ISSN 0006-2979, Biochemistry (Moscow), 2006, Vol. 71, No. 10, pp. 1060-1072. © Pleiades Publishing, Inc., 2006.
Published in Russian in Biokhimiya, 2006, Vol. 71, No. 10, pp. 1312-1327.

REVIEW

Quantitative Proteomics and Its Applications for Systems Biology

S. Ivakhno[1]* and A. Kornelyuk[2]

[1]*University of Edinburgh, School of Informatics, Appleton Tower, Crichton Str., Edinburgh, EH8 9LE, Scotland, UK; fax: (44) 131-651-1426; E-mail: s0567096@sms.ed.ac.uk*
[2]*Institute of Molecular Biology and Genetics, Department of Protein Engineering, 150 Academician Zabolotny Str., Kiev 03143, Ukraine; fax: (38) 044-526-0759; E-mail: kornelyuk@imbg.org.ua*

Received September 6, 2005
Revision received January 8, 2006

Abstract—Here we discuss the current state of research in the rapidly growing field of quantitative proteomics and its applications to systems biology. Quantitative proteomics can be successfully used for characterizing alterations in protein abundance, finding novel protein–protein and protein–peptide interactions, investigating formation of large macromolecular complexes, and elucidating temporal changes in organellar protein composition and phosphorylation in signal transduction cascades. Further, quantitative proteomics can directly compare activation of entire signaling networks in response to individual stimuli and discover critical differences in their circuits that account for alterations of cell response. Maturation of proteomic bioinformatics applications and continuous improvements in proteomics and related genomics and transcriptomics technologies now allows us to investigate cellular mechanisms at the integrative system level.

DOI: 10.1134/S0006297906100026

Key words: quantitative proteomics, stable isotope labeling, tandem mass spectrometry, proteome informatics, systems biology, protein interactions, DNA microarray

The increasing availability of fully or partially sequenced genomes for a variety of organisms has lead us to a new era of biological research aimed towards the systems-level understanding of biological processes. This new approach in biology, termed systems biology, has as its goal an eventual understanding of the information flow from genes to biological function, with explicit treatment of an organism's response to perturbations at the molecular level [1]. Systems biology attempts to integrate data from diverse high-throughput and bioinformatics technologies, such as genome sequencing, DNA microarray and SAGE (Serial Analysis of Gene Expression)-based approaches for mRNA profiling, SNP (Single Nucleotide Polymorphism) genotyping, high-throughput RNAi screens and gene-knockdown, yeast two-hybrid system, chromatin immunoprecipitation, text mining of biological literature, and LC-MS/MS-based proteomics to construct explicit mathematical models of a cell's regulatory and metabolic mechanisms (Fig. 1). One potential promise of systems biology is to create a blueprint in which more conventional one gene/protein investigations can be carried separately and later integrated into the whole model along with high-throughput biological data. Ultimately, this ambitious undertaking will rely heavily on our capability to examine in a massively parallel fashion the identity, concentration, function, and interaction of a wide variety of biological macromolecules [2].

High-throughput DNA sequencing techniques have already enabled us access to fully characterized genomes of many organisms. These genome data, stored in large databases containing the nucleotide sequence code and gene annotations, provide us with basic foundations for studying biological systems. A lot of useful information about particular species can be inferred just from exploring its genome, which is especially true for prokaryotic genomes [3]. But sequence information alone is insufficient for understanding the biology of a given organism. Data on mRNA expression, protein interaction, protein localization, and dynamics of signaling pathways is needed before we can appreciate the computational complexi-

Abbreviations: CDIT) culture derived isotope tag; 2DE) two-dimensional electrophoresis; EGF) epidermal growth factor; EGFR) epidermal growth factor receptor; ICAT) isotope-coded affinity tag; LC-MS/MS) liquid chromatography/tandem mass spectrometry; MALDI-MS) matrix-assisted laser desorption-ionization mass spectrometry; PDGF) platelet derived growth factor; RNAi) RNA interference; SILAC) stable isotope labeling with amino acids in cell culture; TBP) TATA binding protein; XML) extensible markup language.

* To whom correspondence should be addressed.

