CONFERENCE REPORT

IDrugs 2008 **11**(11):
© The Thomson Corporation ISSN 1369-7056

From Genome to Proteome: Integration and proteome completion – 8th Siena Meeting

31 August - 4 September 2009, Siena, Italy

Åsa M Wheelock[1] & Craig E Wheelock[2]

Address

[1]Karolinska Institutet, Lung Research Laboratory L4:01, Division of Respiratory Medicine,
Department of Medicine, SE-171 76 Stockholm, Sweden
Email: asa.wheelock@ki.se

[2]Karolinska Institutet, Division of Physiological Chemistry II, Department of Medical Biochemistry and Biophysics,
Scheeles väg 2, SE-171 77 Stockholm, Sweden
Email: craig.wheelock@ki.se

Correspondence can be addressed to either author

Introduction

The Siena Meeting is designed to cover the breadth of proteomics-based research, with a particular focus on clinical applications as well as emerging technologies. As in previous years, the 2008 meeting attracted a mix of academic, clinical and industrial scientists, providing a unique opportunity for interactions among these researchers and an excellent environment for research discussions and the formation of new collaborations. The coverage of topics ranged from a bioinformatics recreation of the human proteome to the simultaneous quantification of changes in an entire proteome. Specifically, several technological advances in proteomics, with particular focus on proteome quantification methodologies were addressed. Current bottlenecks in the field were also examined, including biomarker verification strategies and the development of bioinformatics resources.

A complete human proteome

Denis Hochstrasser (University of Geneva) opened the meeting by recapitulating the evolution of proteomics research, with specific focus on the huge advances made since the inception of the Siena Meetings in 1994. This progress was emphasized during the initial plenary session in a presentation entitled 'A complete set of annotated human proteins in UnitProt/SwissProt' given by Amos Bairoch (Swiss Institute of Bioinformatics). At the Fourth Siena Meeting in 2000, Dr Bairoch had discussed the launch of the Human Proteome Initiative (*www.expasy.ch/sprot/hpi/*), the goal of which is to annotate all human protein sequences as well as their mammalian orthologs. The annotated human proteome in SwissProt, consisting of 20,325 human protein entries from 20,400 human protein coding genes, was completed in September 2008, with the caveat that many genes encode for the same protein. In the context of the annotation, the word 'complete' was defined as the annotation of all protein-coding genes with a Human *Genome* Organisation Gene Nomenclature Committee (HGNC) gene symbol as well as all proteins in databases and identified through literature mining. The breakdown of the complete proteome was presented in terms of documented evidence at the protein level (11,453; 56.3%) and the transcript level (8068; 39.7%), inference from homology (273; 1.4%), and in terms of proteins that are predicted to exist (140; 0.7%) as well as proteins with uncertain sequences (391; 1.9%). Although this progress is promising, it is important to note that these annotations represent only the beginning of the work of the Human Proteome Initiative. An important task is to examine post-translational modifications (PTMs), SNPs and other distinguishing characteristics. For example, over 46,000 SNPs have been identified by the International HapMap project, of which 21,000 have been linked to disease (the aim of the International HapMap project is to develop a public resource that will help to associate genes with human disease and response to pharmaceuticals). In addition, there are approximately 60,000 experimentally determined or predicted PTMs and > 30,000 domain annotations. To add to this complexity, it is also important to determine, among other factors, protein function, domain structure, subcellular location and polymorphisms. However, the real bottleneck in annotating the human proteome process is the annotation process itself, which is performed manually, and hence is extremely time-consuming. Based on the continuation of manual annotation the process is expected to take years to converge to a comprehensive set of the human proteome. Another issue in the annotation process is that of nomenclature and tracking, for which the Swiss Institute of Bioinformatics is coordinating with the HGNC to ensure only a single entry exists for every known human protein-coding gene. This meeting session revealed that knowledge on the function of many proteins is low, and for the majority of proteins the exact role played by the individual components of the protein complex is unknown. These challenges are augmented by the need for complete annotated proteomes of non-human species.

