld *et al.*), for the handling of large datasets. It was shown that BioClust reaches the throughput of specialized systems. It is considered for the reprocessing of HUPO brain proteomic project (BPP) data.

23.10

New Approaches Towards Integrated Proteomics Databases and Depositories

Christian Rohlff

Oxford Genome Sciences, Oxford, United Kingdom

Abstract: Since the determination of the accessible portions of the human genome, two key points have emerged—first, it is still not certain which regions of the genome code for proteins, and second, the number of discrete protein-coding genes is far fewer than the number of different proteins. This talk will highlight the "post-genomic" issues that proteomics is now addressing and will discuss how these data can be integrated effectively with genomics data. Providing effective bioinformatics solution are key for these very complex data. High-throughput and/or high-output technologies create many challenges, including a lack of common protocols, data formats and representation, the inability to access and understand the information created by other people to avoid repetition of their work, facilitate data comparison, exchange and verification. Consequently, for proteomics to continue its current growth rate, there is a need for new approaches to ease data management and data mining. Oxford Genome Sciences has created a new platform that integrates all clinical, experimental information, experimental expression data and has advanced data pipelines for fusing mass-spectrometry data and summarizes and presents them in a biologically relevant manner. The availability of such bioinformatics solutions are crucial for Proteomics technologies to fulfil their promise of adding further definition to the functional output of the human genome. The "Oxford Genome Anatomy Project" or OGAP will provide a framework for integrating molecular, cellular, phenotypic and clinical information with experimental genetic and proteomics data. OGAP's aim is to provide a data integration framework for all protein expression data from different platforms such as one and two dimensional gel electrophoresis, ICAT and other LC based techniques and associated biological/clinical information and reference data about genomes, biological pathways and other relevant information to act as a biological reasoning platform. OGAP's objective is to aide the understanding of the size and diversity of the human proteome at the tissue, disease and protein isoforms levels in a context where it can be readily accessed to biological reasoning. Several models to make OGAP accessible to both academic and commercial R&D will be discussed.

23.11

Deriving Better Specificity Models for Trypsin to Improve Protein Identification by Tandem Mass Spectrometry

F. Schütz[1], E. A. Kapp[2], R. J. Simpson[2], and T. P. Speed[1]

[1]Division of Genetics and Bioinformatics, WEHI, Parkville, Australia; and [2]Joint Proteomics Laboratory, Ludwig Institute for Cancer Research/WEHI, Parkville, Australia

Mass spectrometry is now the method of choice for establishing the identity of a protein from unknown samples. In the bottom-up approach, proteins are digested by an enzyme to produce peptides that are then identified using CID tandem mass spectrometry and database searching. In most cases, the enzyme trypsin is used because the peptides it produces generally fragment in a more predictable manner under electrospray ionization conditions. Trypsin is generally assumed to cleave after Lysine or Arginine residues, except if followed by Proline. Slightly more complicated rules have been devised for predicting trypsin cleavage, however these rules currently only yield binary answers (cleave or no cleavage). Tandem MS database search algorithms use these rules to reduce the number of potential peptides that the algorithm has to consider, thus dramatically reducing the search time.

Using a manually curated database of approximately 12,000 tandem MS ESI-IT spectra (hosted by the Joint ProteomicS Laboratory, Melbourne), we have derived new models for deducing the cleavage specificity of trypsin. Instead of a binary prediction, our models yield a score indicating the propensity for cleavage. Using this model, many cleavages that would be considered as being "missed" by trypsin using the rules described above can actually be predicted.

While these results are interesting in themselves, they can also be used to improve the identification of proteins by tandem MS and database searching. Until recently, the scores calculated by tandem MS database search algorithms were mainly based on the comparison of an experimental spectrum with sequences from the database. Several groups are now incorporating additional experimental information, such as RP-HPLC retention time and/or pI into the results so as to reduce false-positives and increase the number of true-positives. We will demonstrate how our trypsin models can be used in this context.

