FEBS Letters 583 (2009) 1703–1712

ELSEVIER

FEBS Letters

journal homepage: www.FEBSLetters.org

Minireview

Bioinformatics analysis of mass spectrometry-based proteomics data sets

Chanchal Kumar, Matthias Mann *

Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany

ARTICLE INFO

Article history:
Received 26 February 2009
Accepted 18 March 2009
Available online 21 March 2009

Edited by Miguel De la Rosa

Keywords:
Quantitative mass spectrometry
Computational proteomics
Bioinformatics
Systems biology
SILAC
MaxQuant

ABSTRACT

Proteomics has made tremendous progress, attaining throughput and comprehensiveness so far only seen in genomics technologies. The consequent avalanche of proteome level data poses great analytical challenges for downstream interpretation. We review bioinformatic analysis of qualitative and quantitative proteomic data, focusing on current and emerging paradigms employed for functional analysis, data mining and knowledge discovery from high resolution quantitative mass spectrometric data. Many bioinformatics tools developed for microarrays can be reused in proteomics, however, the uniquely quantitative nature of proteomics data also offers entirely novel analysis possibilities, which directly suggest and illuminate biological mechanisms.

1. Introduction

Biological systems function via intricately orchestrated cellular processes in which various cellular entities – RNAs, metabolites and proteins participate in a tightly regulated manner. Proteins are at the 'executive core' of these cellular events, and their altered behaviors have been implicated in myriad disease pathologies, which also makes them by far the major class of drug targets. Therefore, understanding the structure, dynamics and interactions of proteins has been at the heart of biomedical research from its very inception. Due to limitations of biochemical methods and allied technologies, such studies have traditionally been carried out on single proteins, rather than the entire population of expressed proteins in a cell or tissue, the 'proteome' [1]. The discipline of proteomics was initially equated with two-dimensional gel electrophoresis, a low resolution technology that can only analyze the most abundant proteins in a sample. In recent years mass spectrometry (MS) has become a powerful technology to study proteins on a large-scale [2,3]. Combined with innovative experimental strategies [4], and advances in computational methods [5,6], MS-based proteomics now enables global study of cellular proteomes.

This relatively novel development has led to a surge of qualitative and quantitative data at the proteome level, which has posed analytical challenges hitherto unseen by protein researchers. The mapping of complex proteomics data to biological processes has become impossible by manual means, and the need for computer-aided data analysis is essential for further progress of the field. Proteomics is today at the same crossroads that genomics was at a decade ago in terms of tackling these challenges. Bioinformatics, the scientific field dealing with analyzing large numbers of genes or their transcripts, in fact emerged largely from that challenge [7]. It has evolved to deal with a multitude of different biological data types and should now be well-equipped to aid proteomics [8]. Indeed, proteomics researchers are already actively collaborating with bioinformaticians for comprehensive functional analysis and systematic knowledge mining of complex data sets.

We subscribe to the definition of bioinformatics as a mean for functional analysis and data mining of data sets leading to biologically interpretable results and insights. In this review we highlight recent advances, results, and challenges in proteome based bioinformatics research. Thus the scope of our review is downstream of the related and partially overlapping field of '*computational proteomics*' which blends mathematical, computational and statistical algorithms to address key problems related to protein identification and quantitation from raw mass spectrometry data.

2. Bioinformatics for qualitative proteomics

Until a few years ago proteomics was largely a qualitative discipline. The proteomic experiment typically consisted in identifying as many proteins as possible in a protein complex, organelle or cell or tissue lysate. In the course of obtaining the protein identities of any protein mixture, an enzymatic digestion step is usually employed, yielding a large collection of proteolytic peptides that are then analyzed by 'shotgun' proteomics [9]. This is illustrated in

* Corresponding author. Fax: +49 89 8578 3209.
E-mail address: mmann@biochem.mpg.de (M. Mann).

0014-5793/$36.00 © 2009 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2009.03.035

the upper part of Fig. 1. The peptide inventory, which can include normal peptides or peptides bearing post-translational modifications such as phosphorylation, has been bioinformatically analyzed for various purposes as symbolized by the left part of the figure. The focus of qualitative proteomics was on the correctness and the depth of analysis but the result of the experiment was typically simply a list of proteins. As proteome catalogs started getting larger, they were nevertheless unyielding to manual analysis due to the sheer numbers of proteins, and the immediate challenge was to obtain biological insights into the system being studied (right part of Fig. 1).

The peptides created and collated in shotgun proteomics projects are not functional biological entities and they are therefore usually only of interest for the technology of proteomics itself. Nevertheless, they can be mined for physiochemical and amino acid residue patterns using machine learning approaches, which form the basis for various classification and prediction routines. This can be useful, for example, in predicting which peptides are likely to be detected in proteomics experiments and which are unique for the parent protein – so called proteotypic peptides [10,11]. These peptides can then be specifically targeted by specialized mass spectrometric techniques such as multiple reaction monitoring (MRM) during an analysis in which one is only interested in monitoring the levels of selected proteins [12]. The 'PeptideAtlas' has been created for this purpose and extracts proteotypic peptides and their associated fragmentation spectra from a large number of submitted proteomic data sets [13].

Peptides sequenced in proteomics projects can be mapped onto the positions in the genome that code for them. In this way the peptides provide evidence that the gene is actually expressed and is not, for example, a pseudogene. This is important because a large fraction of the predicted genes in the genomes of eukaryotes do not yet have any direct experimental protein information associated with them. Peptide atlases have been used to find novel transcripts, and to refine gene models, in principle leading to augmented genome and proteome annotations [14,15]. A new sub-discipline of bioinformatics called 'comparative proteogenomics' has now emerged from such endeavors, which proposes to harness MS-based proteomics data sets in conjunction with DNA sequence data sets for large-scale genome and proteome annotation [16]. So far efforts in this area have mined large-scale and usually low resolution data. However, in our opinion, genome annotation should only be done with very high accuracy data that has extremely low error rates. Such data can now readily be produced by the last generation of high precision mass spectrometers [3].

Additionally, the peptide identification information can serve as a very rough indication of protein quantity in the sample. The basic idea is that the abundance of each protein scales with the number of identified peptides. One approach, termed peptide or spectral counting uses the number of times that peptides belonging to a protein are fragmented as a proxy for its abundance [17,18]. Kislinger et al. elucidated the proteome of six mouse tissues in this way, followed by bioinformatics analysis of tissue specific proteome function and regulation [19]. In a related approach, Ishihama et al. showed that the absolute amount of protein in the sample studied correlates with the exponential of the Protein Abundance Index (peptides observed by MS divided by peptides that are potentially observable) [20]. In a more advanced approach Lu et al. used the peptide sampling information in a Bayesian framework to define absolute protein expression measurements (APEX). This measure provided an estimate of the relative contributions of transcriptional and translational-level gene regulation in yeast and *Escherichia coli* [21]. Nevertheless, these semi-quantitative approaches are falling out of favor because modern, high resolution

