

28.9

Proteomics Identification of Beneficial Actions of Grape Seed Extract on AD-linked Protein Targets

Jessy Deshane¹, Lisa Chaves¹, Landon Wilson^{1,3}, Marion Kirk^{1,3}, Stephen Barnes^{1,3}, Sreelatha Meleth², and Helen Kim^{1,3}

Departments of ¹Pharmacology & Toxicology and ²Biostatistics, ³UAB Mass Spectrometry & Proteomics Shared Facility, University of Alabama at Birmingham, Birmingham, Alabama 35294

Behavior studies have demonstrated neuroprotective effects of botanical preparations high in anti-oxidants including soy isoflavones and blueberries, however the molecular basis of these effects has not been understood. Recent reports have identified multiple protein alterations in Alzheimer's disease brains, but which are involved in the pathogenesis of the disease is unknown. Using a combination of 2D electrophoresis, MALDI-TOF and LC-tandem mass spectrometry, as well as western blots of 2D gels, we have identified several proteins in rat brain that are altered following ingestion of diets supplemented with grape seed extract. The observed changes included both differential protein expression and alterations in modifications. Statistical analysis carried out on spot intensities without reference to the images confirmed each of the changes. Several of the protein changes detected were previously identified as differentially expressed or modified in AD brain. The changes we noted among these latter proteins were in the opposite direction of the changes noted in AD brain, relative to non-disease human brain. This is the first conclusive identification of specific protein targets for a botanical, as well as the first identification of such targets that are specifically linked to a disease. The fact that the direction of the changes is counter to that observed in disease suggests that ingestion of GSE has specific neuroprotective activity.

28.10

Proteomics Applied to Antibacterial Drug Discovery

Ruth A. Van Bogelen

Molecular Sciences & Technologies, Pfizer Global Research & Development, Ann Arbor, Michigan 48105

In 1975 O'Farrell published a method to separate a complex mixture of cellular proteins such that individual polypeptides could be detected and quantified. Physiologists immediately saw the potential of this global protein profiling method as a means to study physiology at a molecular level. The potential was realized within the first ten years with examples of the classification of proteins into functional groups and into regulatory networks. Recently, it has been demonstrated that changes in protein expression can be correlated with physiological and genetic variation. Despite the success, the work was largely ignored because of the difficulty in determining which gene encoded the polypeptides being cataloged. During the last ten years, methods for global profiling of RNA (immediately linked to their cognate gene) have been developed and are widely used. In addition, mass spectrometry methods have been adapted to handle proteins from gels and has advanced to the point that the 200 most abundant protein spots on the gels can be identified.

Does knowing the genes that encode the molecules monitored with these global expression methods aid in the interpretation of the data? Global profiling reveals cell behavior at a molecular physiological level. How much of the cell's behavior has previously been characterized at the molecular level? Most investigators use only a small portion of the global profiling data. The typical "human response" to a report of a global profiling data set is immediate "data reduction". What and how can we expect global RNA and protein profiling contribute to biology?

29.1

A High Throughput Protein Identification Expert System for 1D PAGE—MALDI-TOF MS Proteome Inventory Technique

Petr Lokhov and Vadim Govorun

Institute of Biomedical Chemistry, Moscow, Russia

Matrix assisted laser desorption-ionization time-of-flight MS is most conventional MS technique for proteomics. The available search systems can significantly identify proteins using MALDI-TOF mass-spectra of their proteolytic digests. However, the prior high performance protein separation (2D-PAGE, multidimensional chromatography) is a necessary condition for automatic high throughput analysis. A substitution of two-dimensional separation for more simple one-dimensional, e.g., SDS-PAGE, meets a need of mass-spectra analysis for proteolytic digest of protein mixture. In view of the fact that the interpretation of mixed digest mass-spectra is time-consuming and labor-intensive even for skilled artisan it is difficult to develop a high throughput system for proteomic study using simple separation methods.