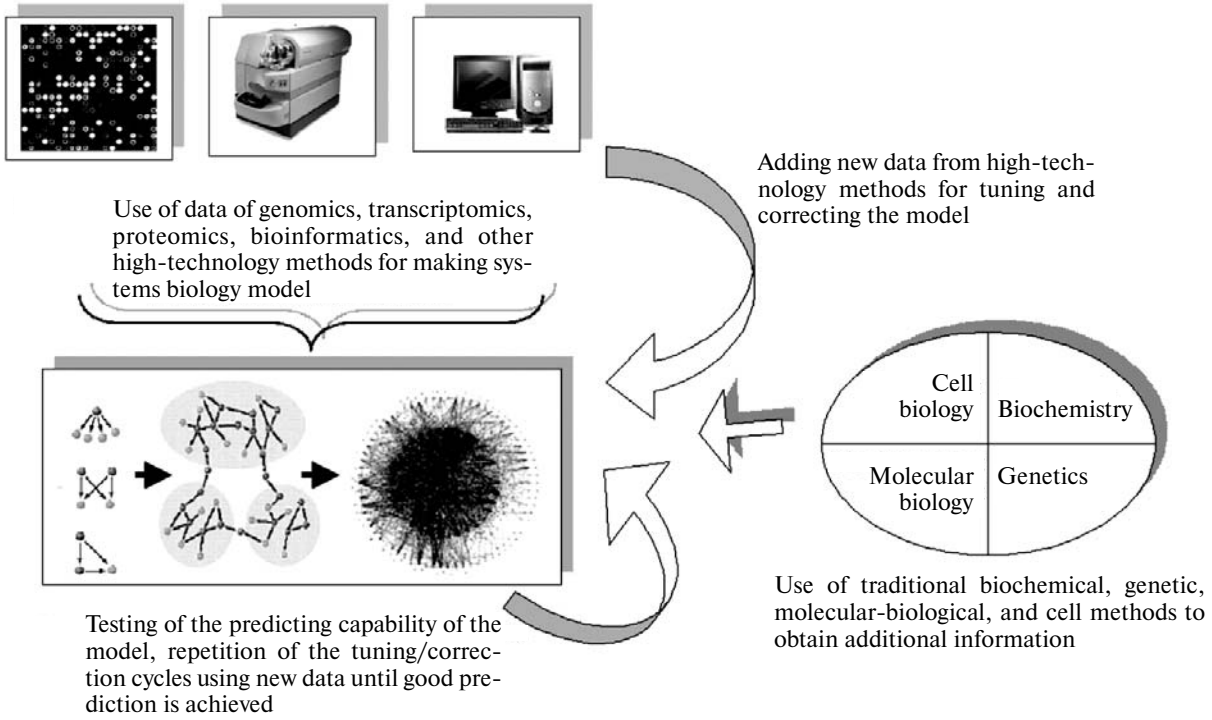

Fig. 1. Principles of systems biology model building. A model is built from analyzing and comparing large-scale data obtained by genomic, proteomic, and bioinformatic methods. Usually a model has several levels of complexity, incorporating information on distinct motifs, subnetworks, and networks of protein interactions, signaling pathways, and transcription regulation. An additional level of refinement, not available in large-scale datasets, comes from data on individual proteins/genes obtained by conventional biochemical and genetic investigation. However, a model is not encumbered with unnecessary details and only incorporates information that is necessary to produce correct and relevant predictions.

ty of living cells. The power of high-throughput approaches in functional genomics is exemplified by DNA microarray technologies. They have been successfully used for elucidation of a cell's transcriptional response to various perturbations [4], identification of signature genes in different cancer subtypes [5], and metabolic profiling [6]. However, because proteins are the predominant functional macromolecules, the identity of potentially expressed proteins at a given time defines the functional state of the cell. Since significant molecular control is exercised at the level of translation initiation, post-translational modifications, and mRNA turnover, the investigation of proteome dynamics is a vital requirement for understanding of the cell's regulatory mechanism.

One of the most promising approaches for measuring proteome dynamics, quantitative proteomics, relies on the use of stable isotope labeling techniques for relative quantification of protein abundance. In this article, we will review the major technologies for mass spectrometry based quantitative proteomics and its contributions to the field of systems biology. Several excellent and comprehensive reviews concerning proteome analysis, interpretations of tandem mass spectra, and applications of mass spectrometry based proteomics have already been written [7-11]. Our objective here is not a comprehensive survey of tandem mass spectrometry applications in proteomics, but instead a review of the current state of research in the rapidly growing field of quantitative proteomics. We first describe the most popular techniques and recent developments for relative quantitative proteome analysis by isotope labeling and mass spectrometry, where we will also consider their relative merits and pitfalls. We then discuss principles of quantitative proteomics data analysis and its role in proteome bioinformatics research. The most widely used programs for tandem mass spectra database searches and protein quantification will be described along with relevant details of their performance. Finally, current application of quantitative proteomics for analysis of protein abundance, finding novel protein interactions in macromolecular complexes, investigation of dynamic changes in organelles' protein composition, and temporal dynamics in the signal transduction cascades will be described.

STABLE ISOTOPE LABELING TECHNIQUES IN QUANTITATIVE PROTEOMICS

Tandem mass spectrometry based methods for quantitative proteomics are fundamentally different from the

more conventional methods of two-dimensional electrophoresis (2DE) and matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS). The later rely upon the rather irreproducible nature of 2DE separation followed by identification and quantification of protein spots using silver staining or fluorescent dyes and MALDI-MS. Unfortunately, the large range of protein expression levels limits the ability of the 2DE-MS approach to analyze proteins of medium to low abundance, and thus the potential of this technique for proteome analysis is likewise limited [12].

Tandem mass spectrometry based quantitative proteomics (quantitative proteomics) offers an alternative approach to comparing protein abundances in different states with much broader application potential. The method is based on a simple approach of relative protein quantification, which is similar in statistical design to cDNA microarray protocols utilizing competitive mRNA hybridization. Both methods perform quantification indirectly by computing ratios of abundances in different states (this similarity in data representation allows comparison of data generated by two technologies [13]). In quantitative proteomics, proteins from various mixtures can be distinguished in the MS spectrum by means of stable-isotope tags—chemical groups or natural amino acids, which are encoded in two states: one labeled with a stable isotope and another that is not. Stable isotope labeling does not change physical or chemical properties of proteins/peptides apart from conferring distinct mass difference (m/z) in their MS spectrum. Typically, peptides labeled with stable isotope show 6-10 dalton shifts from natural unlabeled ones but co-elute on a chromatographic column. Most quantitative proteomics methods that use stable isotopes include the following four steps: 1) differential isotopic labeling of two or more separate protein mixtures; 2) digestion of the combined labeled protein mixtures followed by separation of the resulting peptides by liquid chromatography; 3) analysis of the separated peptides by automated tandem mass spectrometry; and 4) automated database searching to identify the peptide sequences (and hence the proteins from which they were derived) followed by computation of relative protein abundances from the MS data. Thus, tandem mass spectra are used to determine the identity of peptides and mass spectra are used for relative protein quantification (c.f. iTRAQ^{TM} reagent from Applied Biosystems (USA) uses a different strategy discussed below). In the next section, we will consider these steps in more detail, concentrating on various methodologies and protocols and exploring their strengths and weaknesses.