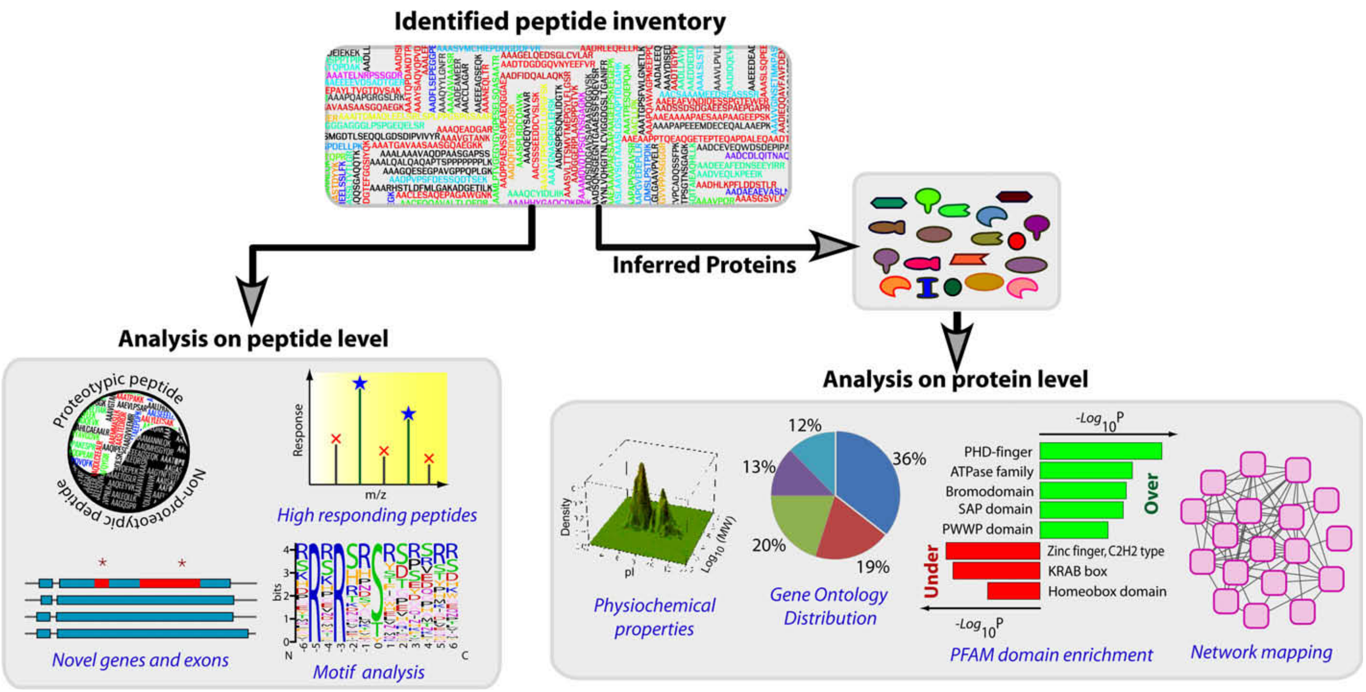

Fig. 1. Bioinformatics analysis paths for qualitative proteomics. The peptide inventory identified by 'shotgun proteomics', (left part of figure) can be mined for patterns by machine learning approaches–such as proteotypic peptides. The identified peptides can be mapped to genomic coordinates for identifying novel ORFs and for augmenting genome annotations. Post translationally modified (PTM) peptides such as phosphopeptides can be analyzed for sequence motifs. The right hand side shows analysis directions after the peptide identifications are consolidated into protein identifications. The proteins can be examined for their physiochemical properties to uncover MS sampling and identification biases towards acidic or basic proteins or high molecular weight proteins in a sample, for example. On the functional level these proteins can be integrated with annotational databases like Gene Ontology and PFAM to find enriched biological processes, functions, cellular components and protein domains. Additionally these proteins can be mapped to network and pathway databases (STRING, KEGG) to visualize them in their modular functional contexts.

instruments allow direct comparison of peptide signals across experiments, in so called 'label-free' and much more accurate quantitative proteomics (see below).

Mass spectrometry is especially well suited to analyze post-translational modifications (PTMs) on peptides. In contrast to unmodified peptides, these PTM-bearing peptides are of great biological interest because they reveal sites of functional changes to a protein. Large-scale studies of phosphorylation, ubiquitination, acetylation and many other PTMs are now possible, especially if the modified peptides can be specifically enriched with respect to unmodified ones [22–24]. In phosphorylation studies, the site specific information is used to extract enriched sequence motifs, which in turn provides insights into proteome regulation by upsteam kinases, modular protein domain mediated interactions, and also a basis for prediction of novel PTM sites [25–27]. Recently, an approach called NetworKIN was reported that mines large-scale phosphorylation data sets in the context of the protein–protein interaction network topology to predict kinase substrates in phosphorylation networks [28]. The analysis of PTMs by mass spectrometry – especially in a quantitative format as described below – is increasing exponentially and will be one of the main contributions of MS-based proteomics to biology.

Several types of bioinformatic analyses are almost invariably performed on a measured proteome. Analyzing proteins for sequence features like transmembrane domains and signal peptides can provide clues about features of studied sample – for instance in cases where the membrane proteome is enriched, or where one needs to determine the fraction of secreted proteins [29]. This also helps in ascertaining and correcting experimental or MS-identification related sampling biases in proteome catalogs [30]. For example, Shi et al. used a kernel density based approach to correlate the isoelectric point (*pI*) and molecular weight feature space of an MS-based mouse liver proteome to show that it was much more unbiased than earlier 2D-gel based studies, and was largely representative of complete mouse proteome (with Pearson correlation of 0.98) [31]. The same approach was subsequently applied to compare in-gel digestion and isoelectric focusing separation methods in a study of a Drosophila cell line [32].

On the level of protein catalogs, the bioinformatics analysis typically involves integration of proteome data with annotational databases, such as Gene Ontology (GO) [33], protein domains (InterPro, PFAM) [34] and pathway database (KEGG) [35] – to determine if any of these properties are over or underrepresented. This type of analysis may directly yield functional insights into the data set and is easily accomplished using standard tools. DAVID [36], GoMiner [37], Cytoscape [38] plug-ins like BINGO [39] are examples of readily available software that can be used. In our laboratory, we typically employ Bioconductor [40] within the R statistical platform [41]. This requires some more programming experience but offers broader capabilities and flexibility in analysis.

Adachi et al. studied the proteome of 3T3-L1 adipocytes and performed advanced bioinformatic mining of a qualitative data set [42]. The proteome was first mapped to GO, KEGG and InterPro databases, providing insights into adipocyte biology and functions. Statistical enrichment tests were performed to find significant over-represented GO and InterPro categories, which in turn related to signal transduction, redox system, protein transport, translation, transcription, protein degradation, fatty acid synthesis, and phospholipid biosynthesis; all characteristic functions of adipocytes. Putative biological functions were assigned to more than 50% of un-annotated proteins identified in 3T3-L1 cells through sequence similarity based annotation transfer [43]. Additionally, a novel tool for functional association of proteins to protein-models prototypical of the function of interest (insulin mediated vesicular traffic in this case) was employed [44]. This led to the association of several proteins in the adipocyte proteome with this function, at least one of which was later independently validated [45].

Combining proteomic data sets with complementary 'omics' data sets such as transcriptome data can reveal interesting facets of cellular functions. In an early example, Mootha et al. compared the mitochondrial proteome with tissue microarray data to show that for mitochondrial proteins on a bulk level mRNA expression levels are correlated with protein detection and their abundances. Exploiting the much more readily available microarray data sets, they also found that mitochondrial proteins show tissue specific patterns of expression and regulation to a much greater extent than previously recognized [46]. Furthermore, they characterized key transcriptional regulators of mitochondria organelle biogenesis using expression neighborhood analysis, which identified these proteins by the co-regulation of their messages with the messages corresponding to the mitochondrial proteome.

In a mouse liver organellar proteome study Foster et al. used regulatory motif analysis to elucidate key players mediating organelle biogenesis [47]. Calvo et al. integrated protein identification information in mitochondria along with a collection of functional genomics data sets in a naive Bayesian method to predict novel mitochondrial candidates in humans [48]. Recently, Pagliarini et al. used tissue wide mitochondrial proteome data from mouse as an input for phylogenetic profiling across 42 eukaryotic species and identified 19 new candidates of respiratory chain complex I (CI). One of them *C8orf38* was directly implicated in a lethal CI deficiency [49]. Graumann et al. for the first time compared the mouse embryonic stem cell (mESC) proteome with a genome wide chromatin state map of mESC to show near perfect correlation between protein expression and the presence of active rather than repressive chromatin marks [50]. Mapping proteome data to pathway databases like KEGG and network databases like STRING [51], MINT/IntAct [52] and HPRD [53] can provide valuable clues about the presence of signaling pathways and functionally interacting modules of interest.

3. The nature of quantitative proteomics data

Functional insight most often requires quantitative comparison between two or more biological states. In the last few years proteomics has been catapulted into the realm of high-throughput 'omics' technologies mainly due to significant advances in two aspects, first the development of accurate methods of proteome-wide quantitation, and second the development of computational proteomics algorithms and software for efficiently harnessing this quantitative proteomic data.