23.12

Correlation Analysis Between the Breeding Value of Carcass Traits and Spot Intensity on Two-dimensional Gel Electrophoresis of Skeletal Muscle in Hanwoo (Korean Brown Cattle)

K. S. Seo[1], H. B. Yoon[1], D. H. Yoon[1], H. G. Lee[2], and S. H. Kim[3]

[1]Animal Genetic Evaluation Division, National Livestock Research Institute, Cheonan, Korea; [2]School of Agricultural Biotechnology, Seoul National University, Seoul, Korea; [3]Department of Biology, Kyung Hee University, Seoul, Korea

In order to investigate the genetic marker associated with economic performance in Hanwoo (Korean Brown Cattle), proteomic approach was used. Cattles were raised according to the performance test guideline in Korea. Longissimus dorsi muscles were obtained from twelve out of bulls finished progeny test. By two-dimensional electrophoresis, total 123 spots on each gel were detected and compared with the reference gel to be evaluated. Based on the comparison of spot intensity among gels, significant spots were selected and determined the correlation analysis with the breeding value of carcass weight, eye muscle area, and marbling score. The Animal Model with BLUP as a mixed linear model estimated individual breeding value. 9, 16 and 7 spots in high correlation coefficient of above 0.5 were determined on carcass weight, eye muscle area, and marbling score, respectively. Among them, spot #119 in the breeding value of carcass weight was 0.66 in correlation coefficient. For the breeding value of eye muscle area, the correlation coefficient of spot #122 was 0.70. In addition, spot #18 showed 0.58 in that of marbling score. These results suggest that some proteins selected by proteomic analysis would be useful candidate markers for improving the economic performance of cattle.

23.13

Protein Knowledge Meta Model

Amandeep S. Sidhu[1], Tharam S. Dillon[1], Baldev S. Sidhu[2], and Henry Setiawan[1]

[1]Faculty of Information Technology, University of Technology Sydney, Australia; and [2]Board of Studies for Biology, Punjab State Education Department, India

The scope of public protein data sources ranges from the comprehensive, multidisciplinary, community informatics center, supported by government public funds and sustained by team of specialized, to small data sources by individual investigators. The content of protein databases varies greatly, reflecting the broad disciplines and sub-disciplines across life sciences from proteomics and cell biology, to medicinal and clinical trails to ecology and biodiversity. Data elements in public or proprietary protein databases are stored in heterogeneous data formats ranging from simple files to fully structured database systems that are often ad hoc, application specific and vendor specific. Scientific Literature, images and other free-text documents are commonly stored in unstructured or semi-structured formats (plain text files, HTML or XML files, binary files). Information Integration of protein data sources must consider the following characteristics:

1. Diverse protein data are stored in autonomous data sources that are heterogeneous in data formats, data management systems, data schema and semantics.

2. Analysis of protein databases requires both database query activities and proper usage of computational analysis tools.

3. A Broad Spectrum of Knowledge Domains divides traditional Protein Domains in Molecular Biology.

Information Integration in Proteins faces challenges at technology level for data integration architectures and at semantic level for Meta Data specifications, maintenance of data provenance and accuracy, ontology development for knowledge sharing and reuse, and Web representations for communication and collaboration. In this paper, the proposed Semantic Protein Map addresses the following Information Integration Challenges for Protein Data –(1) Semantic Meta –Data integration of PDB, SwissProt and OMIM and (2) Knowledge Sharing & Reuse by using terminology from a shared ontology description, for Web Collaboration.

23.14

A Network of Protein-Protein Interactions in *Leptospira interrogans serovar Lai*

Jingchun Sun[1], Jinlin Xu[1], Qi Liu[2], Tieliu Shi[2], and Yixue Li[2]

[1]School of Life Science & Technology, Shanghai Jiaotong University, Shanghai, China; and [2]Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Leptospirosis is a potentially serious and contagious bacterial illness that affects the liver and kidneys in humans and some animals, which is caused by bacteria of the genus *Leptospira,* a corkscrew-shaped bacterium (spirochete). It is transmitted through contacting with infected urine or water, such as streams used by animals (like rats or goats). With the *Leptospira interrogans serovar Lai* genome *essentially* completed, we can study the organism from a whole-genome standpoint.

Understanding the cellular mechanisms and interactions between cellular components is instrumental to the development of new effective drugs and vaccines. Protein-protein interaction network, involving many aspects of cellular biology, is important to systematically understand the biological mechanisms such as molecular mechanisms of metabolic or signal pathway and the construction of complexes. The availability of increasing completely sequenced genomes makes the *in silico* or experiment based reverse proteomics methods to detect protein-protein interaction popular.