Two major variants of stable isotope labeling have been reported: metabolic and chemical labeling. In metabolic labeling, cells usually incorporate stable isotope through growth on isotope-enriched media. In chemical labeling isotopes are incorporated by attachment of various chemical groups after cell lysis and protein purification. These methods can also be considered as *in vivo* versus *in vitro* labeling. As proteins are combined at the earliest possible step of cell culture, *in vivo* labeling is characterized by smaller experimental error [14].

It should be noted that quantitative proteomics via stable isotope labeling is not the only mean of comparing protein abundances. Several other approaches are currently undergoing technological development and should be soon available for use. Among them, we should mention (i) the method of measuring the number of times a given peptide appears on the MS/MS spectrum and (ii) the relative quantification based on the ratios of areas under ion current chromatograms for a single peptide in two different LC-MS runs [15, 16]. The salient feature of these approaches is that they require high reproducibility between different chromatographic runs, and therefore their applicability is limited to the comparison of similar samples.

Metabolic stable-isotope labeling. The first approach developed for *in vivo* stable isotope labeling utilized media containing ^{15}N. In this procedure, yeast cultures are grown in two separate media, one containing ^{15}N. Cells are then pooled together, proteins extracted, separated by gel-electrophoreses, digested to peptides, and quantified on MS [17]. Modified versions of this method also exist, which directly connects microscale two-dimensional chromatography to tandem mass spectrometry (termed MudPit). This latter approach was successfully used for quantification of yeast proteins [18]. Cells grown in media enriched in ^{15}N were used as an internal standard for all quantitative measurements. These internal standards were mixed with cells from different conditions early during the sample preparation so that any protein loss during cell lysis, digestion, and measurement were accounted for by their respective ^{15}N-labeled protein. Since this method uses an internal standard, comparison of protein abundances in different samples is achieved by estimating ratios of ratios [19]. Unfortu-nately, methods which use bare stable isotopes have several drawbacks. First, these methods, while applicable to bacteria and yeasts, do not perform well on mammalian cell cultures, which poorly incorporate stable isotope. Although some investigators have reported alternative procedures for incorporation of stable isotopes into higher eukaryotes, for instance by feeding them with metabolically labeled yeast, these methods still lend themselves poorly to standardization [20]. Second, different proteins incorporate unequal amount of stable isotopes equivalent to the number of nitrogen atoms they posses. The direct consequence of this is that labeled and unlabeled peptides have a variable mass shift in the MS spectra, which complicates automatic comparison of peptide abundance.

The second methods of *in vivo* labeling, SILAC, take into account shortcomings of ^{15}N labeling [21, 22]. SILAC (Fig. 2a) relies on the incorporation of amino

acids with substituted stable isotopic nuclei. In this approach, two groups of cells are grown in culture media that are identical except that the first contains the "light" and the other the "heavy" form of a particular amino acid (for example, L-leucine or deuterated L-leucine). By selecting for labeling of essential amino acids, cells are forced to use them and consequently 100% efficiency in label incorporation can be achieved. An additional advantage of using amino acids is that they can be labeled by several different stable isotopes, for instance, ^{13}C and ^{15}N. This allows three cell cultures to be processed in parallel (one encoded with ^{13}C-labeled amino acid, one with ^{13}C and ^{15}N, and an unlabeled one), thus increasing the number of conditions that can be compared in one experiment. One serious limitation of SILAC is that the method cannot be used to estimate protein abundance ratios for tissue samples, for example in identification of protein markers in cancer tissues [23].

An alternative approach based on the use of the culture derived isotope tags (CDITs) for quantitative tissue proteomics has recently been proposed to overcome the limitation of SILAC [24]. In this method, stable isotope-labeled cultured cells are used as global internal standards (Fig. 2b). Tissue samples to be compared are mixed with cultured cells early in the process to control for variations during sample preparation. After protein extraction and separation, digested proteins are analyzed by mass spectrometry to identify and quantify peptides. The ratio between the two isotopic distributions (one from a tissue sample and one from isotope-labeled cells) can then be determined from the MS spectra. Changes in protein level in two tissue samples are estimated by calculating the ratio of two ratios, canceling in this way systematic errors of internal standard intensities (two or three different amounts of labeled cultured cells can be added into tissue samples to increase the number of quantified proteins). Ishihama et al. [24] showed that cells in cell culture express more than 97% of proteins found in tissues (using mouse brain and Neuro2A cell culture as an example), and the ratio of a target peptide, which does not have a corresponding labeled peak in cultured cells, can be obtained by using the peak ratio against an isotope-