While mass spectrometry is not inherently quantitative, this limitation has been successfully overcome by introduction of stable isotopes into the molecules to be identified. Stable isotope labeling has been done for decades in small molecule MS and can be performed in proteomics either by chemical modification of peptides after tryptic digestion or by metabolic labeling of intact proteins during cell culture [54]. For example, iTRAQ is a commonly used technique in quantitative proteomics in which amino groups of peptides (lysine side chain and the N-terminus) are chemically labeled by isotopically different forms of the derivatizing agent. The most widespread metabolic labeling technique is SILAC (stable isotope labeling by amino acids in cell culture) [55,56]. As the name indicates, SILAC incorporates the heavy labeled amino acid into the entire proteome in the course of normal cell metabolism and proliferation. SILAC therefore does not require any chemical derivatization. It is generally considered the most accurate quantitation strategy because all peptides of a protein are labeled and because processing of proteins occurs after samples have already been combined and therefore cannot contribute to any quantitation error. In our laboratory, the SILAC approach has enabled comprehensive quantitation of the yeast proteome [57], and now facilitates routine measurement of expression changes of 4000–6000 proteins in more complex eukaryotic cells [32,50].

Alternatively, protein quantification without the use of isotopic labels is emerging as a practical approach in MS-based proteomics. This 'label-free quantitation' is especially important where isotope labeling is not feasible or scalable, for example in many instances in which tissues or clinical samples are measured. A precursor of quantitative proteomics for patient samples is the SELDI method, in which a low resolution MALDI spectrum is taken to be indicative of the state of the proteome [58]. However, despite an extensive literature on patient classification using these patterns, the accuracy of such results and of the underlying data have been questioned [59,60]. In contrast, high resolution instruments are now making it possible to directly compare the integrated peptide ion signals between experiments. With the inclusion of sophisticated MS data (signal) processing algorithms and advanced statistical procedures accurate label-free proteomics appears to become feasible, which may herald the beginning of successful clinical and in vivo tissue proteomics endeavors.

Modern proteomic experiments generate gigabytes of data for a typical experiment. While advanced algorithms for protein identifications have been in vogue for nearly a decade [61,62], robust algorithms for extracting protein quantitation information from the multidimensional MS data structures have only recently started to emerge. Retrieving protein identification and quantitation from MS data is an intensive multi-level algorithmic endeavor, now studied in the above mentioned sub-discipline of '*computational proteomics*'. It spans a gamut of computational, statistical and machine learning algorithms especially applied or developed for peptide and protein identification and quantitation. In comparison to efforts aimed at microarray data, the lack of standardized and comprehensive quantitation software for MS data has been one of the major challenges and bottleneck for proteomics. Empirical methods like spectral counting have been employed for protein quantitation (see above). Even methods that took the great complexity of proteomics data [63] into account, were inherently of low accuracy because they were developed for low resolution MS data. Therefore, they fail to deliver when MS data is highly resolved and fine grained as generated by the latest generation of mass spectrometers. Our laboratory has developed MaxQuant – a suite of integrated algorithms specifically designed for high resolution, quantitative MS data based on state-of-the-art data reduction, correlation analysis and graph theory [6]. Mueller et al. recently reviewed other existing computational proteomics frameworks and software [64].

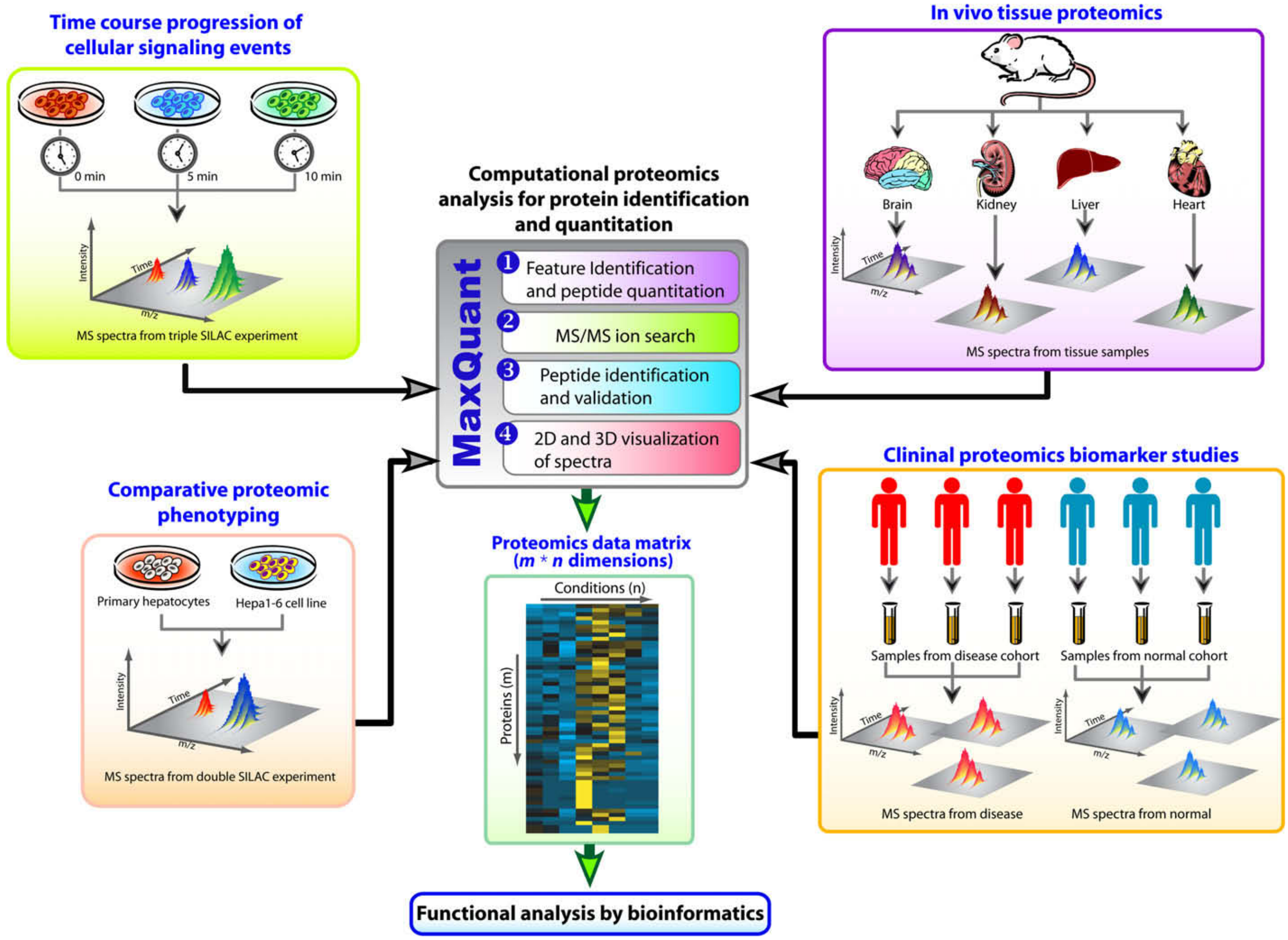
Time course progression of cellular signaling events
0 min
5 min
10 min
Intensity
Time
m/z
MS spectra from triple SILAC experiment
Computational proteomics analysis for protein identification and quantitation
MaxQuant
1 Feature Identification and peptide quantitation
2 MS/MS ion search
3 Peptide identification and validation
4 2D and 3D visualization of spectra
In vivo tissue proteomics
Brain
Kidney
Liver
Heart
Intensity
Time
m/z
MS spectra from tissue samples
Comparative proteomic phenotyping
Primary hepatocytes
Hepa1-6 cell line
Intensity
Time
m/z
MS spectra from double SILAC experiment
Proteomics data matrix (m * n dimensions)
Conditions (n)
Proteins (m)
Clinal proteomics biomarker studies
Samples from disease cohort
Samples from normal cohort
Intensity
Time
m/z
MS spectra from disease
MS spectra from normal
Functional analysis by bioinformatics

Fig. 2. Quantitative data generation in MS based proteomics. Contemporary mass spectrometry based proteomics combines sophisticated experimentations, advanced MS instrumentation and computational proteomics platforms to generate high dimensional data sets. These data sets can come from either isotopic-labeled samples (such as SILAC) (left side of the figure) which typically are used in following temporal trajectories of cellular signaling events or in comparative proteomic phenotyping experiments across two or more levels. Proteomes that were not isotopically labeled (right side of the figure) like tissue or clinical samples can be analyzed by a 'label-free' approach. Computational software such as MaxQuant enables parallel processing of these complex data sets generating a multidimensional data matrix which contains a wealth of information on peptide and protein identity, PTMs and their quantitative ratios.