Validation of biomarkers

The identification of novel selective biomarkers with sufficient sensitivity to detect and stage disease represents one of the major expectations of proteomics. However, as highlighted by Leigh Anderson (Plasma Proteome Institute), the field has yet to deliver. Dr Anderson suggested that limited budgets for biomarker R&D, rather than an actual biological limitation, were responsible for the declining rate of clinical biomarker registration. According to Dr Anderson's investigations, the total annual R&D budget for biomarker discovery is less than 1% of the corresponding budget for drug R&D. Assuming that a selective and sensitive biomarker is as challenging to identify as a novel drug target, the investment gap could explain the lack of novel biomarker discovery. Furthermore, while the rate of biomarker discovery is low, investments in subsequent validation steps are even lower, and verification of validated biomarkers is essentially non existent. The failure to follow up on promising biomarker leads was explained by a reluctance to invest in validation and verification steps. Since academia is traditionally not funded for this type research, academic researchers do not usually have the means to validate and verify potential biomarkers following their discovery.

Reflecting Dr Anderson's discussion on the lack of validated biomarkers, the session on biomarker discovery at the meeting was dominated by presentations on the development of new tools to validate potential biomarkers rather than discussions of validated biomarkers. Peter Schulz-Knappe (CEO of Proteome Sciences plc) highlighted a related internal divide within the industrial sector. Diagnostic companies that follow up on biomarker leads are typically too antibody-focused, to the extent that if antibodies against epitopes specific for the biomarker in question cannot be produced, the product is not developed further. In contrast, pharmaceutical companies, which are generally better equipped to develop purpose-based assays, are less interested in diagnostic biomarker products. Dr Schulz-Knappe presented the tandem mass tag (TMT) technology, a novel tool for reliable absolute comparisons of biomarker abundance between different laboratories, platforms and workflow designs, developed by Proteome Sciences. The TMT technology is related to the iTRAQ (isobaric tag for relative and absolute quantitation) peptide tagging technology in that amines are covalently labeled with isobaric tags. The TMT sixplex set (ie, a set of six isobaric tags) is specifically designed to function as an internal standard and spiked into all samples to enable independent comparisons, for example, in multisite validation studies.

Geneviève Choquet-Kastylevsky (bioMerieux Inc, France) described a protein reaction monitoring (PRM) system based on a robust platform for automated sample preparation, fragmentation and fractionation prior to liquid chromatography with tandem mass spectrometry (LC-MS/MS) analysis. PRM is an alternative to validating protein fragmentation by ELISA and has demonstrated comparable results between the two methods in quantification of the PSA biomarker model system. Another LC-MS/MS-based validation platform for colorectal cancer biomarkers was presented by Volker Kruft (Applied Biosystems, Germany); candidate markers of colorectal cancer predicted from gene expression analyses could be screened by selective MS/MS detection of proteotypic peptides derived from target proteins.

Plasma proteome biomarkers

Dr Anderson provided an in-depth discussion on several limitations of the current paradigm for clinical biomarker discovery. The declining trend in FDA approval of new plasma protein biomarkers, despite the high amount of global proteomics research, was highlighted. High-throughput proteomic strategies are expected to lead to the discovery of novel clinical biomarkers; however, to date, the total number of successful clinical biomarkers discovered by proteomics is zero. Dr Anderson commented that the field of biomarker discovery is too drastically under-resourced to succeed. A critical bottleneck in this area is the ability to verify potential biomarkers. To address this issue a new approach to verification has been proposed that utilizes multiplexed panels of specific candidate assays based on hybrid immunomass spectrometric detection, that is, a technique denoted SISCAPA (Stable Isotope Standards and Capture by Anti-Petptide Antibodies). SISCAPA detection uses anti-peptide antibodies immobilized on 100-nanoliter nanoaffinity columns to quantify peptides in complex digests. The specific peptides of interest are enriched along with spiked stable isotope-labeled internal standards of the same sequence, thereby reducing the overall peptide background signal. Following elution from the column, peptides are quantified by electrospray MS. A limitation of SISCAPA detection is that only sequence-defined (ie, predetermined) analytes can be detected, but the method does offer the possibility of increased sensitivity. The extremely small volumes of plasma required for this technique provide a potential platform for the systematic verification of large numbers of biomarker candidates from a multitude of samples. This strategy could enable testing of the exciting hypothesis that molecular biomarkers exist for all disease states. To this end, the creation of a library of specific analytes for all human proteins as a resource for the global biomedical community was proposed – this is technically feasible and would provide analytical access to the entire human proteome.