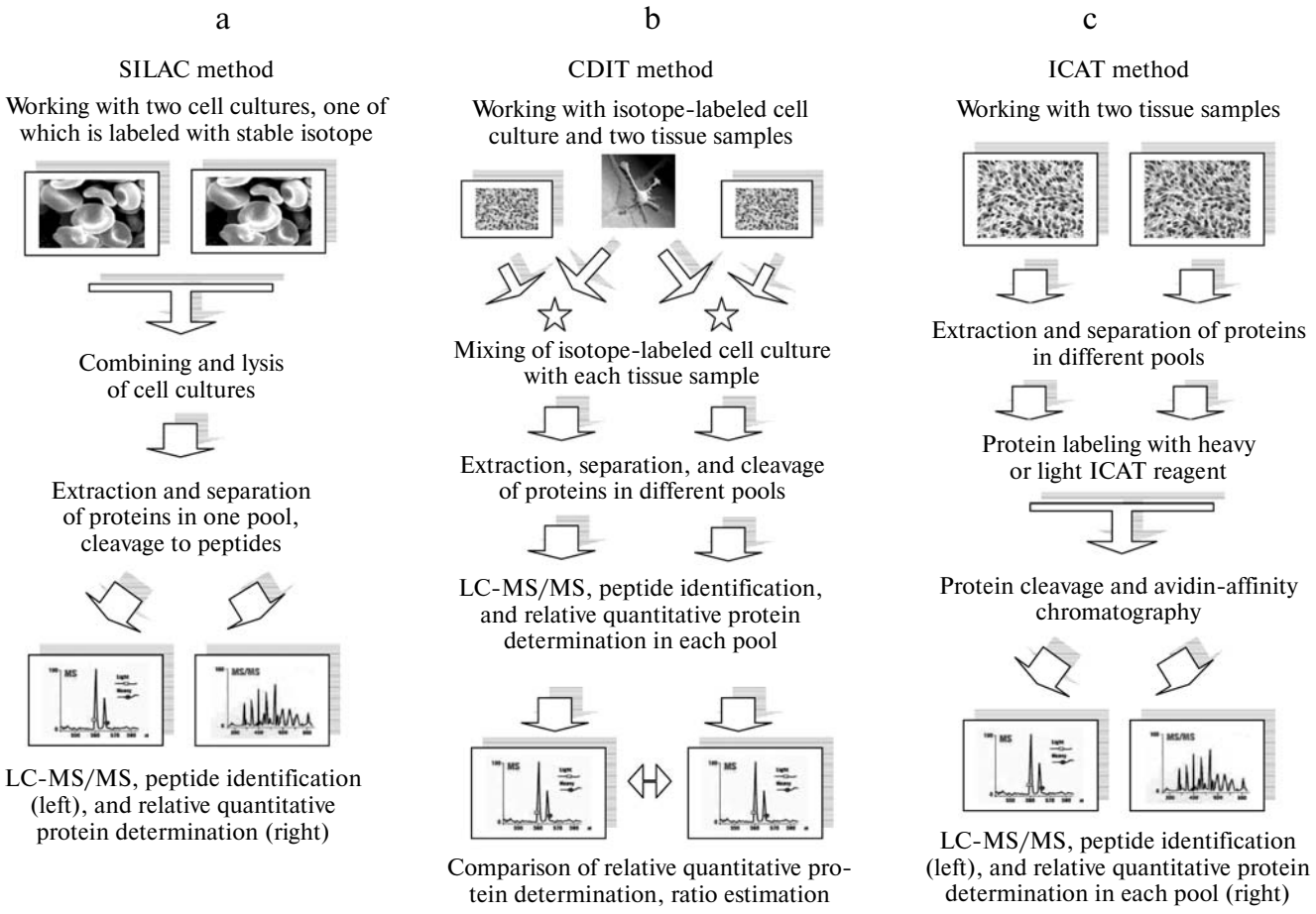

Fig. 2. Schematic representations of quantitative proteomics methods: SILAC (stable isotope labeling with amino acids in cell culture) (a), CDIT (culture derived isotope tag) (b), and ICAT (isotope-coded affinity tag) (c). The difference between methods resides in the point where stable isotope label is introduced, the nature of stable isotope tag, specific purifications steps necessitated by the tag, and algorithms for calculation of relative peptide abundance.

labeled peptide of different sequence, but with the closest retention time in LC/MS.

Chemical methods of stable isotope incorporation. *In vitro* labeling methods involve incorporation of the stable isotopic tags onto selective sites on peptides via *in vitro* chemical reactions. Three variants of *in vitro* labeling methods exist: those that isotopically label target peptides at (i) amino- or (ii) carboxyl-terminal and those (iii) that label specific amino acid residues, such as cysteine, lysine, tyrosine, etc.