We attempted to analyze MALDI-TOF mass-spectra of protein mixture digests automatically with employment of artificial intelligence (AI) for their interpretation. Mass-spectra analysis results received using Bayesian probability, published MOWSE algorithm, digest pattern estimate, correlations between calculated and received peptide length distribution, as well as protein masses proposed from SDS-PAGE image, were used as input data for AI. Previously manually analyzed 150 protein digest spectra from SDS-PAGE of *H. pylori* proteins were employed as the training database for AI. The study result has shown the capability of MLP topology neuronal network-based AI to identify up to 95% proteins found in spectra previously. We consider that this approach may be used for the development of high throughput systems based on combined use of 1D-PAGE and MALDI-TOF-MS, the approach being able in many cases to substitute more expensive and laborious systems with two- and multidimensional protein separation.

29.2

An Integrated Data Management System for Protein Informatics

Ali Pervez, Dana Heath, Oliver Sauer, Ahmed Elbaggari, Bruce Sadowick, and Tom Slyker

Bio-Rad Laboratories

High throughput proteomics has driven the need for a centralized databases that allows for the storage, management and tracking of data that is generated from the proteomics workflow. The challenge is to be able to integrate disparate and heterogeneous sources of data to derive a biological result. Numerous approaches have been taken with the ultimate goal of being able to develop a truly integrated system, that allows scientist and researchers to seamlessly move between the different data sets that are stored within the database. Bio-Rad's WorksBase Software for proteomics is the first gel based system that allows the seamless integration of image and mass spectrometry information. The primary goal of the software is to make databases more friendly and accessible to biologists. WorksBase works with Bio-Rad's PDQuest software as well as data that is generated from both ms and ms/ms mass spectrometers. Public data bases used for protein identification include SwissProt, TrEMBL and NCBI.

Our Poster will discuss these issues and share our approach.

29.3

PATIKA: An Integrated Visual Environment for Collaborative Construction and Analysis of Cellular Pathways

Emek Demir¹, Ozgun Babur¹, Ugur Dogrusoz², Atilla Gursoy², Gurkan Nisanci², Rengul Cetin-Atalay¹, and Mehmet Ozturk¹

¹Department of Molecular Biology and Genetics and Center for Bioinformatics, Bilkent University, Ankara 06533, Turkey; and ²Computer Engineering Department and Center for Bioinformatics, Bilkent University, Ankara 06533, Turkey

Availability of the sequences of entire genomes shifts the scientific curiosity towards the identification of function of the genomes in large scale as in genome studies. In the near future, data produced about cellular processes at molecular level will accumulate with an accelerating rate as a result of proteomics studies. In this regard, it is essential to develop tools for storing, integrating, accessing, and analyzing this data effectively.

We define an ontology for a comprehensive representation of cellular events. Our ontology enables integration of fragmented, incomplete pathway information and supports manipulation and incorporation of the stored data, as well as multiple levels of abstraction. Based on this ontology, the architecture of an integrated environment named PATIKA (Pathway Analysis Tool for Integration and Knowledge Acquisition) is generated. PATIKA is composed of a server-side, scalable, object-oriented database and client-side editors to provide an integrated, multi-user environment for visualizing and manipulating network of cellular events. This tool features automated pathway layout, functional computation support, advanced querying and a user-friendly graphical interface.

We expect that PATIKA will be a valuable tool for rapid knowledge acquisition, microarray generated large-scale data interpretation, disease gene identification, and drug development. Microarray technology generates gene expression profiles at an unparalleled detail and speed. However the usefulness of this large-scale raw data is limited, unless it is translated into a network of cellular events as provided by PATIKA. Being able to perform complex queries on the pathways, researchers could find drug target candidates and predict potential side effects *in silico*.