Here we combined four *in silico* methods, phylogenetic profiles method, gene neighbor method, gene fusion and operon method, to construct the protein-protein interaction network of *Leptospira interrogans serovar Lai* and analyzed it in detail with the help of functional annotation of gene products. The genome-wide networks for *Leptospira interrogans serovar Lai* contain 4785 functional linkages of 1345 proteins, including 1039 known proteins and 306 unknown proteins. In terms of the topological characteristics that provide quantitative insight into basic organization, the network is a scale free network: its degree distribution fits the power low distribution ($\lambda=1.47$) and its average connectivity is about 3.56 and its average distance is 6.4. The network includes most functional modules such as cell motility, lipid transport and metabolism, biosynthesis of amino acids, out-membrane proteins and other potential virulence factors, as described in the previously published results (Nature 2003). Proteins LA1001 and LA1002 (Cell Research 2004), a new tonix-antitoxin module, linked to each other in our network. According to the guilt-by-association principle, which states that functions of an unknown protein can be obtained based on its interacted protein(s) of known function, 154 uncharacteristic proteins were assigned functions. Therefore, the network would be expected to aid us, not only to identify possible interacting protein pairs, but also to infer protein functions and new functional modules.

23.15

The Application of Protein-Protein Interaction in the Liver Cancer Research

Yuan-ping Tang[1], Xue-pong Duan[1], Rong Zeng[2], Yi-xue Li[3], and Tie-liu Shi[1,3]

[1]Life Science College, Shanghai University, Shanghai; [2]Proteomics Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai; [3]Bioinformatics Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Hepatocellular carcinoma (HCC) is one of the most common malignancy causing the death in our country. Chronic hepatitis B virus (HBV) or chronic hepatitis C virus (HCV) infection and alcoholic cirrhosis are all prone to result in HCC.

The genesis and the development of the HCC are involved in multiple-factors, multiple phases. One of the most important incident mechanisms is the mutations of the multiple-genes which result in activating mitosis or repressing growth and inactivating apoptosis in liver cell. The protein-protein interaction (PPI) plays the important role in the process of HCC, because current research indicate that most of the human diseases, such as diabetes, cancers, senile dementia etc. result from the dysfunction of the proteins and the changes in the PPI network.

Based on the BIND, DIP, HPRD databases and other information, we have built the PPI network in the human cell. After classifying the proteins with their functions according to GO, we discovered that the PPI network has the same characteristics as the PPI network in yeast cell–functional related proteins forming the module.

Combining the PPI network with the HCC candidate genes from ACEVIEW, we have systematically analyzed the roles of the related genes in the PPI network. Firstly, those genes have been divided into two groups—an up-regulated group and a down-regulated group based on the microarray data and Unigene database; Secondly, those genes are incorporated into the PPI network; lastly, those differential expressed proteins are classified under HCC and normal cell using GO as the standard.

The results from the PPI networks in the two groups showed that the genes related to mitosis and the receptors of the tumor necrotic factors have high connectivity to other proteins, this confirmed that mutations in those genes have the major effect on the cancer cell generation and development. In the tumor tissue, a lot of the high connectivity protein-protein interactions are related to oncogenes such as CCNA2 protein. Interestingly, we have found that several proteins, PLK1, TOP2A, OK/SW-cl.56, have the similar high connectivity as the CCNA2 protein, a lot of the proteins (over 20) they connect with are related to liver cancers. The PLK1 is related to proliferation and has the protein serine/threonine kinase activity; the function of TOP2A is a DNA topoisomerase, and OK/SW-cl.56 is CTP binding, and activates MHC class I protein binding chaperon. Those three proteins share the same expression patterns with the CCNA2 based on the microarray data, therefore, it implies that they could be related to liver caner. We also have found that the ion transport channels in tumor cells and normal cells show great differences, the expression level of many proteins related to ion transport channels (such as sodium, potassium etc.) correlate to their interaction. The expression of those proteins decreases in the tumor tissue. In the contrast, the glutamate gated ion channel activity increase to certain extent. Besides that, the proteins containing the endopeptidase activity decrease dramatically in the cancer cell.

23.16

A System for Proteomic Data Management and Post-planned Analysis

Juhui Wang, Christophe Caron, Alain Trubuil, Michel-Yves Mistou, and Christophe Gitton

INRA, Domaine de Vilvert, Jouy-en-Josas

Although proteomic analysis is intrinsically an iterative and incremental process (information is acquired gradually by researchers in different projects), quite few are current biological data management systems which take account of this reality. Most of them treat the experiment generated data as static and unchangeable: data are never reconsidered, or seldom, whereas technology becomes more powerful or other researchers have brought information on data correction And yet, post-planned analysis which involves multiple iterations and subsequent re-investigations of previously prepared data might bring tremendous benefits.