The earliest chemical methods for stable isotope incorporation (ICAT) was based on the use of cysteine-specific, isotopically labeled biotin affinity tags (containing eight deuterium atoms) covalently linked to proteins in paired proteome samples [25]. Here, the tagged protein pools were combined and digested to peptides (Fig. 2c). Biotinylated peptides (containing cysteine) were then isolated from the complex peptide mixture by avidin affinity chromatography and subsequently analyzed by LC-MS/MS. As with other isotope labeling techniques, protein abundance ratios were calculated from relative peak intensities of light and heavy versions of labeled peptides. The early version of ICAT had two potential disadvantages. First, the biotin affinity tag remained linked to the peptides throughout the analysis, leading to large shifts in MS and changes in MS/MS spectra relative to unlabeled peptides. Further, ICAT reagent exhibits relatively large differences in chromatographic elution times between the deuterated and nondeuterated label versions. These deficiencies have been corrected through development of a new cleavable-ICAT reagent, which uses modified cleavable biotin affinity tags labeled with C^{13}. Unfortunately, cleavable-ICAT introduces an additional purification step, making the whole process irreproducible and time consuming. A significant improvement to the ICAT strategy, which reduces the effect of the multiple purification steps, is the incorporation of solid-phase capture in the isotope tagging approach [26]. In this method, labeled leucine covalently linked to a solid-phase support via a photocleavable linker is used to isolate and label cysteinyl peptides in global proteome samples. After combining beads from paired samples, the bound peptides are released by photolysis and analyzed by LC-MS/MS. Unlike the original ICAT approach, the only chemical modification that remains on peptides at the time of LC-MS/MS analysis is an isotopically labeled (or unlabeled) leucine residue. Apparent advantages of the method are enhanced sensitivity, reduced sample handling, and facilitation of extensive sample washing protocols *prior to* peptide elution [26].

In addition to cysteine labeling by the ICAT reagent, other amino acids have also been reported for enrichment. Kuyama et al. described an approach for enriching tryptophan-containing peptides by modifying the tryptophan residues with isotopic $^{12}C/^{13}C$ 2-nitrobenzenesulfenyl chloride (NBSCl) [27]. Similarly, isotopically labeled phosphotyrosine-containing peptides were used as isotopic tags after enrichment through an anti-phosphotyrosine antibody-based purification procedure [28]. Another effort to target specific amino acid-containing peptides, reported by Cagney et al., incorporated labeled lysine-containing peptides by a method termed mass-coded abundance tagging (MCAT) [29]. The MCAT procedure specifically labels the ε-amino group of lysine by guanidinylation using O-methylisourea. The labeled sample is then compared with the non-labeled sample to determine relative quantities. This is not a strict stable isotope labeling approach, since the internal standard (unlabeled peptide) is chemically different from the labeled sample by more than just isotopic atoms, and the physicochemical difference between the labeled and unlabeled peptides may reduce the accuracy of the quantification.

The ICAT method selectively purifies the cysteine-containing peptides and thus dramatically reduces sample complexity, allowing detection and quantification of non-abundant proteins. In the human proteome, about 26.6% of the total tryptic peptides contain at least one cysteine residue and they cover 96.1% of the human proteome [30]. Therefore, application of cysteine-based enrichment theoretically reduces sample complexity by at least fourfold while missing fewer than 4% of the proteins. To obtain complete protein coverage, ideal targets for introducing an isotopic tag are the N- or C-termini of peptides. N-Terminal isotope-encoded tagging (NIT) specifically incorporates stable isotopes at the N-termini of peptides by firstly converting the lysine residue to homoarginine using O-methylisourea [31]. An alternative approach introduces stable isotopes specifically to the C-termini of peptides via biochemical reactions using enzymes such as trypsin. During trypsin proteolysis, ^{16}O or ^{18}O isotopes can be incorporated into the C-termini of peptides in the presence of ^{16}O- or ^{18}O-containing water. The relative quantity of proteins is determined by the ratio of ion intensities of ^{16}O- or ^{18}O-labeled peptides measured by MS [32]. A pitfall of this method is that the possible loss or incomplete incorporation of the isotopic labels due to the use of enzyme complicates the quantification. While N-terminal or C-terminal peptide labeling approaches are expected to have complete protein coverage, it remains a challenge to apply them to global proteome-wide quantitative profiling due to the high sample complexity.

A completely different approach to protein quantification is taken by the iTRAQ (isobaric tags for relative and absolute quantification) reagents developed by Applied Biosystems (USA), where protein abundance ratios are calculated from MS/MS spectrum [33]. These reagents target N-termini and the lysine residues of proteins and allow quantification of the relative protein expression in up to four different experimental conditions. Each reagent contains three groups: reporter, bal-

ance, and reactive groups. The reporter groups of the four iTRAQ reagents have molecular weights of 114, 115, 116, and 117 daltons depending on differential isotopic combinations of $^{12}C/^{13}C$ and $^{16}O/^{18}O$ in each individual reagent. The balance group ranges in mass from 28 to 31 daltons to ensure the combined mass of the reporter and balance groups remains constant (145 daltons) for all four reagents. Therefore, peptides labeled with different isotopes are isobaric and are chromatographically indistinguishable, a factor that is important for accurate quantification. During collision-induced dissociation, the reporter group ions fragment from the backbone peptides, displaying distinct masses of 114 to 117 daltons. The intensity of each of these peaks represents the quantity of small reporter group fragment and is proportional to the amount of peptide in the sample. Other peaks in the MS/MS spectrum are used to identify peptide sequences. The idea of the iTRAQ protocol is quite simple—the quantity of each peptide in each sample is the peak intensity of its corresponding reporter fragment in the MS/MS spectrum. Pro QUANT software for peptide quantification with these reagents is available commercially from Applied Biosystems.

The most significant advantage of this technology is that it allows labeling of up to four different samples within a single experiment. It is useful for quantifying proteins from multiplex samples, such as those in a time course study, replicate measurements of the same sample, or simultaneous comparison of normal, diseased, or drug-treated samples.