These computational proteomics frameworks need to handle the very large size of MS data sets that can readily be generated today. Complex experimental schemes need to be accommodated – enabling parallel processing of samples generated by isotope-labeled and non-labeled proteomic experiments in replicates and cross-over (isotope-swapping) experiments (Fig. 2). Accurate quantitation, associated statistic, and quality control metrics need to be generated and reported for thousands of proteins in each project. One end product of such endeavors is a resultant matrix containing expression values of thousands of proteins across many conditions but also containing information pertaining to the peptides identified, their uniqueness to the protein and so on. Such matrices are much more complex than data structures generated by microarrays, and pose at least as difficult analysis and interpretation challenges that accrue from their high dimensionality [65]. Furthermore, the wealth of proteomic information needs to be mapped onto existing biological knowledge to generate new insights. The general task for bioinformatics in this context is to provide the framework for systematic knowledge mining of such proteomics data sets thereby mapping them back onto their biological context.

4. Bioinformatics for quantitative temporal, spatial, interaction, phenotypic and proteomics

The technological ability of obtaining accurate global quantitative proteome data by MS has infused new vigor into proteomic studies in which cellular system trajectories are studied at the level of proteome dynamics and post-translational modifications. Such data sets are now regularly generated at different levels of cellular function, organization and architecture and it is possible to systematically investigate proteomes from various perspectives. These include quantitative protein–mRNA expression correlation and translational regulation effects on protein expression (Fig. 3A), comparisons of the 'proteomic phenotype' across various cellular or physiological states (Fig. 3B), temporal kinetics of proteins and their post-translationally modified peptides (Fig. 3C), spatial distribution of proteins across sub-cellular organelles (Fig. 3D), and interaction partners of proteins mediating a pathway or cellular process (Fig. 3E). Such diverse proteomic data sets have in turn opened exciting vistas for their analysis by bioinformatics and machine learning algorithms.

4.1. Quantitative gene-expression at message and protein levels

Quantitative mass spectrometry in principle provides similar data than a microarray experiment; however, there are also important differences. Due to technological challenges, quantitative proteomic data sets have until recently been much smaller than microarray data sets and they were often biased against low abundance proteins. Nevertheless, proteomics has for some time been able to obtain a relatively complete view of quantitative changes in sub-proteomes such as organelles, for which there is no microarray equivalent. If the proteomics experiment is meant to provide a quantitative map of the gene-expression differences between two cellular states then it has the same goal as a microarray experiment (or nowadays 'deep RNA sequencing' [66]) and in principle the same bioinformatics tools can be used. Even in this case proteomics may be closer to functional relevance because usually mRNA is only an intermediate but not itself the active molecule in the cell. This is true even when we consider regulatory mechanisms such as those mediated by microRNAs, because their ulti-

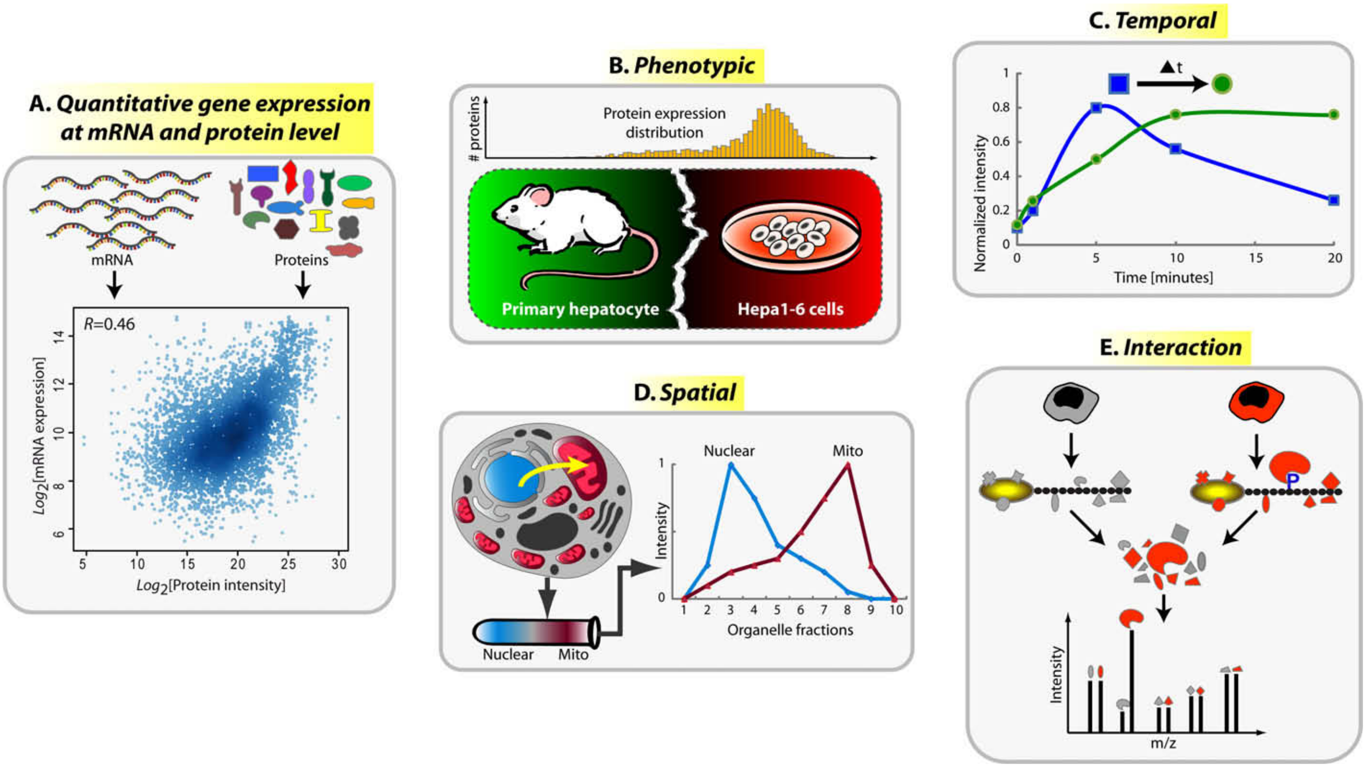

Fig. 3. Quantitative proteome analysis on multiple levels of cellular functions. Quantitative proteomics can now study cellular systems and functions from various perspectives of cellular organization. Typical questions that have now become feasible to answers include: (A) How does gene expression correlate quantitatively at mRNA and proteome levels? (B) What are the phenotypic differences between a different but related cell at the proteome level? (C) On a temporal scale how is a substrate phosphorylated by an upstream kinase? (D) What is the distribution of protein populations in organelles like mitochondria and nucleus versus the cytoplasm in any given cellular condition? (E) What are the specific protein-protein interactions mediated by a change in phosphorylation state of a protein?

mate outcome is still a change in protein amount. As exemplified below, comprehensive quantitative proteomics has now become possible and microarray and proteomics data sets can both be analyzed in a systems-wide manner.