29.4

ProteinScape: An Integrated Bioinformatics Platform for Proteome Analysis

Herbert Thiele¹ and Martin Blüggel²

¹Bruker Daltonik GmbH, D 28359 Bremen, Germany; and ²Protagen AG, D 44227 Dortmund, Germany

A software platform ProteinScape has been developed which is able to handle multiple complex Proteome studies. The database system is structured by the central elements biological sample, 2D gel electrophoresis, protein spot, protein identification and post translational modifications. 2D gel images serve as one navigation tool for visualization of the DB content. Special emphasis is laid on sample treatment and mass spectrometric protein identifications. Project specific parameters can be defined freely to customize the platform for user specific needs.

Mass spectrometric data from either peptide mass fingerprinting or from peptide fragmentation fingerprinting experiments at the level of peak lists is imported into the database from MALDI and ESI instruments. MS data is filtered from contaminants, Na⁺/K⁺ adducts, polymers, peak detection errors and is recalibrated by the detected contaminants as an internal standard. ProteinScape triggers different search engines for peptide mass fingerprinting and peptide fragmentation fingerprinting searches against sequence databases and calculates a meta-score value of the different search engine dependent scores. Judging the correct identification can be done fully automated or in manual mode. Sequence database search scenarios combining several search parameter sets can be defined freely.

A flexible retrieval system is developed based on a set of search strategies for samples, mass spectrometric data and proteins. ProteinScape as an integrated database system for bioinformatics establishes links to different knowledge libraries. Cross-project analysis and storage of data as well as the analysis thereof build up valuable inhouse knowledge library.

29.5

PROCSY—PROtein Characterization SYstem

Osnat Sella-Tavor, Assaf Wool, and Zeev Smilansky

Compugen Ltd., Tel Aviv 69512, Israel

Mass-spectrometry in proteomics currently focuses on high throughput protein identification. Protein characterization, however, is still an art. This is due to the great complexity of protein structure. This complexity arises from well-known sources, the main ones being alternative splicing, mutations, cleavages and PTMs. In order to understand protein function and regulation proteome analysis should not be limited to protein identification but should aim at protein characterization.

PROCSY aims at improving this situation by using a new approach. PROCSY utilizes the incomplete protein databases for high-confidence identification of a protein sufficiently similar to the one being characterized, and uses this entry as a template for subsequent characterization. The proteins are digested with several proteases in parallel, followed by PMF analysis using a MALDI-TOF MS. The resulting information can allow prediction of deviations from the database structure, even in cases where such differences were not expected. PROCSY can predict mutations and modifications as well as splice variance and cleavage sites. While the technique of MS analysis with several proteases is not in itself novel, its usage for predictive characterization is a new approach. Huge amounts of raw information are generated, requiring intricate computational analysis to assign a confidence level to the predictions and filter out false positives. Manual analysis of this data is prohibitively difficult. PROCSY software allows this method to be used in a high throughput environment. At the cost of tripling the digestion stage in the standard MS protocol, PROCSY can automatically provide a large amount of important characterization information.

29.6

The Proteome Analysis Database

Manuela Pruess, Alexander Kanapin, Youla Karavidopoulou, Paul Kersey, Virginie Mittard, Isabelle Phan, Florence Servant, and Rolf Apweiler

EMBL Outstation—The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

The EBI's Proteome Analysis Database (<http://www.ebi.ac.uk/proteome/>) has been set up to provide comprehensive statistical analyses of the predicted proteomes of fully sequenced organisms. The analysis is compiled using the InterPro and the CluSTR database as well as GO Slim, and it is performed on non-redundant complete proteome sets of SWISS-PROT and TrEMBL entries. It thus provides a new perspective on families, domains and sites of the proteins from each of the complete genomes. Complete proteome analysis is available for 83 proteomes, spanning eukaryotes, archaea, and prokaryotes. For *Homo sapiens*, two non-redundant proteome sets have been prepared, one of SWISS-PROT + TrEMBL entries, and one of SWISS-PROT, TrEMBL and Ensembl entries. The latter is the combination of the SWISS-PROT and TrEMBL non-redundant human set and additional non-redundant peptides predicted by Ensembl. The current human set contains 33,451 sequences, comprising 25,879 SWISS-PROT/TrEMBL sequences and 7,572 additional Ensembl sequences.