PROTEOME BIOINFORMATICS

We have considered metabolic and chemical mechanisms of isotope labeling and protein purification routines. However, bioinformatic analysis incorporates an important dimension in quantitative proteomics experiment.

Peptide identification by database search. Protein identification in tandem mass spectrometry is made by virtue of database search [34]. In LC-MS/MS, gas phase peptide ions undergo collision-induced dissociation (CID) with molecules of an inert gas such as helium or argon. At low-energy CID, fragmentation mainly occurs along the peptide backbone bonds generating characteristic *b*-ions and *y*-ions and neutral losses of water and ammonia in MS/MS spectra [9]. The fact that fragmentation patterns are strongly dependent on the chemical and physical properties of the amino acids and sequences of the peptide allows comparison of MS/MS spectra to spectra generated from public proteomic and genomic sequence database [35]. In contrast to peptide sequencing by database search, *de novo* sequencing programs try to interpret the sequence of peptide from MS/MS spectra alone. We will not consider them in details here, but rather refer readers to an excellent review [36]. However, users should be warned that use of *de novo* sequencing is only justified when high quality MS/MS spectra are available and appropriate databases do not contain the sequences of the proteins of interest (as will be the case for the organisms with unsequenced genomes or when no sequenced homolog genome exists).

The goal of a tandem mass spectral database search is to identify the best sequence match to the spectrum [37]. For MS/MS with high signal-to-noise ratio and uniform fragmentation it is reasonably straightforward to identify the correct sequence match. In situations where a tandem mass spectrum is of poorer quality or when the peptide ion undergoes unusual fragmentation, MS/MS spectra analysis may benefit from the use of multiple search algorithms. A number of algorithms and scoring models have been developed to assess the likelihood of a match. Most of the algorithms have an identical step where proteins in the sequence database are artificially cleaved with proteases and an MS/MS spectrum is calculated for each peptide. Several basic approaches have been developed to model MS/MS spectra matches to sequences in the databases: descriptive, interpretative, and probability-based, which will be described in detail below. It should be noted, however, that algorithms will come up with peptide identification even for data acquired from non-peptide ions, peptides from incomplete proteolytic digestion or from poor-quality MS/MS with low signal levels and signal-to-noise ratios. To decrease the number of false-positive identifications, classification methods have been developed, which sort the good quality spectra from bad quality spectra. These methods rely on the relative normalized intensity of the peaks in MS/MS spectra and further assume that bad quality spectra are not amenable to correct peptide identification [38]. They should be used whenever possible to prevent large false-positive error rate in peptide identifications.

SEQUEST is an example of a program that uses a descriptive model for peptide fragmentation and correlation matching to a tandem mass spectrum [39]. It uses a two-tiered scoring scheme to assess the quality of the match between the spectrum and amino acid sequence from a database. The first score, preliminary score (*S*p), is an empirically derived value that restricts the number of sequences analyzed in the correlation analysis. *S*p sums the peak intensity of fragment ions matching the predicted sequence ions and accounts for the continuity of an ion series and the length of a peptide. The second score, XCorr, is a correlation score of the experimental to theoretical spectra. To calculate XCorr a theoretical spectrum is generated from the predicted fragment ion. In the theoretical spectrum the products of the main ion series are assigned an abundance index of 50, a window of one atomic mass unit around the main fragment ions is assigned an index of 25, and water and ammonia losses are assigned intensity of 10. Then theoretical and normal-

ized experimental spectra are cross-correlated to obtain similarities between the spectra (normalization takes into account the fact that larger peptides produce higher correlation scores). Consequently, users can identify the best-matched peptide by comparing XCorr scores for different peptide assignments.

Interpretative approaches in tandem mass spectral database search are based on manual or automated interpretation of a partial peptide sequence from a tandem mass spectrum and incorporation of that sequence into a database search. As with the above SEQUEST example, matches between the sequence and the spectrum are scored using probabilities or correlation methods. Peptide Search is the earliest and most widely used program in this category [40], which makes use of the fact that fragmentation spectra usually contain at least a small series of easily interpretable sequences. This series constitute an amino acid tag. The lowest mass in the series contains information about the mass unit distance to one of the peptide termini, the highest mass – about mass unit distance to the other peptide terminus. Thus, the spectra can be decomposed into three parts—the amino terminal mass, stretch of amino acid sequence, and carboxyl terminal mass. The whole construct can then be matched against sequences in the database, using additional information, if desired, such as identity of the proteolytic enzyme that was used.

In probability based methods no *a priori* determined probabilities are used. These methods generate a model that relates sequences to a spectrum and estimates a peptide identification score from the model. Thus, in the simplest models match frequencies of *b*- and *y*-ions are determined and used to calculate probability of correct peptide sequence identification by multiplying probabilities of individual fragment matches. Mascot is the most widely used database search program employing probabilistic approaches [41]. However, due to its commercial nature its algorithm has not been published.