Named PARIS (Proteomic Analysis and Resources Indexation System), the system we developed here seeks to address this requirement. It automatically takes data from the gel image analysis softwares, and stores the raw and processed data in a relational database suitable for advanced exploration. The system also manages informations about experiments, protein expression and genomic data, and allows the user to search and analyze a large gel collection. It supports visual verification of the analysis results and provides tools for advanced, cross multi-experiment, multi-experimenter data exploration. Implemented in Java, the system is platform independent, accessible to multiple users through Internet. It is also scalable for use for one or many laboratories, and suited to inter-community collaboration work.

AVAILABILITY: PARIS can be tested and downloaded at www.inra.fr/bia/J/imaste/paris.

23.17

Discovery and Identification of Pathogenic Related Proteins in Pathogenic Germs

Y. L. Wang, B. P. Li, L. Wang, J. J. Yue, L. Liang, and P. T. Huang

Department of Microbial Genomics, Beijing Institute of Biotechnology, Beijing, P.R. China

Pathogenic germs often cause various epidemic diseases, which are threatened for human health and lives. More and more researches have been conducted in the discovery and identification of pathogenic related proteins and its function. Although many achievements were achieved through the methods of experiment science, the researches were restricted by the scale. For super-large-scale research, bioinformatics method is desiderated. In this paper, the similarity comparative method based on minus model of sample database was reported, which could efficaciously discover and identify the pathogenic related genes in the super-large-scale. Two test sample databases were constructed, including pathogenic germs and nonpathogenic germs. The databases of pathogenic candidate proteins and orphan proteins were achieved by this method. In order to verify the result's accuracy and the method's correctness, two kinds of proof databases were constructed. The first was the database of virulence factors of pathogenic germs proved by experiment. The second was the type III secretion gene cluster, so the horizontal transfer genes could be verified. In our study, 12,274 pathogenic candidate proteins and 7,100 orphan proteins were discovered, the ratio of accuracy was above 80%. In other words, the similarity comparative method based on minus model of sample database could correctly discover 80% proved pathogenic related genes, and correctly discover the pathogenic related genes obtained by horizontal transfer in substance. In conclusion, the pathogenic related genes were systematic analyzed, the higher ratio of accuracy was obtained. Our work had contributed to decreasing the scale of progress experiment, and to offering the theory for experiment sciences.

23.18

A Visualization and Analysis Tool for Discovery from High-throughput Proteomics

B. M. Webb-Robertson, S. L. Havre, M. Singhal, M. S. Lipton, M. F. Romine, and G. A. Anderson

Pacific Northwest National Laboratory, Richland, WA, USA

Modern biology is interested in understanding the dynamics of a cell at a global level. Of proteomic technologies, mass spectrometry (MS)-based approaches hold the greatest promise for high-throughput applications. This movement to high-throughput is creating a wealth of data on the dynamic nature under which proteins are expressed in a system. At the level of peptide/protein identification the data is typically evaluated using simple spreadsheets that are difficult to manipulate and compare. Of key importance in proteomics is the ability to analyze protein expression between experimental conditions and over time. Thus, the development of techniques and software to support the analysis of MS-based peptide/protein identifications is necessary. We have developed a prototype system call Peptide Permutation and Protein Prediction Tool (PQuad) that is an interactive visualization and statistical analysis tool for understanding relationships in one or more experimental proteomic datasets at a global level. Using PQuad a scientist can explore experimentally identified peptide/protein datasets in the context of existing experimental or predicted biological information associated with the organism under study. PQuad allows the user to observe the results at multiple resolutions –spanning from a chromosomal level down to individual nucleotides and amino acids. In addition to general analysis of a single experimental dataset we have demonstrated the tool on applications such as differential proteomics and gene discovery.

23.19

Evaluation of Within-Disease, or Biological, Variability Using Unsupervised Methods

C. N. White, J. Koopmann, D. W. Chan, and Z. Zhang

Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD, USA

Biological variability is frequently mentioned in cancer research studies, but analysis of within-disease biological variability is rarely completed in concert with the comparison of disease and non-disease groups. Such analysis could be completed using unsupervised analysis methods, which, ironically, are frequently only used to compare the disease and non-disease groups. We hypothesize that identifying similarities among disease samples prior to, or in combination with, the comparison between groups will improve hypothesis development and testing.