The Proteome Analysis Database also presents structural information for each proteome, including structure information and protein length distribution. Moreover, precomputed comparisons with selected proteomes are provided, and users are enabled to perform their own interactive proteome comparisons between any combination of organisms in the database, to run a FASTA similarity search against a complete proteome, and to download a proteome set or a list of InterPro matches for a given organism. Furthermore, with IPI, the International Protein Index, a top level guide to the main databases that describe the human and mouse proteomes is provided.

29.7

Quantitative Analysis of Two-dimensional Gel-separated Proteins Using Isotopically Marked Alkylating Agents and Matrix-assisted Laser Desorption/Ionisation Mass Spectrometry

Daniela Cecconi¹, Sylvie Gehanne², Lucia Carboni³, Pier Giorgio Righetti¹, Enrico Domenici³, and Mahmoud Hamdan²

¹Department of Agricultural and Industrial Biotechnologies, University of Verona, Verona, 37134, Italy; ²Computational, Analytical and Structural Sciences, Discovery Research, GlaxoSmithKline, Verona, 37100, Italy; and ³Center of Excellence for Drug Discovery in Psychiatry, GlaxoSmithKline, Verona, 37100, Italy

We describe an approach for relative quantification of proteins within a mixture. The method is based on the differential labelling of the mixture by acrylamide and deuterium-labelled [2,3,3'-d3]-acrylamide to alkylate proteins prior to 2D-PAGE. The tryptic digests of proteins were subjected to MALDI-TOF analysis and the relative peak heights of cysteine-containing peptides were used to quantify proteins. This approach was tested for the quantification of proteins within an artificial mixture of standard proteins and for proteins observed in a 2D-map of rat serum. A good correlation was found between the measured ratios derived from MALDI-TOF data and those calculated prior to 2D analysis via known mixing ratios of the two alkylating reagents. This procedure has proved to be effective for comparative measurements of protein abundances.

29.8

A New Approach Based on Fuzzy Logic and Principal Component Analysis for the Classification of 2D Maps: Application to a Lymphoma Case Study

Francesca Antonucci¹, Emilio Marengo¹, Elisa Robotti¹, and Pier Giorgio Righetti²

¹Department of Agricultural and Industrial Biotechnologies, University of Verona, Verona, 37134, Italy; and ²Department of Agricultural and Industrial Biotechnologies, University of Verona, Verona, 37134, Italy

Our method for the comparison of 2D maps, can be summarised in 4 steps: (a) image digitalisation; (b) fuzzyfication of the digitalised map in order to account for the variability of the 2-D PAGE separation; (c) analysis by Principal Component Analysis of the previously-obtained fuzzy maps (for reducing the system dimensionality); (d) classification analysis (Linear Discriminant Analysis), for separating the samples contained in the dataset. The method was applied to 4 maps from reactive human lymph-nodes and 4 from mantle cell lymphoma tissues. The study was performed on different values of the method parameters, in order to investigate the best parameter set. Principal Component Analysis and Linear Discriminant Analysis allowed the separation of the two classes of samples without any misclassification.

29.9

Knowledge Discovery to Understand and Predict Peptide Mass Fingerprinting Spectra

Pierre-Alain Binz¹, Steven Gay¹, Denis F. Hochstrasser^{2,3}, and Ron D. Appel^{1,2,3}

¹Swiss Institute of Bioinformatics; ²Geneva University Hospital; and ³Geneva University

Mass spectrometry is today an almost unavoidable and very efficient tool in proteomics. With the increasing acquisition rate of mass spectrometers, one of the major issues remains the improvement of accurate, efficient and automatic peptide mass fingerprinting (PMF) identification tools. Most of the current tools are scoring protein entries by comparing experimental masses of protein digests with theoretical peptide masses obtained by in-silico digestion of protein sequences. They can also take into consideration experimental pI or Mw, mass tolerance, presence of modified amino acids, among others. However, these identification tools seldom use peak intensities as parameter as there is currently no formal model predicting the intensities based on the physico-chemical properties of peptides. We have used datamining methods such as classification and regression methods to find correlations between peak intensities and the properties of the peptides composing a PMF spectrum. These methods enabled us to obtain decision and model trees that can be directly used for prediction and identification of PMF results. The work performed permitted to lay the foundations of a method to analyze factors influencing the peak intensity of PMF spectra.