Analysis of tandem mass spectral database search results. As with many other database search applications, the main challenge in MS/MS database search is not finding the best match in the database, but rather determining whether the best match was assigned correctly. Obviously, a human expert by observing a particular MS/MS spectrum can easily verify the correctness of peptide assignment. Unfortunately, in large-scale quantitative proteomics projects, it is becoming impossible to confirm correctness of each peptide assignment; therefore, high significance is placed on the use of appropriate scoring schemes for database searches [42]. In the simplest method, separation of correct from incorrect peptide assignments is achieved by applying *ad hoc* filtering criteria based upon database search scores and some properties of the assigned peptides. For example, using cut-off threshold values for SEQUEST such as "XCorr" score it is possible to short-list the best peptide assignments. With few exceptions [43], false identification error rates obtained from the application of filtering criteria are not estimated and reported, which makes comparison of results from different experiments or groups an extremely difficult task. Consistent and reliable interpretation of MS/MS data to enable the comparison of results from different experimental groups requires robust statistical methods to validate peptide assignments to MS/MS spectra. It should be noted that advantages of robust statistical approaches have been already realized in other high-throughput fields. For example, statistical base-calling models have been developed for the estimation of errors in raw DNA sequences obtained using large-scale DNA sequencing [44].

Several supervised classification methods for post-database search validation of MS/MS spectra to peptide assignments have recently been developed, with underlying statistical methods based on linear discriminant analysis [45] and support vector machines [46]. Here the supervised approach implies that a program has been trained on manually validated peptide sets to distinguish between correct and incorrect peptide assignments. The algorithm finds features that are most dissimilar in correct versus incorrect peptides and then uses them to build a classifier for use on new samples. We should mention that fully supervised classification algorithms might not produce accurate results when applied to datasets that are significantly different from those used for training and will depend on the quality of acquired MS/MS spectra, complexity of the analyzed samples, and differences in the experimental protocols used to generate distinct datasets. Thus, training datasets should be used to find features that discriminate between correct and incorrect peptide assignments from the data itself, but additional approaches should be used to derive numerical characteristics of each classifier [47].

One statistical model developed to overcome limitations of supervised learning is implemented in the software tool PeptideProphet [45]. It is based on the expectation maximization (EM) algorithm by which it derives a probabilistic mixture model of correct and incorrect peptide assignments from the data. It executes in two successive steps. First, it uses the observed information about each assigned peptide in the dataset and learns to distinguish correct from incorrect peptide assignments; second, it computes a probability for each assignment being correct. Therefore, it does not rely on a training set and in this way represents a fully unsupervised approach of machine learning. Features used to discriminate between incorrect and correct peptide assignments in PeptideProphet include database search scores, difference between measured and theoretical peptide mass, the number of termini consistent with the type of enzyme used, and the number of missed cleavage sites. If the database search tool outputs more than a single score useful for distinguishing correct from incorrect peptide assign-

ments, all such scores are combined into a single discriminant score in such a way that correct and incorrect peptide assignments are optimally discriminated. To conclude, the use of probability scores in PeptideProphet allows estimation of both the total number of correct identifications and the false-positive error rates, making this program very useful for large-scale quantitative proteomics experiments. Another type of software used for validation of MS/MS database search results deals with the issue of peptides to proteins assignment. A typical output of database search programs includes assignment of MS/MS spectra to a list of identified peptides. The ultimate goal of LC-MS/MS experiment is identification of proteins in the mixture from the list of identified peptides. However, assembling peptides into proteins is not straightforward. This challenge is analogous to shotgun fragment assembly of genomic sequences where overlapping DNA segments must be ordered to recreate the original sequence, which is complicated by the presence of repeats and gaps between individual fragments. Similarly, the presence of degenerate peptides whose sequence is present in more than one entry in the protein sequence database makes it difficult to determine the corresponding proteins present in the sample [46]. Such cases often result from the presence of homologous proteins, splicing variants, or redundant entries in the protein sequence databases, and are particularly abundant in large databases for higher eukaryotes. Some search programs, like MASCOT, can automatically group peptides according to their corresponding protein entries. However, if multiple datasets of MS/MS spectra are acquired and processed at different times, which is often the case in large-scale proteomics experiments, the peptide-protein assignment will not be possible.

ProteinProphet addresses this issue by providing probability that a protein is present in the sample through combination of probabilities that corresponding peptide assignments are correct [48]. Peptides corresponding to single-hit proteins are penalized, whereas those corresponding to multi-hit proteins are rewarded. The amount of adjustment depends on the sample complexity and the number of acquired MS/MS spectra, and it is learned from the data using the expectation maximization algorithm. The model handles degenerate peptides by sharing each such peptide among all its corresponding proteins to derive a minimal protein list sufficient to account for the identified peptides. Those proteins that are impossible to differentiate on the basis of identified peptides are grouped together. By these means, ProteinProphet produces accurate probabilities of the presence of a protein and can discriminate between correct and incorrect protein identifications including identifications based on a single peptide.

Calculation of relative protein abundance ratios. There are fewer software programs for relative quantification of proteins with stable isotope labeling than for MS/MS spectra interpretation. This is explained by the later advent of stable isotope labeling techniques and their more specialized nature compared to the ubiquitous applications of tandem mass spectrometry. Among open-source programs for protein quantification, the most widely used are ASAPRatio [49] and MSQuant [50] (Pro ICAT software is available from Applied Biosystems for the cleavable ICAT reagent). The RelEx program has also been developed for data generated by shotgun proteomics that uses N^{15} labeling [19]. In this review, we will consider the algorithm of ASAPRatio as an example of computational approaches used for protein abundance estimations. We should note that other quantitative proteomics programs use the same principle whereby relative ratio of peptide abundance is computed from ion current chromatogram (c.f. iTRAQ). In a nutshell, the programs approximate the area under MS spectra for a single peptide at different elution time points and compare the area of labeled and unlabeled peptides to compute peptide ratio. ASAPRatio does these computations in four steps.