To tag or not to tag

An entire section of the meeting was devoted to the controversial debate of whether to use covalent or metabolic labeling in MS-based proteomics applications or whether to conduct label-free analysis; the general consensus was to use labels or tags. The main advantage for labeling is that it can be used to correct for experimental variation in every step of the often complex multistep analyses conducted in proteomics platforms. Several pre-existing and novel labeling methods were evaluated; these included the aforementioned TMT system, and an alternative to the classical stable isotope labeling with an amino acid in culture (SILAC) method, which was designed to detect labeled peptides on a background of non-labeled peptides. The alternative method was presented by Mara Colzani (University of Lausanne), who explained that the method was specifically designed to detect newly synthesized and secreted proteins. Dr Colzani highlighted observations from Yohann Couté (University of Geneva) that the SILAC medium alone induces differential expression at both the mRNA and protein levels, suggesting that a method in which cells are cultured in a non-SILAC medium can provide significant improvements in accuracy for a range of *in vitro* model systems as well as primary cells. iTRAQ-based labeling was used in an interesting approach by Emoke Bendixen and coworkers (University of Aarhus) to investigate the expression of specific gene-protein pairs across various tissues. This study was the first to correlate the results from iTRAQ with those from 454 sequencing (a large-scale parallel pyrosequencing system developed by 454 Life Sciences), which is of particular relevance because combined analyses of transcriptome and proteome data often result in very low correlations between transcripts and proteins. The across-tissue expression ratios from complementary DNA microarrays, 454 sequencing and iTRAQ revealed a positive correlation between transcript and protein ratios (Pearson's correlation), with the most abundantly expressed gene-protein pairs producing the highest correlation scores.

Evolution of structural and sequence proteomes

Sung-Hou Kim (University of California, Berkeley) discussed the classification and evolution of proteomes. Dr Kim and coworkers have conducted a global mapping of the 'protein-sequence universe' (a collection of all known protein sequences) and 'protein-structure universe' (a collection of all known protein structures) to understand the demographics of all protein structures. This data can then be utilized to identify evolutionary relationships among the different demographic groups and to subsequently infer evolutionary relationships. For example, there are approximately 500 to 40,000 genes in an organism, with > 1.7 x 10^7 species, which translates to > 1 x 10^{10} to 1 x 10^{12} protein species. To date, Dr Kim and coworkers have investigated the structure of approximately 3 x 10^8 protein sequences, of which there are 1 x 10^7 non-redundant proteins and 3 x 10^5 sequence clusters. Together these data comprise the known 'protein-structure universe', that is, a collection of all known protein structures. The structural proteome of an organism can be mapped onto the protein-structure universe, providing knowledge on the evolution of protein structure and function. Dr Kim and coworkers have also developed alignment-free methods to compare whole genome sequences (both coding and non-coding) to construct a sequence proteome. This mapping can also provide a molecular taxonomy of proteins. Dr Kim offered the provocative suggestion that there is potential to create a new model of evolution that accounts for all features of whole chromosomes and not just individual genes.