To evaluate our novel method of analysis and compare it with existing clustering and unsupervised pattern recognition methods, we analyze the biological variability within benign pancreatic diseases. We re-evaluated a dataset collected for the identification of biomarkers for pancreatic cancer. In addition to pancreatic cancer and normal healthy controls, 60 benign pancreatic disorder samples were collected, including 26 pancreatitis patients and 34 patients with miscellaneous disorders of the pancreas and surrounding GI tract. Using several unsupervised analysis methods, including hierarchical clustering, self-organized maps, and a method developed for this study, we evaluate the accuracy of each method's reported grouping of pancreatitis. The outcome of such unsupervised analysis could help researchers gain new insights into the comparison of disease and non-disease groups.

23.20

Multi-modality of pI Distribution in Whole Proteome, Natural Selection or a Mathematical Fun?

Songfeng Wu[1,2], Ping Wan[1,2], Jianqi Li[1], Dong Li[1], Yunping Zhu[1], and Fuchu He[1]

[1]Department of Genomics and Proteomics, Beijing Institute of Radiation Medicine, Beijing, P.R. China; [2]these authors contributed equally to this work.

Isoelectric point and mass weight are two indexes used to separate different proteins in 2D PAGE. The modalities of pI distribution are usually thought to be caused by the proteins with different subcellular location. Our studies presented a different result that the modalities are the results of the discrete *pK* values for different amino acids. The special amino acid composition and different MW distribution also contribute to the special distribution. The multi-modality feature (including three obvious and one ambiguous modalities) of pI distribution was observed in predicted proteomes of animals, plants, bacterium, archaeas and even in random proteome. Though, subcellular location is related to the pI values, our analyses reveal that neither the multi-modality distribution nor the distribution bias of pI values comparing with the random proteome is caused by subcellular location. The blank neutral region of pI distribution is caused by the absence of amino acids with neutral $pK_{R,}$ which results in the absence of neutral proteins with low MW. One of the possible explanations is that the selection of those amino acids with ionizable side chain might be restricted by the requirement of special pH environment during the origin of life, and they might influence the pH of the solution in cells, too.

23.21

wMOWSE, a Weighted MOWSE Scoring Algorithm for Cross-species Peptide Mass Fingerprinting Database Searching

Heming Xing and John P. Pirro

Bioinformatics Research, Charles River Proteomic Services, Worcester, MA, USA

Peptide mass fingerprinting provides rapid protein identification using mass spectrometer as the primary analytical technique coupled with bioinformatics. This relies on the presence of protein sequence in the current database. As genome sequencing projects continue to add more and more sequence data into the sequence database, proteins from poorly characterized organism will increasingly be identified using cross-species comparison to proteins from well characterized organisms. In this study, the application of cross-species protein identification using peptide mass fingerprinting has been investigated. More than 7000 human/mouse protein orthologous pairs are used to study the performance of cross-species PMF database searching. When sequence identity of ortholog pairs drops below 70% virtually no tryptic peptide of molecular weight between 700 and 3000 were conserved, which is consistent with previous small-scale study. MOWSE program was also tested for the performance of cross-species protein identification using PMF. Mouse proteins are theoretically digested using trypsin and PMF searches are done against human protein sequence database. In 37% of the cases, human ortholog hits are ranked within top 5. As expected in 42% of the cases, human ortholog hits are ranked out of top 50. To improve the performance of cross-species PMF searches, a new algorithm-wMOWSE, which considers the conserved domains, is being developed to recognize homologous proteins across species boundaries. In 71% of the cases, wMOWSE-based PMF search returns the human ortholog hit as No.1, compared to only 23% when using MOWSE-based PMF search. In conclusion, wMOWSE is more sensitive for cross-species PMF search and reduces the rate of false positive hits.