29.10

A Proteomic Approach in the Analysis of Thyroid Hyperplastic Goiter

M. Eugenia Schininà¹, Giuseppina Mignogna¹,
Alessandra Giorgi¹, Francesca Cancellario d'Alena²,
Stefania Scarpino², and Maurizio Simmaco²

¹Dipartimento di Scienze Biochimiche 'A. Rossi Fanelli'; and ²II
Facoltà di Medicina e Chirurgia Università La Sapienza, Via di
Grottarossa, 1035-00189 Roma, Italy

The pathogenic mechanisms underlying most thyroid dysfunctions are still largely unknown. Aim of this project is to provide a complete analysis of the thyroid proteome in specific functional and pathological states.

Prior to investigate a specific physiological or pathological effect on protein expression and/or PTMs in a given pathological tissue, it is necessary to establish a resting protein expression profile under controlled conditions. By improving solubilization of protein extracts we were able to produce 2DE gels, from normal thyroid tissue and tissue obtained from patients with hyperplastic goiter, with a very good protein representativity in the entire pH and mass range. Proteins were mainly identified by MALDI-TOF MS. Briefly, stained spots were excised, in gel digested and analyzed by peptide mass mapping. MALDI-TOF spectra were processed via the Data Explorer software. Proteins were unambiguously identified by searching against a comprehensive nonredundant sequence database using the program ProFound.

We can therefore envisage that this approach will be of great help not only for the clarification of the molecular structure of a thyroidal network. The comparison of normal and pathological tissues it will be expected to identify changes in protein expression, localization and post-translational modifications (PTMs), the latter known to regulate protein activity, localization and metabolism.

29.11

Proteome-related Activities of the EBI Sequence Database Group

Alexandra van den Broek and Rolf Apweiler

EMBL Outstation—The European Bioinformatics Institute (EBI),
Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD,
United Kingdom

The EBI Sequence Database Group (<http://www.ebi.ac.uk/npsg/>) is responsible for the production and maintenance of the SWISS-PROT and TrEMBL protein sequence databases, the InterPro protein families and domains database, the CluSTR database, and the Proteome Analysis project.

SWISS-PROT is a curated, non-redundant protein sequence database, which provides a high level of annotation and integration with other databases. TrEMBL, its computer-annotated supplement, contains translations of all coding sequences in the EMBL Nucleotide Sequence Database not yet included in SWISS-PROT. Currently there are 8398 human entries in SWISS-PROT and 50244 in SPTR (data September 2002).

InterPro is an integrated documentation resource for protein families, domains and functional sites. It has already been used for the proteome analysis of a number of completely sequenced organisms including preliminary analyses of the human genome. CluSTR (Clusters of SWISS-PROT+TrEMBL proteins) offers an automatic classification of SWISS-PROT + TrEMBL proteins into groups of related proteins. The Proteome Analysis database provides comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms.

HPI (Human Proteomics Initiative) is a major project by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). It aims to annotate all known human sequences according to SWISS-PROT standards, and contains 113470 entries (data September 2002). (<http://www.ebi.ac.uk/swissprot/hpi/hpi.html>)

29.12

InterPro Database

Maria Krestyaninova¹, Nicola J. Mulder¹, Rolf Apweiler¹,
and InterPro Consortium²

¹EMBL Outstation—The European Bioinformatics Institute (EBI),
Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD,
United Kingdom; and ²International

The exponential increase in the submission of nucleotide sequences to the nucleotide sequence database by genome sequencing centres has resulted in a need for rapid, automatic methods for classification of the resulting protein sequences. There are several signature and sequence cluster-based methods for protein classification, each resource having distinct areas of optimum application owing to the differences in the underlying analysis methods. In recognition of this, InterPro was developed as an integrated documentation resource for protein families, domains and functional sites. The member databases: PRINTS, PROSITE, Pfam, ProDom, SMART and TIGRFAMs form the InterPro core. Related signatures from each member database are unified into single InterPro entries. Each InterPro entry includes a unique accession number, functional descriptions and literature references, and links are made back to the relevant member database(s).