Quantifying a complete proteome

Matthias Mann (Max-Planck Institute for Biochemistry), whose research has led to numerous technological breakthroughs in the field of proteomics, presented the closing lecture of the meeting. Of particular interest was the report that it is now possible to quantify the entire proteome of a simple system. This accomplishment has been an important goal of the proteomics field for the last 30 years, and was achieved by the development of a technology comprising novel algorithms that increased the accuracy of peptide mass measurements and the proportion of peptides identified. This technology makes possible a microarray-like investigation of the proteome, enabling large-scale profiling of systemic fluctuations that are important for systems biology strategies that quantify the effects of perturbations at the systemic level. Comprehensive quantification of the proteome is an important step toward establishing a robust, reproducible high-throughput platform that will enable advances in proteomics similar to the genomic advances created by microarrays. A particular focus throughout this conference was the SILAC procedure. In this method, cells are metabolically labeled by stable non-radioactive amino acids (eg, ^{2}H, ^{13}C and ^{15}N labels), which render the proteomes of these cells distinct from cells containing only unlabeled amino acids. In a SILAC experiment, two groups of cells are grown in culture media that were identical except for the presence of the label in one of the media. Because only essential amino acids are labeled, the cells in the media containing the label are forced to incorporate the labeled amino acid. Consequently, all SILAC-labeled proteins were fully labeled, in contrast to chemical modification methods that may not achieve 100% labeling efficiency. Generally, SILAC experiments are conducted in cell culture, but Dr Mann and coworkers have demonstrated that whole animal labeling is possible, thereby enabling the study of any tissue.

The SILAC strategy is being combined with high resolution MS to quantify the phosphoproteome (publicaly available in the PHOSIDA database; *http://www.phosida.com/*). Protein phosphorylation is a ubiquitous PTM that affects approximately one third of all proteins, with abnormal phosphorylation often leading to the development of severe disease. A statistical algorithm termed 'PTM score' was developed to determine sites of modification on phosphopeptides, which are automatically calculated in the proprietary MSQuant software. Currently this strategy can detect 6600 phosphorylation sites in HeLa cells in a single proteomics experiment. The described approaches are being combined to develop methods that can be applied to individuals. In one strategy, cell lines corresponding to the tissue of interest were SILAC-labeled and then mixed with case and control cells to derive a quantitative proteome. In a second strategy, novel algorithms for 'label-free quantification', in which the MS signal of each peptide was compared across several patients, were utilized. This method provided similar accuracy to microarray technology, enabling systemic profiling of clinical samples at the proteome as well as the genome level. Taken together, these advances demonstrate the potential to develop analytical methodology for routine quantification of entire proteomes.

Summary

The 2008 Siena Meeting covered the full range of proteomics research and demonstrated the power of technology to quantify changes in an entire proteome. Moreover, the importance of bioinformatics support for these research strategies was equally stressed. A vital point repeatedly raised was the requirement for tools to analyze the large datasets that are being routinely collected. Several researchers are utilizing a variety of currently available pathway analysis tools, for example, Ingenuity and the KEGG database. Numerous other resources (both commercial as well as custom) were also presented. These tools have proved useful for some analyses, but there is still a general lack of pathway analysis software capable of managing large-scale input (eg, simultaneous transcriptomics, proteomics and metabolomics data) and with the ability to examine flux dynamics or multiple dosing parameters. These capabilities will become increasingly important as experiments shift from cell-based models to animal systems. As the acquisition of large-scale datasets continues to increase, the requirement for appropriate software tools for data management and analysis will become acute. The intended application of many proteomics research projects is the identification of clinical biomarkers of disease, but several limitations in this approach were presented. In addition, the therapeutic concept of a magic bullet directed against a specific target was presented as unrealistic because many diseases are polygenic. Accordingly, it is clear that although much has been accomplished in the field of proteomics, it has yet to deliver. The 9th Siena Meeting was scheduled for August 29 to September 2, 2010, and the first validated biomarker derived from proteomics strategies to be presented at that meeting is waited for with much anticipation.