23.22

Classification of Multiple Cancer Types by Using a Two-Stage Multi-Bottleneck-Based Classification Method

Xuejian Xiong, Guojun Yuan, and Kian Lee Tan

Singapore-MIT Alliance, Nation University of Singapore, Singapore

Most gene expression data classification problems belong to the multi-class categorization. The relationships among different classes are complex and implicit, and there are noises in each class. Therefore, a multi-bottleneck-based (TMB) classification method is proposed here to classify multiple cancer types using their gene expression data. There are multiple bottlenecks corresponding to multiple classes, and the bottlenecks are used to cut the relationships among classes. Each bottleneck is the abstract representation of its corresponding class, and preserves the maximum information of the class. Therefore, the multi-bottleneck-based problem can be formalized as that of finding a bottleneck of a class that is totally different from bottlenecks of other classes. The mutual information between bottlenecks should be minimal, while the mutual information between each pair of class and its bottleneck should be maximal. The TMB method is realized by a two-stage way. In its first stage, a class-by-class learning scheme is applied. As a result, the "sub-optimal" bottleneck of each class is generated separately. Due to this independent generation, the relationships among different classes are not accounted for. Therefore, in the second stage, a minimum-mutual-information approach based on the Jensen-Shannon divergence is developed to improve the discriminability of the bottlenecks. Several well-known cancer data sets are analyzed using the proposed TMB method, e.g. the small round blue cell tumors (SRBCT) data, the lymphoma cancer data, etc. In the experiments, the important genes are selected using leave-one-out cross validation. The results show the advantages of the TMB method over previous methods, especially when there are more than two cancer types. For the SRBCT data, the cross validation accuracy reaches 100% when 17 genes are selected.

23.23

Analysis of ”Dead“ Proteome: Identification of Pseudogenes Through Whole-Genome Expression Profiling

Alison Yao[1], Weiniu Gan[1], Rosane Charlab[1], Gennady Merkulov[1], Richard Mural[1], and Peter Li[2]

[1]Celera Genomics and [2]Applied Biosystems, Rockville, MD, USA

A whole-proteome of an organism has a “live” and a “dead” part, i.e. genes and pseudogenes. Defining the number of genes in a genome provides a foundation for the estimation of the “live” proteome size. However, the determination of this is substantially hampered by the misincorporation of pseudogenes into gene collection in addition to the prevalence of alternative splicing. Pseudogenes themselves are important sequences for the study of molecular evolution. This makes it important to correctly identify pseudogenes in gene annotation. Pseudogenes are commonly identified based on structure disablements such as frameshifts or premature stop codons and lack of intron. However, not all pseudogenes have these structure features. It is reported that some pseudogenes have intact coding regions without obvious disablements but do not appear to be expressed. Many pseudogenes, even with structure disablements, have been evidently proved to be transcribed. In addition, a pseudogene in one individual can be functional in another due to polymorphism. Because of these complex characteristics, it is not a trivial matter to distinguish pseudogenes from functional genes. We designed a computational process to systematically detect pseudogenes through the profiling of expressed DNA and protein sequences. This process comprises of three steps: 1). Mapping sequences of ESTs and mRNAs from RefSeq set and GenBank mRNA set onto the genome by blastn/sim4, and then collocating with transcripts and genes. 2). Assigning rankings to the resulting alignments primarily based on sequence identity. This ranking separates out the evidence alignments to a primary hit (best hit) and secondary hits. 3). Constructing an expression profile based on supporting evidence for each gene and calculating the frequency of best hit. Pseudogenes were defined based on a set of criteria through the profiling data. We applied this method to the Celera human and mouse annotation gene set. The resulting pseudogenes had gone through manual curation by expert annotators and proved to be highly accurate. The resulting data in human and mouse were also used to support the probe and primer design of Applied Biosystems gene expression assays and microarrays and significantly increased assay success rate.

23.24

Proteomic Analysis of Normal Human Kidney Glomerulus and Construction of XML-Based Database

Yutaka Yoshida[1], Bo Xu[1], Kenji Miyazaki[2], Ken’ichi Kamijo[2], Akira Tsugita[2], Eishin Yaoita[1], and Tadashi Yamamoto[1]

[1]Division of Structural Pathology, Institute of Nephrology, Graduate School of Medical and Dental Sciences, Niigata University, Niigata, Japan; and [2]Proteomics Research Center, Fundamental and Environmental Res. Labs., NEC Corp., Japan