The latest release (release 5.1 July 2002) of InterPro contains 5629 entries describing 4280 families, 1240 domains, 95 repeats and 15 post-translational modifications. Currently the combined signatures in InterPro cover more than 80% of all proteins in SWISS-PROT and TrEMBL, an increase in over 15% since the conception of InterPro. New features of the database include improved searching capabilities and enhanced graphical user interfaces for visualisation of the data. InterPro has been applied to a number of completed genomes. The database is available via a webserver (<http://www.ebi.ac.uk/interpro>) and anonymous FTP (<ftp://ftp.ebi.ac.uk/pub/databases/interpro>).

29.13

Metabolic Network from Proteomic Data Using Graph Theory

A. Joulie¹, C. Gitton², A. Guillot², S. Corteel², M. Y. Mistou²,
and D. Barth¹

¹PRIISM CNRS, UVSQ, 45 avenue des Etats Unis, 78035 Versailles
Cedex, France; and ²INRA, Unité biochimie et Structure des
protéines, Equipe Proteome, 78352 Jouy-en-Josas Cedex, France

The structure of the metabolic network can be described through graph theory. Until now this formalism has been used for networks deduced from genomic data (Jeong et al. (2000)). We apply a similar approach to analyze the data obtained from the proteomic approach. We used data sets obtained from the 2D-gel analysis of a Gram-positive bacteria: *L. lactis* cultivated in various media in which the nitrogen or carbon sources were varied. We find that the metabolic network reconstructed from metabolic enzymes detected on 2D-gels is a small-world network.

This approach allows also to analyze the data. We compare the graphs generated by different experiments. We show that the difference graphs are smaller connected networks. This automatically indicates the metabolic pathways activated in a condition and not in another. We will apply separation algorithms to detect metabolic pathways and compute their interactions using the experimental data.

29.14

KAST: A Protein Network Analysis System for the Spots in Two-dimensional Electrophoresis Gel Image

Min-Seok Kwon^{1,2}, Sang Yun Cho¹, Lang Ho Lee¹,
Soh Yang Ha¹, and Young-Ki Paik^{1,2}

¹Yonsei Proteome Research Center; and ²Department of Biochemistry, Yonsei University

It becomes very important to efficiently manage and analyze all the spot informations obtained from two-dimensional electrophoresis gel. Simple table formats of spot information are not sufficient to represent the inter-relationship between the spots and their biological significance. To this end, we have developed KAST (Knowledge Analysis System for the Spots on Two Dimensional Electrophoresis) by which one can organize any correlated spots as a group and construct a network between spots of interests for further analysis. The data obtained from KAST can be reported in a graphic form or a map type, thereby representing all the spots in the context of correlation family. KAST can also be coupled with YPRC-PDB (Cho et al. (2002) *Proteomics*, in press) or any type of proteome DB in order to construct an integrated proteome DB.