Comparison of transcript and proteome level changes has been a topic of immense interest since the days when microarrays and proteomics technologies were still in their infancy [67,68]. These early studies as well as recent ones using much more advanced instrumentation have generally reported that at the global level proteome and transcriptome exhibit poor correlation, while this correlation is better for changes associated with a perturbation of the system under study [32,57]. Quantitative proteomic data now facilitate accurate assessment of transcriptome–proteome correlation and helps in estimating the apparent level of technical and stochastic noise in the respective data sets. In a recent proteome wide study of haploid vs. diploid yeast Godoy et al. compared the SILAC based protein ratios with mRNA expression and found that for strongly regulated entities of the pheromone pathway the correlation between protein and mRNAs was relatively high with correlation coefficient of 0.68 (Fig. 4A).

In a study of proteome changes upon dsRNA mediated knockdown of the chromatin remodeling factor ISWI in Drosophila Bonaldi et al. showed that the quantified proteome covered a broad spectrum of biological processes and functions. In terms of physiochemical properties (*pI*, molecular weight) these functions were comparable to the complete Drosophila proteome ruling out bias in MS-based identification. This study again showed minimal correlation for steady state levels of message and protein but enhanced correlation between regulated proteins and their corresponding transcripts. When only considering robustly regulated genes, the correlation coefficient reached 0.8. Despite this good correlation, the actually fold change of the proteins was not predictable from microarray or RT-PCR data.

With the availability of high quality quantitative proteomic data it is now feasible to move beyond the question of correlation and explore differences in protein expression caused by translational regulation. In a recent study Selbach et al. studied global protein expression changes induced by miRNA mediated gene regulation for five distinct miRNAs in HeLa cells using 'pulsed SILAC' [69]. Correlating protein expression as a function of miRNA binding sites feature on the target mRNAs they found that protein downregulation is strongly influenced by the presence of 3′UTR seed sequence motifs, and repression was more pronounced for conserved seed sites. Moreover, they showed that only seed-containing mRNAs (of the proteins profiled) with at least one mismatch were overall repressed at the protein level, and repression was more pronounced for conserved seed sites. Additionally, comparing protein expression with cognate transcript expression showed that a multiplicity of a miRNA-binding site in the same 3′UTR exerts a stronger direct effect on protein production than on mRNA levels. Another study on the same subject found, in common with the Selbach et al. paper, that single miRNAs have a widespread but relatively mild effect on proteins targets [70].

4.2. Function from quantitative proteome comparisons

SILAC is a very potent and accurate technique for comparing relative proteome change where two or more phenotypic states need to be compared, for instance before and after knock-down of a gene of interest or to compare two related by divergent cell types. Typically, such studies provide relative proteome ratios for thousands of proteins, and their goals are twofold – first to enumerate global properties of the quantified proteome, second and more importantly to access systems level differences in terms of biological functions, regulation and pathway modules.

Global comparisons of cellular systems and organisms have so far been done on transcriptome, interactome and epigenome levels [71–74]. Advanced quantitative proteomics adds a novel and very pertinent dimension to these comparisons which should allow molecular phenotypes to be discerned directly at the proteome level. Pan et al. measured the quantitative proteomic difference between Hepa1–6, a transformed hepatocyte cell line, and primary hepatocytes from mouse by 'proteomic phenotyping' [75]. This approach combines the strength of high accuracy quantitation by SILAC with a novel bioinformatics algorithm to provide a functional portrait of significant phenotypic differences (Fig. 4B). In this first proteome level comparison of a cell line vs. the corresponding primary cells more than 4000 proteins were quantified with high accuracy and reproducibility (Pearson correlation across biological replicates 0.95). Their quantitative distribution exhibited broad spectrum of proteome changes (from -10 to +5 on $\log_2$ scale) and was non-normally distributed with substantial negative skew. Consequently, conventional methods of assigning up and downregulation would have failed because they implicitly assume an underlying Gaussian distribution of fold changes. Instead, the complete quantitative expression map was divided into five quantiles of expressions and analyzed by bioinformatics for enriched KEGG pathways, and GO biological process and cellular component of each of the subpopulations. After consolidation and hierarchical clustering of statistically significant categories separately in KEGG and GO a clear functional phenotype emerged, which provided interesting insights into divergent and shared features of the two cell types. For example, the result showed that Hepa1–6 cells were deficient in mitochondria, have lost many of the specific functions typical of hepatocytes in vivo – for instance drug metabolizing enzymes (DMEs), complement production, and synthesis of extracellular matrix. Conversely, the cell line shifts most of its resources into functions associated with cellular proliferation, but maintains important cell signaling pathways. This study and the associated bioinformatics provides a good starting point as to which functions can be studied in a the cell line and which are better studied in the primary cells or in vivo. It also serves as a template for investigations of cellular phenotypes across various conditions.

4.3. Bioinformatics of temporal proteomics

Just as many microarray studies measure not only static differences between two cellular states but also the kinetics of a biological process, proteomics can also directly measure kinetics. For example, using the SILAC technology, three cellular states can be isotopically labeled and be perturbed for three different time points [76]. Repeating the experiment several times with a common reference point yields five point or higher kinetics. This scheme was used to directly study the change in the nucleolar proteome upon small molecule inhibition of the proteasome [77]. Functionally or physically interacting protein groups had similar kinetics and the experiment supported the notion that the nucleolus is not purely an activity based cellular structure dedicated to transcribing rRNA genes.

White and co-workers have additionally linked temporal proteomics to a matrix of downstream outcomes after epidermal growth factor receptor (EGFR) activation [78]. They studied the time course profiles of quantitative tyrosine phosphorylation changes using iTRAQ, and the resultant temporal data for 78 sites was clustered using self organizing map(SOM) thereby revealing interesting modules in EGFR signaling network. In one of the first systems level phosphoproteomics study of in vivo signaling events, Olsen et al. employed SILAC to generate quantitative time course data of phosphoproteome changes across five time points after EGF stimulation [79]. Apart from very high identification confidence in the phosphopeptides, they provided statistical confidence