The proteome of normal human kidney glomerulus was analyzed by 2-DE and identification through MALDI-TOF MS and/or LC-MS/MS. Glomeruli, which were highly purified from kidney cortices with no apparent pathologic manifestation, were separated by 2-DE using 26 cm IPG strips (pH 3–10) and 25×20 cm separation gels. From 2-DE gels of 5 normal subjects, a synthetic gel image was created on which all the identification results were annotated. Nearly 350 protein spots, representing 212 proteins, were so far identified, which were grouped into 18 lager categories on the basis of Gene Ontology (GO) terms. Although most of proteins identified include cell structural proteins, metabolic enzymes, and protein metabolism, significant number of proteins implicated in signal transduction (25), cell cycle and proliferation (10), and stress response (13) were also identified suggesting the usefulness of our database in elucidating biological processes altered under different physiological or pathological conditions. In addition, we have specified proteins abundantly expressed in the glomerulus by statistical analysis of proteins differentially expressed in the glomeruli, cortex and medulla of normal human kidney, and have detected 204 proteins preferentially expressed in the glomeruli. A database of normal glomerular proteome has been constructed by an XML-based editor (HUP-ML) designed for construction of proteome database. The database includes annotations such as protein name and synonyms, accession number of protein database, observed and theoretical pI and Mw, accession number of cDNA database, gene name, GO classifications, and other probable candidates or co-migrated proteins. The database will be submitted on a Web site for public access.

23.25

Merlion: A Proteomics Database Query System

T. You[1*], S. L. Lo[1*], Q. Lin[1], S. B. Joshi[1], M. C. M. Chung[2], and C. L. Hew[1]

[1]Department of Biological Sciences, National University of Singapore, Singapore; [2]Department of Biochemistry, National University of Singapore, MD 7, Level 5, Singapore; *these authors contributed equally to this work

As proteomics research continues to advance to a high-throughput end, involving the use of various instruments and technologies, the complexity to standardize the way to store, exchange and disseminate proteomics data grows exponentially each year. International efforts have been made with several proposals of proteomics data representations emerged. Among them, PEDRo stands out as an excellent data model which systematically represents both methods and experimental results in proteomics. Here we report the development of the Merlion Query System, a customization of PEDRo. This Apache Cocoon based system uses XML, XSL and XSP technologies to allow users to search for proteins by name, SWISS-PROT accession number, protein description, full text, spot serial number, molecular weight and pI, and experiment dates. Query results can be sorted in several ways including the accession number, the spot serial number and the protein name. The modularity of the system allows user to customize the query functions by revising the source code easily. SVG is used to implement gel image display and navigation by clicking on a spot. These features made Merlion a very practical and highly useful data query tool for the proteomics research community.

23.26

The Integrated Proteomics Exploring Database

GuangYong Zheng[1,2*], QuanHu Sheng[2*], HuaYong Xu[3*], ShaoYou Yang[1], BoShu Liu[1], HaiWei Fan[1], Lei Zhang[2], Long Li[2], Hao Tan[1], Chuan Wang[1], JingKang Guo[3], Rong Zeng[2], and YiXue Li[1]

[1]Bioinformatics Center, ShangHai Institute for Biological Sciences, Chinese Academy of Sciences, ShangHai, P.R. China; [2]Research Center for Proteome Analysis, Institute of Biochemistry and Cell Biology, ShangHai Institute for Biological Sciences, Chinese Academy of Sciences, ShangHai, P.R. China; and [3]School of Life Science, ShangHai University, ShangHai, P.R. China; *these authors contribute equally to this work

Proteomics is an increasingly important area of basic and medical research. How to store, manage and exchange huge data, information is urgent for proteomics research. Integrated proteomics exploring database (IPED) was created to provide a solution for these needs. The database consists of three parts: IPED web interface, IPED client software and IPED server software. Important message is distributed through web interface. You can browse special information of proteomics experiment via web interface. As for IPED client software, you can use it to create XML data file quickly for exchange. IPED server software is responsible for data processing, data loading and repository management. General proteomics standard (GPS) and PEDRo model were used as primary rules to construct the database. Sample information, Experiment information, one-dimension (1D) gel image, chemical treatment information, LC column information, two-dimension (2D) gel image, association information obtained by mass spectrometry and identified protein information are stored in the database. IPED has several special features compared with other biological database system: 1 it gives necessary information for repeat proteomics experiment, including information of sample, detail of sample treatment, particular process step and result; 2 Graphic interface is offered so you can browse information mentioned above conveniently; 3 XML format is used to exchange data for the database, so the exchange data file is platform independent; 4 Using this integrated system for proteomics experimental data and information, you can collect experimental data and related information, deal with data and load raw data to repository, and message is released through IPED web interface .In conclusion, the database is efficient for treating proteomics experiment and related information, and providing a initial base for proteomics data mining and analysis.