29.15

YPRC-PDB: An Integrated Proteome Database for Two Dimensional Electrophoresis (2-DE) Data Analysis and Laboratory Information Management System

Sang Yun Cho¹, Kang-Sik Park¹, Jung Eun Shim¹,
Min-Seok Kwon^{2,3}, Gil Hong Joo¹, Won Suk Lee¹,
Joon Chang⁴, Hoguen Kim⁴, and Young-Ki Paik^{2,5}

¹Yonsei Proteome Research Center; ²Yonsei Proteome Research Center; ³Department of Biochemistry, Yonsei University; and ⁴Yonsei University College of Medicine

We describe an integrated proteome database, termed Yonsei Proteome Research Center Proteome Database (YPRC-PDB) which can store, retrieve and analyze various information including 2-DE images and associated spot information that were obtained during studies of hepatocellular carcinoma (HCC). YPRC-PDB is also designed to perform as a laboratory information management system that manages sample information, clinical background, and conditions of both sample preparation and 2-DE, and entire sets of experimental results. It also features query system and data-mining applications, which are amenable to automatically analyze expression level changes of specific protein and directly link to clinical information. The user interface is web-based, so that the results from other laboratories can be shared effectively. Especially, the master gel image query is equipped with a graphic tool that can easily identify relations between the specific pathological stage of HCC and expression levels of potential marker protein on the master gel image. Thus, YPRC-PDB is a versatile integrated database suitable for subsequent analyses. The information in YPRC-PDB is updated easily and it is available to authorized users on the World Wide Web (<http://yprcpdb.proteomix.org/~damduck/>).

29.16

Advanced Proteome Data Analysis for High Throughput Proteomics—Towards a Brain Proteome Database

Martin Blüggel¹, Johan Gobom², Andreas Wattenberg¹,
Sonja Bailey¹, Christian Scheer¹, Gerhard Körting¹,
Daniel Chamrad¹, Helmut E. Meyer¹, and Joachim Klose³

¹Protagen AG, Emil-Figge-Str. 76A, 44227 Dortmund, Germany;
²Max-Planck-Institute for Molecular Genetics, Ihnestraße 73, 14195 Berlin, Germany; and ³Charité, Institute for Human Genetics, Augustenburger Platz 1, 13353 Berlin, Germany

Automation has improved dramatically within the last year in different Proteomics techniques. Multidimensional LC-ESI-MS/MS systems are used for shotgun proteomics in protein and PTM identifications. MALDI-mass spectrometry has overcome the bottleneck of high throughput Peptide Fragmentation Fingerprinting (PFF) due to new MALDI-TOF-TOF instrumentation. Additionally, laboratory automation for 2D gel based Proteomics have improved with automated spot picking, protein digestion and sample preparation for MALDI-MS. On the other hand bioinformatics for Proteomics is still at the very beginning. Managing Proteome data, enhanced mass spectra interpretation algorithms, correlations within a Proteome study and correlation to public knowledge libraries, is a mayor task in high throughput Proteomics. We developed an integrated database system with advanced algorithms for mass spectrometry data interpretation for these automated technologies which allows to manage, interpret and correlate Proteome data. We apply these techniques to a mouse brain Proteome project, in which we combined highest resolution 2D gel electrophoresis with automated and high sensitive protein identification techniques. The identified proteins are classified in regard of their participation in biological processes, their cellular distribution and their molecular function according the gene ontology classifications. This classification gives insights of the part of proteome studied today by high resolution 2D gel electrophoresis. Identification of significant over- and underrepresentation of classified proteins is done by comparison to a reference database. These results are essential for 2D gel based proteome study design e.g. for proteomics driven drug target discovery.

29.17

Comparison of Mass Spectral Data Search Results for Several Search Programs Using EST Database

Kyung-Hoon Kwon, Seung Il Kim, Mioak Kim, Jong Shin Yoo, and Young Mok Park

Korea Basic Science Institute, 52 Yeoeun-dong, Yusung-gu, Daejeon, South Korea

Even we are living in post-genome era, EST databases are still popular to study proteomics for identifying proteins. There are several database search programs for identifying proteins by comparing Peptide Mass Fingerprinting (PMF) as well as tandem mass spectra generated from mass spectrometers with peptide molecular weights of EST databases. In this study, we will compare popular search programs, such as MS-Fit, MS-Tag of ProteinProspector of UC San Francisco, MASCOT Search of Matrix Science Ltd., and TurboSEQUENT of Thermo Electron Co., to find out the characteristics of each program for identifying proteins using PMF and tandem mass data and EST databases. We found EST sequences which had got the highest score by the search programs from each mass spectrum produced by ESI/Q-TOF/MS and MALDI/TOF-TOF/MS. As a method for evaluating the programs, we used the peptide sequences calculated by De novo sequencing as a reference and compared it with EST sequences which are obtained from the search programs.