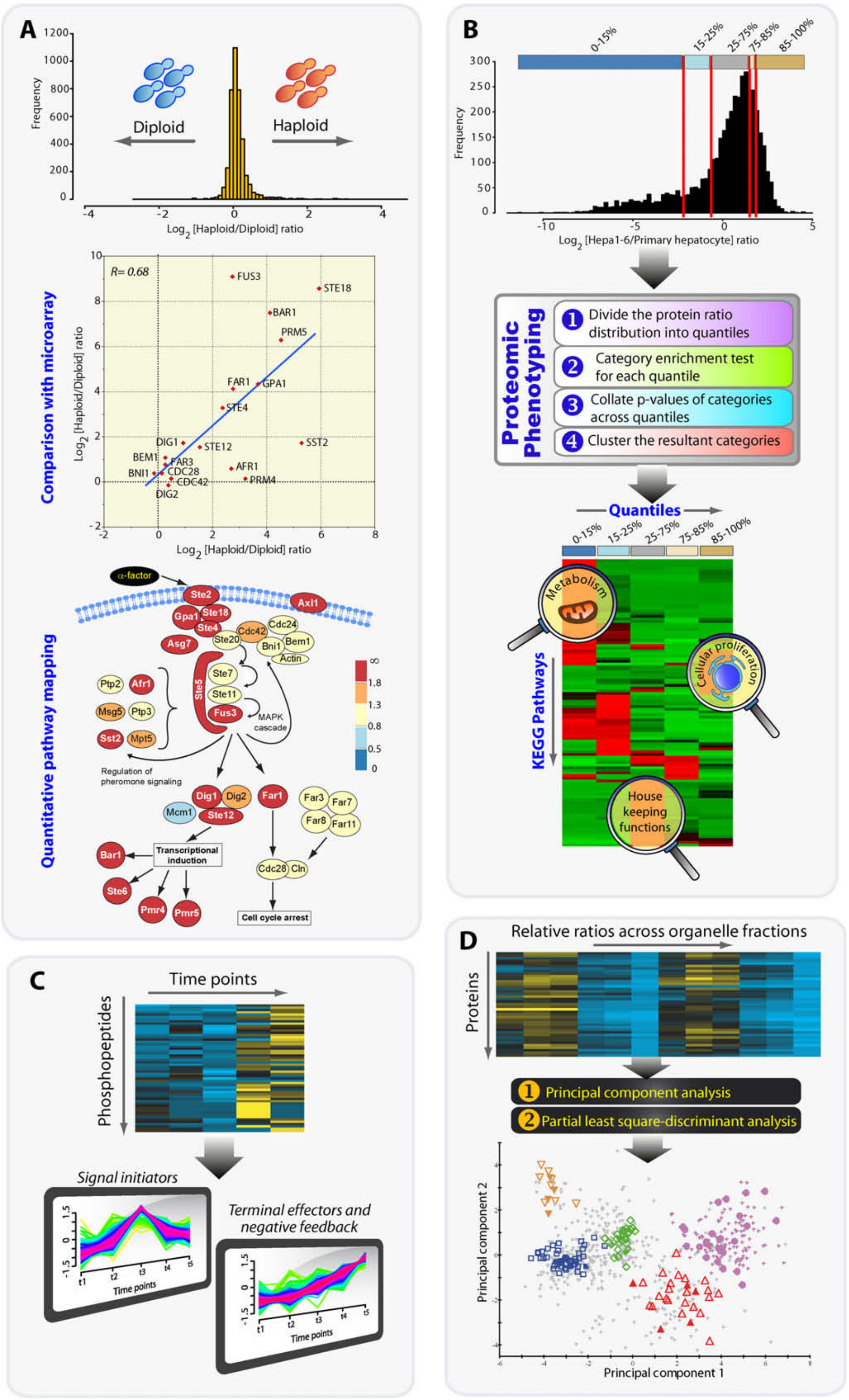

Fig. 4. Exploring quantitative proteomics data using bioinformatics and machine learning. (A) Accurate quantitative proteomics enables gene expression comparison at protein and mRNA levels thereby providing insights into translational and transcriptional regulations. In a global yeast proteomic study the quantitative proteomic ratios correlated significantly with the mRNAs for modulated entities of the pheromone pathway (middle part), while the global steady state correlation was minimal. The quantitative proteome provided a systems wide view into the concomitant up regulation of pheromone signaling pathway components (bottom part). (B) 'Proteomic phenotyping' provides a novel bioinformatics approach to functionally analyze the quantitative expression map emerging from proteomic comparison of primary cells and cell line. (C) Time course profiles of phosphopeptides obtained after EGF stimulation were clustered using fuzzy c-means clustering to discern cluster patterns which were mapped to representative biological functions. (D) In organellar proteomics of Arabidopsis the relative quantitative protein ratios across membrane fractions were analyzed to assign novel proteins to various organelles of the endo-membrane system.

to the location of the phosphogroup in the peptide by probabilistic phosphosite assignment. Analyzing the data for Gene Ontology enrichment revealed that the majority of phosphorylation events occur in the nucleus and cytoplasm, whereas mitochondria, for example, are underrepresented. The data was furthered analyzed using fuzzy c-means clustering to elucidate common patterns of regulation. Key proteins with similar temporal profiles were automatically grouped into clusters which turned out to form biologically relevant functional classes (Fig. 4C). The results of this analysis were made available to biomedical community via a database called PHOSIDA [27] which has since then matured into a comprehensive data repository for phosphorylation related information.

4.4. Bioinformatics of spatial proteomics

Mapping the sub-cellular localization of proteins is one of the first steps towards understanding their function in particular cellular contexts and is a prerequisite for systems-wide understanding of the cell [80,81]. The task for organellar proteomics in particular is to provide an inventory of the proteins making up the individual organelles. Initially such studies were performed purely qualitatively, which may have been adequate when MS could only identify the major, purified components. However, with increasing MS sensitivity minute contaminations are easily detected and consequently there are no 'pure' organellar preparations anymore. Therefore a quantitative dimension has become essential. Initial quantitative organellar proteomics studies include the work of Foster et al. who used 'protein correlation profiling' (PCP) to determine the integrated peptide intensity profiles for each protein across sub-cellular fractions. Correlation with organelle specific marker proteins led to the localization of more than 1,400 proteins to cytosolic organelles in mouse liver [47]. Through bioinformatic analysis based on the concept of gene-expression neighborhood (basically the Euclidian distance in gene expression space) they elucidated key regulators of organelle biogenesis. Additionally, by comparative genomics for cis-regulated, conserved elements they obtained novel organelle specific enriched motifs–for instance binding motifs of Myc upstream of the genes encoding ER resident proteins. In a technique termed Localization of Organelle Proteins by Isotope Tagging (LOPIT), iTRAQ based relative quantitative protein ratios are obtained across multiple sub-cellular fractions and analyzed by multivariate data analysis algorithms. Dunkley et al. applied LOPIT to the Arabidopsis membrane proteome and determined the relative quantitative protein ratios across membrane gradient fractions by principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) (Fig. 4D) [82]. This analysis assigned 527 proteins to organelles of endomembrane system and also predicted novel candidates in Golgi apparatus, endoplasmic reticulum, vacuolar and plasma membrane.

4.5. Bioinformatics of interaction proteomics

Cellular processes are generally carried out by stable or transiently assembled protein complexes, therefore identification of protein-protein interaction (PPI) provides insights into cellular dynamics and function. Recent advances in proteomics and allied technologies have lead to a deluge of protein interaction data and resulting interaction networks and databases [83]. They have also spawned a field of computational biology in which mathematical network analysis is applied to molecular and cell biology. One of the key analytical directions in the PPI field is understanding the network behavior of these interactomes using concepts from 'Network Biology' [84], and to gain insights into cellular organization (as complexes or modules) and function by bioinformatics and machine learning methods [85,86]. A prominent question in all these studies is ascertaining the degree of true versus false interactions (i.e. functionally valid vs. background interactions), which is a key requirement for formulating biologically testable hypothesis [87,88]. MS based quantitative proteomics is specifically suited for providing highly accurate information of protein interactions but only when it explicitly incorporates a quantitative dimension [89]. The quantitative ratios are used not to determine the strength of interaction but to discriminate background binders from specific binders. This dimension has yet to be incorporated into bioinformatics modeling of interaction networks.

MS is also the only generic technology that can provide data on interactions mediated by specific post translational modification (PTM) such as phosphorylation in a close to *in vivo* situation [90–92]. Our laboratory has recently used SILAC for the systematic study of the phosphotyrosine interactome of the EGFR receptor family as well as the insulin receptor family and its substrates IRS-1 and IRS-2 [93,94]. Miller et al. combined bioinformatics approaches to first decompose 481 unique tyrosine phosphopeptides by sequence similarity to known ligands of SH2 and PTB domains [26]. The partition around medoids (PAM) based clustering of these phosphopeptides, which were transformed as vectors of 11 SH2 and PTB weight matrices led to four novel motifs (in addition to 16 known motifs). Subsequently, SILAC was employed to confirm 15 novel phosphosite mediated interactions with SH2 or PTB. Rinner et al analyzed protein complexes through the integration of label-free, quantitative mass spectrometry and computational analysis [95]. Using this approach the authors identified known and novel interactors of FoxO3A with 14-3-3 proteins.

5. Data analysis, visualization and presentation in quantitative proteomics

Bioinformatics tools and resources for genomics, functional genomics, and structural genomics efforts are aplenty, and have a large user and developer base. MS-based proteomics is a recent member of the 'omics' clan and is starting to attract considerable attention from the biomedical informatics community. Many of the analysis algorithms and tools developed for functional genomics are being leveraged in proteomics related bioinformatics applications. Most common among them are the tools for unsupervised and supervised clustering and visualization. Hierarchical clustering and *k*-means algorithm especially have found numerous applications in modern proteomics due to their simple algorithmic assumptions, and intuitively clear and interpretable visualization [96]. As complex proteomics data sets have started to be generated, they require sophisticated analytical treatments and visualization capabilities. With high dimensionality comes the need of statistically robust machine learning techniques [97] which are now being applied in proteomics and statistical and graphical aspects of these data sets are being explored. The R framework, which in the past decade has emerged as the lingua franca for open source statistical and machine learning research, has already permeated the microarray community through its bioinformatics chapter called Bioconductor [40,41]. We have adapted R and Bioconductor for most of our proteo-bioinformatics analyses due to its expansive capabilities in data mining, statistical procedures, and excellent graphics. Moreover, R is open source; enables quick prototyping and thereby delivers results faster without the need to reinvent the wheel. We foresee that as bioinformatics for proteomics matures, R and Bioconductor will become an indispensable toolkit in proteomics research. Additionally, proteomics will continue to benefit from experiences and challenges faced by the microarray and related gene expression communities. This process, too, is facilitated by an open, active and collaborative environment such as R and Bioconductor embody.

6. Challenges and perspectives

In this post-genomic era proteomics along with sister 'omics' disciplines provides the foundations on which the promise of systems biology will be realized. The next steps in this direction are consolidation and integration of data sets and information across different layers of the 'omics' hierarchy, ultimately leading to physiologically accurate and clinically relevant in silico models of biological processes and systems. Future bioinformatics activities in proteomics will increasingly focus on integrative systems biology, as there are many aspects of cellular function which can only be answered by adopting this approach. Currently, the challenges one faces during such analyses involve seemingly mundane issues such as the mapping of identified proteins to genomic and microarray identifiers. Due to many-to-many mapping between proteins and their corresponding genes this problem is further compounded. There are now frameworks like BioMart [98] which provide mapping from protein to genomic identifiers, but which still harbor inconsistencies due to error propagation from legacy issues during automated data integration. Biomedical research until recently has been genomic oriented so most of the ontologies and annotation databases are still 'gene centric', and often fail to capture protein specific characteristic and functions, which arise from the greater diversity of proteins compared to genes. On a more abstract level, it is clear that proteomics and other large-scale 'post-genomic' technologies will profit tremendously from further investments into accurate and detailed Gene (or more accurately Protein) Ontologies. Indeed, the more comprehensive such ontologies are, the easier it will be to infer function directly from quantitative proteomics data sets.

Proteomics has finally started to attract the attention of key stakeholders in bioinformatics database development process and we envisage that in future these resources will be tuned to the needs of proteomics research. More scientific investments are required to generate unified and comprehensive proteomics database along the lines of microarray databases and in this respect the recent strong interest of the National Institutes of Health (NIH) in addition the longstanding one of the European Bioinformatics Institute (EBI) are positive factors. However, the key factor here is how to control the accumulation of false positives, for example in organellar databases. As mentioned above for genome annotation, we envision that only extremely accurate proteomics data sets, and not any data sets contributed by the proteomics community, should be used for this purpose.

As proteomics is further integrated into the fabric of contemporary biomedical ecosystem it will also become a key driver of bioinformatics research. Although from the data structure and analysis points of view parallels can be drawn with genomics and transcriptomics data sets, proteomics data sets are still unique in their constitution and underlying assumptions. Therefore, complete exploitation and optimal harnessing of this uniqueness will necessitate development of novel analytical and bioinformatics approaches. In due course proteomic related bioinformatics activities will create its own identity and evolve as a new stream of biomedical informatics research with exciting avenues.

Acknowledgments

We would like to thank all the members of the department of proteomics and signal transduction at Max Planck Institute for Biochemistry. Special thanks to Chunaram Choudhary, Jesper V. Olsen, and Christian Kelstrup for inputs on the contents and the figures. This work was partially supported by the EU research directorate through the PROSPECTS and INTERACTION PROTEOME projects (7th and 6th framework, respectively).

References

[1] Wilkins, M.R. et al. (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Biotechnology (NY) 14, 61–65.
[2] Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. Nature 422, 198–207.
[3] Mann, M. and Kelleher, N.L. (2008) Precision proteomics: the case for high resolution and high mass accuracy. Proc. Natl. Acad. Sci. USA 105, 18132–18138.
[4] Malmstrom, J., Lee, H. and Aebersold, R. (2007) Advances in proteomic workflows for systems biology. Curr. Opin. Biotechnol. 18, 378–384.
[5] Colinge, J. and Bennett, K.L. (2007) Introduction to computational proteomics. PLoS Comput. Biol. 3, e114.
[6] Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372.
[7] Hagen, J.B. (2000) The origins of bioinformatics. Nat. Rev. Genet. 1, 231–236.
[8] Perez-Iratxeta, C., Andrade-Navarro, M.A. and Wren, J.D. (2007) Evolving research trends in bioinformatics. Brief Bioinform. 8, 88–95.
[9] Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M. and Yates 3rd, J.R. (1999) Direct analysis of protein complexes using mass spectrometry. Nat. Biotechnol. 17, 676–682.
[10] Mallick, P. et al. (2007) Computational prediction of proteotypic peptides for quantitative proteomics. Nat. Biotechnol. 25, 125–131.
[11] Fusaro, V.A., Mani, D.R., Mesirov, J.P. and Carr, S.A. (2009) Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nat. Biotechnol. 27, 190–198.
[12] Lange, V., Picotti, P., Domon, B. and Aebersold, R. (2008) Selected reaction monitoring for quantitative proteomics: a tutorial. Mol. Syst. Biol. 4, 222.
[13] Deutsch, E.W., Lam, H. and Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep. 9, 429–434.
[14] Desiere, F. et al. (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol. 6, R9.
[15] Baerenfaller, K. et al. (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science 320, 938–941.
[16] Gupta, N. et al. (2008) Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. Genome Res. 18, 1133–1142.
[17] Pang, J.X., Ginanni, N., Dongre, A.R., Hefta, S.A. and Opitek, G.J. (2002) Biomarker discovery in urine by proteomics. J. Proteome Res. 1, 161–169.
[18] Liu, H., Sadygov, R.G. and Yates 3rd, J.R. (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. 76, 4193–4201.
[19] Kislinger, T. et al. (2006) Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell 125, 173–186.
[20] Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J. and Mann, M. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol. Cell Proteomics 4, 1265–1272.
[21] Lu, P., Vogel, C., Wang, R., Yao, X. and Marcotte, E.M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. 25, 117–124.
[22] Witze, E.S., Old, W.M., Resing, K.A. and Ahn, N.G. (2007) Mapping protein post-translational modifications with mass spectrometry. Nat. Meth. 4, 798–806.
[23] Jensen, O.N. (2006) Interpreting the protein language using proteomics. Nat. Rev. Mol. Cell Biol. 7, 391–403.
[24] Macek, B., Mann, M. and Olsen, J.V. (2009) Global and site-specific quantitative phosphoproteomics: principles and applications. Ann. Rev. Pharmacol. Toxicol. 49, 199–221.
[25] Schwartz, D. and Gygi, S.P. (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. 23, 1391–1398.
[26] Miller, M.L., Hanke, S., Hinsby, A.M., Friis, C., Brunak, S., Mann, M. and Blom, N. (2008) Motif decomposition of the phosphotyrosine proteome reveals a new N-terminal binding motif for SHIP2. Mol. Cell Proteomics 7, 181–192.
[27] Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Oroshi, M. and Mann, M. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol. 8, R250.
[28] Linding, R. et al. (2007) Systematic discovery of in vivo phosphorylation networks. Cell 129, 1415–1426.
[29] Adachi, J., Kumar, C., Zhang, Y., Olsen, J.V. and Mann, M. (2006) The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. Genome Biol. 7, R80.
[30] Brunner, E. et al. (2007) A high-quality catalog of the *Drosophila melanogaster* proteome. Nat. Biotechnol. 25, 576–583.
[31] Shi, R., Kumar, C., Zougman, A., Zhang, Y., Podtelejnikov, A., Cox, J., Wisniewski, J.R. and Mann, M. (2007) Analysis of the mouse liver proteome using advanced mass spectrometry. J. Proteome Res. 6, 2963–2972.

[32] Bonaldi, T., Straub, T., Cox, J., Kumar, C., Becker, P.B. and Mann, M. (2008) Combined use of RNAi and quantitative proteomics to study gene function in *Drosophila*. Mol. Cell 31, 762–772.