29.18

Beyond Data: DiBase™ Bridges the Information Gap—Getting the Picture, not the Pixels

Wafik Farag

SkyPrise, Inc.

Despite data integration efforts, solutions to manage genomics, proteomics, combinatorial chemistry, etc. sources still fall well short. The bottleneck is looking only at data. DiBase™ is an information platform empowering scientist global access to share and control informatics—not only data, but also the sources of data, the process to glue applications and custom functions together to obtain a result, as well as storing those results “on the fly”—eliminating the need for costly and time-consuming schema redesigns. The key is moving integration logic outside data layer into a function layer. Will examine practical examples in building and sharing of such processes, i.e., “data-flow-path,” searching and re-creation of stored results, illustrating the strength of utilizing all of information building blocks. What makes this approach so vastly more cost effective is allowing collaboration via a web-based platform leveraging the minds of many for both short and long-term strategies.

29.19

Using Image Fusion and Positional Indexing to Create Proteome Maps from Collections of 2D Gel Images

Sven Luhn¹, Matthias Berth¹, Michael Hecker², and Jorg Bernhardt^{1,2}

¹DECODON GmbH, Greifswald, Germany; and ²Institute of Microbiology and Molecular Biology, University of Greifswald, Greifswald, Germany

We present an image processing technique that produces comprehensive and accurate visual maps of a proteome's subset that is visible on 2D electrophoresis gels. These maps are annotated, realistically-looking synthetic gel images where each spot as a standard position that can be used as a unifying index for the whole collection. They are complete in the sense that every spot that is visible on any of the gels will be shown on the map. The map shows spots in a rather realistic shape and size, making it well-suited for visual comparison and automated image registration. The standard positions give a much more accurate estimation of a spot's position on a gel than those obtained using theoretical isoelectric point and molecular weight. Due to the high accuracy, positional indexing can be used as a complement to a-priori identifications (such as MS or Edman degradation). Linking spots to their position on the map makes it possible to create expression profiles that span the entire collection, an essential requirement for large-scale mining of protein expression data.

29.20

Laboratory Workflow System (LWS) for Proteomics, a Database System Managing Samples from 2D Electrophoresis to MALDI-MS Protein Identification Results

Anneli Jorsback, Margareta Degerman, and Gunilla Jacobson

Amersham Biosciences, Björkgatan 30, SE-751 84 Uppsala, Sweden

The development of more accurate and higher throughput methods for protein identification and expression increases the need for tools for laboratory data management and accessibility to bioinformatics queries. Proteomics analysis is going from manual to automatic processing and is reaching an audience of new researchers that are non-expert user.

The Laboratory Workflow System (LWS) provides software tools to collect, track and direct processes through laboratory activities. The workflow itself mirrors the activities associated with the experimental process: work request, receipt of sample, two dimensional electrophoresis, image analysis, spot handling, MALDI-MS analysis, and protein identification. The results are maintained in a relational database management system (RDBMS) and the data is available for web-based retrieval to perform subsequent analysis and export to other formats.

The software supports different user roles. Project planning and reporting are performed in the web application. Assistance to monitor in-process activities and project status is given, protocols, consumables, and instruments are managed, and the detailed experimental design for each sample is set up in the management workbench. In the applications module the navigator tool assists the research associate while progressing from one activity to another recording every step in the electronic laboratory note-book. By using barcode technology including barcoded IPG strips and gels the risk of mix-up of samples is minimized and samples are traced throughout the entire workflow.

The use of the LWS is demonstrated in some examples showing *E. coli* and mouse liver projects where the link from sample to gel spots to protein identity is demonstrated.