[33] Ashburner, M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

[34] Hunter, S. et al. (2009) InterPro: the integrative protein signature database. Nucl. Acids Res. 37, D211–D215.

[35] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. Nucl. Acids Res. 32, D277–D280.

[36] Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57.

[37] Zeeberg, B.R. et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 4, R28.

[38] Shannon, P. et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504.

[39] Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21, 3448–3449.

[40] Gentleman, R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.

[41] Team, R.D.C. (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[42] Adachi, J., Kumar, C., Zhang, Y. and Mann, M. (2007) In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. Mol. Cell Proteomics 6, 1257–1273.

[43] Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

[44] Aerts, S. et al. (2006) Gene prioritization through genomic data fusion. Nat. Biotechnol. 24, 537–544.

[45] Ishikura, S., Bilan, P.J. and Klip, A. (2007) Rabs 8A and 14 are targets of the insulin-regulated Rab-GAP AS160 regulating GLUT4 traffic in muscle cells. Biochem. Biophys. Res. Commun. 353, 1074–1079.

[46] Mootha, V.K. et al. (2003) Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. Cell 115, 629–640.

[47] Foster, L.J., de Hoog, C.L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V.K. and Mann, M. (2006) A mammalian organelle map by protein correlation profiling. Cell 125, 187–199.

[48] Calvo, S. et al. (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. Nat. Genet. 38, 576–582.

[49] Pagliarini, D.J. et al. (2008) A mitochondrial protein compendium elucidates complex I disease biology. Cell 134, 112–123.

[50] Graumann, J. et al. (2008) Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins. Mol. Cell Proteomics 7, 672–683.

[51] von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P. and Snel, B. (2003) STRING: a database of predicted functional associations between proteins. Nucl. Acids Res. 31, 258–261.

[52] Chatr-aryamontri, A. et al. (2008) MINT and IntAct contribute to the second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. Genome Biol. 9 (Suppl. 2), S5.

[53] Peri, S. et al. (2004) Human protein reference database as a discovery resource for proteomics. Nucl. Acids Res. 32, D497–D501.

[54] Ong, S.E. and Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. Nat. Chem. Biol. 1, 252–262.

[55] Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A. and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell Proteomics 1, 376–386.

[56] Mann, M. (2006) Functional and quantitative proteomics using SILAC. Nat. Rev. Mol. Cell Biol. 7, 952–958.

[57] de Godoy, L.M., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Frohlich, F., Walther, T.C. and Mann, M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 455, 1251–1254.

[58] Merchant, M. and Weinberger, S.R. (2000) Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. Electrophoresis 21, 1164–1177.

[59] Diamandis, E.P. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. Mol. Cell Proteomics 3, 367–378.

[60] Baggerly, K.A., Morris, J.S. and Coombes, K.R. (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 20, 777–785.

[61] Steen, H. and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. 5, 699–711.

[62] Sadygov, R.G., Cociorva, D. and Yates 3rd, J.R. (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat. Meth. 1, 195–202.

[63] Listgarten, J. and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography–tandem mass spectrometry. Mol. Cell Proteomics 4, 419–434.

[64] Mueller, L.N., Brusniak, M.Y., Mani, D.R. and Aebersold, R. (2008) An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J. Proteome Res. 7, 51–61.

[65] Clarke, R., Ressom, H.W., Wang, A., Xuan, J., Liu, M.C., Gehan, E.A. and Wang, Y. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat. Rev. Cancer 8, 37–49.

[66] Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.

[67] Gygi, S.P., Rochon, Y., Franza, B.R. and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. Mol. Cell Biol. 19, 1720–1730.

[68] Orntoft, T.F., Thykjaer, T., Waldman, F.M., Wolf, H. and Celis, J.E. (2002) Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas. Mol. Cell Proteomics 1, 37–45.

[69] Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008) Widespread changes in protein synthesis induced by microRNAs. Nature 455, 58–63.

[70] Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P. and Bartel, D.P. (2008) The impact of microRNAs on protein output. Nature 455, 64–71.

[71] Alizadeh, A.A. et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511.

[72] van't Veer, L.J. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536.

[73] Sharan, R. et al. (2005) Conserved patterns of protein interaction in multiple species. Proc. Natl. Acad. Sci. USA 102, 1974–1979.

[74] Mikkelsen, T.S. et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448, 553–560.

[75] Pan, C., Kumar, C., Bohl, S., Klingmueller, U. and Mann, M. (2009) Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type specific functions. Mol. Cell Proteomics 8, 443–450.

[76] Blagoev, B., Ong, S.E., Kratchmarova, I. and Mann, M. (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. Nat. Biotechnol. 22, 1139–1145.

[77] Andersen, J.S., Lam, Y.W., Leung, A.K., Ong, S.E., Lyon, C.E., Lamond, A.I. and Mann, M. (2005) Nucleolar proteome dynamics. Nature 433, 77–83.

[78] Zhang, Y., Wolf-Yadlin, A., Ross, P.L., Pappin, D.J., Rush, J., Lauffenburger, D.A. and White, F.M. (2005) Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. Mol. Cell Proteomics 4, 1240–1250.

[79] Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P. and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127, 635–648.

[80] Yates 3rd, J.R., Gilchrist, A., Howell, K.E. and Bergeron, J.J. (2005) Proteomics of organelles and large cellular structures. Nat. Rev. Mol. Cell Biol. 6, 702–714.

[81] Andersen, J.S. and Mann, M. (2006) Organellar proteomics: turning inventories into insights. EMBO Rep. 7, 874–879.

[82] Dunkley, T.P. et al. (2006) Mapping the *Arabidopsis* organelle proteome. Proc. Natl. Acad. Sci. USA 103, 6518–6523.

[83] Gingras, A.C., Gstaiger, M., Raught, B. and Aebersold, R. (2007) Analysis of protein complexes using mass spectrometry. Nat. Rev. Mol. Cell Biol. 8, 645–654.

[84] Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113.

[85] de Lichtenberg, U., Jensen, L.J., Brunak, S. and Bork, P. (2005) Dynamic complex formation during the yeast cell cycle. Science 307, 724–727.

[86] Krogan, N.J. et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440, 637–643.

[87] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein- protein interactions. Nature 417, 399–403.

[88] Yu, H. et al. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322, 104–110.

[89] Vermeulen, M., Hubner, N.C. and Mann, M. (2008) High confidence determination of specific protein–protein interactions using quantitative mass spectrometry. Curr. Opin. Biotechnol. 19, 331–337.

[90] Blagoev, B., Kratchmarova, I., Ong, S.E., Nielsen, M., Foster, L.J. and Mann, M. (2003) A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. Nat. Biotechnol. 21, 315–318.

[91] Schulze, W.X. and Mann, M. (2004) A novel proteomic screen for peptide–protein interactions. J. Biol. Chem. 279, 10756–10764.

[92] Ranish, J.A., Yi, E.C., Leslie, D.M., Purvine, S.O., Goodlett, D.R., Eng, J. and Aebersold, R. (2003) The study of macromolecular complexes by quantitative proteomics. Nat. Genet. 33, 349–355.

[93] Schulze, W.X., Deng, L. and Mann, M. (2005) Phosphotyrosine interactome of the ErbB-receptor kinase family. Mol. Syst. Biol. 1, 0008.

[94] Hanke, S. and Mann, M. (2009) The phosphotyrosine interactome of the insulin receptor family and its substrates IRS-1 and IRS-2. Mol. Cell Proteomics 8, 519–534.

[95] Rinner, O., Mueller, L.N., Hubalek, M., Muller, M., Gstaiger, M. and Aebersold, R. (2007) An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. Nat. Biotechnol. 25, 345–352.

[96] Meunier, B., Dumas, E., Piec, I., Bechet, D., Hebraud, M. and Hocquette, J.F. (2007) Assessment of hierarchical clustering methodologies for proteomic data mining. J. Proteome Res. 6, 358–366.

[97] Hastie, T., Tibshirani, R. and Friedman, J. (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, New York.

[98] Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G. and Kasprzyk, A. (2009) BioMart – biological queries made easy. BMC Genomics 10, 22.