Step 1. Evaluation of a peptide abundance ratio for each peptide identified by MS/MS and database searching.

Step 2. Evaluation of a "unique peptide ratio" for each identified peptide sequence.

Step 3. Evaluation of protein abundance ratio for each identified protein.

Step 4. Evaluation of the significance of abundance change for each identified protein.

In step 1, the Savitzky–Golay filtering method is used to obtain the smoothed chromatogram (for area estimation) and the average signal outside the elution peak is used for estimating the background level. In cases where the signal from one of the peptide pairs is missing, the ratio is assumed to be 1 : 0 or 0 : 1. Often, the same peptide is observed on MS/MS spectra in different charge states, and ASAPRatio accounts for this by determining locations of all possible charge states, and if signal is present, estimating their (weighted by area) averages. In the second step, ASAPRatio computes "unique peptide ratio", where it assigns single ratios for identical peptides eluted from the mass spectrometer in different chromatographic fractions. In the third and fourth steps, the protein abundance ratio is computed from unique peptide ratios and distribution of log-transformed ratios is fitted to normal distribution to calculate statistical significance for each protein identified. Thus, ASAPRatio calculates ratios at protein level and also provides *p*-values for estimation for statistical significance. The latter feature is especially relevant for large-scale quantitative proteomics project, since it obviates manual verification of insignificant ratios (PeptideProphet, ProteinProphet, and ASAPRatio discussed previously are parts of "Protein Identification Pipeline" developed at the Institute of Systems Biology (Seattle, USA), which offers integrative analysis of tandem mass spectrometry experiments and

can be downloaded from the institute's website http://www.proteomecenter.org/software.php).

In our final section on quantitative proteome bioinformatics we will briefly consider the issue of data standards for proteomics experiments (including quantitative). As exemplified by the MIAME (minimal information about microarray experiment) initiative [51], development of data standards is essential to facilitate exchange of data generated by different mass spectrometers. Furthermore, such standards would encourage data submission to online bioinformatics databases and would make information from tandem mass spectrometry experiments available to a wide community of researchers. Although several initiatives were already described [52, 53], there is still no unified framework available for representation of mass spectrometry data. Currently, the most successful endeavor in this area is the mzXML format for vendor-neutral representation of MS data [54]. In mzXML, XML schema and libraries of converters are used to represent MS data from specific tandem mass spectrometers in vendor-neutral style. The mzXML format can represent raw or processed data, offering researchers not directly working in the proteomics field a way to access all the information required for development of data manipulation or data mining algorithms (e.g., noise reduction, peak detection, charge state deconvolution). This creates new opportunities for statisticians and computer scientists. These features of the mzXML format are expected to drive progress and standardization in MS-based proteomics.

QUANTITATIVE PROTEOMICS APPLICATIONS IN SYSTEMS BIOLOGY

Methods of tandem mass spectrometry-based quantitative proteomics have been successfully used for analyzing large-scale changes in protein abundance [55], finding novel protein–protein [56] and protein–peptide interactions [50], studying dynamics of large macromolecular complex formation [57, 58], and elucidating dynamic changes in the protein composition of organelles [59] and phosphorylation in signal transduction cascades [60]. All these approaches pose great promise for systems biology and therefore merit discussion here.

Analyzing changes in protein abundance. Quantitative proteomics for studying changes in protein abundance is akin to DNA microarray technologies and has been used by many researchers employing different isotope labeling techniques and methods of protein separation. The methods of quantitative proteomics go one step further in analyzing regulation of gene expression and take into account variations produced at the post-transcriptional level, such as mRNA degradation and variability in translation initiation. Low correlation between DNA microarray and proteomics data has been reported [61], which can have various explanations. First, it should be noted that application of more "biologically accountable" statistical methods rather than the use of simple correlation coefficients can discover novel congruity between mRNA and protein levels [62]. Second, the lack of correlation arises from the fact that the synthesis of individual protein species is regulated, not only by transcript level, but by *cis* regulatory elements of mRNA molecules that generate individual translation patterns. Therefore, accounting for the differences between mRNA and protein levels has in itself important implications for large-scale analyses of cellular regulatory mechanisms. A newer study approached this problem by using polysome fractionation *prior to* transcript analysis [63]. They created an expression profile for each mRNA molecule as a function of its ribosome loading. Of 816 genes whose protein expression was altered by at least 2-fold, 24% showed 2-fold change in translational efficiency. Regularly, transcript array analysis would have ignored those genes that were regulated solely at the translational level and would have erred quantitatively with those transcripts that showed mixed regulation. Furthermore, examples of regulation at the level of protein and mRNA degradation were found, which would have been missed if DNA microarray and ICAT proteomics experiments were done separately. Translation of the transcriptome is highly diverse both qualitatively and quantitatively, and it is impossible to assume a simple, linear relationship between the level of an mRNA and the rate of synthesis of its encoded protein.

Protein interactions. Another very promising area for application of stable isotope labeling techniques is identification of protein interactions [64]. Tandem mass spectrometry has an intrinsic advantage over other large-scale approaches like the yeast two-hybrid system in that interactions are observed in the cell's native environment and that the large protein complexes can be scrutinized (rather than simple binary interactions predominant in the two-hybrid system). However, since many biological interactions are of low affinity and dependant on immediate protein environment, a typical MS/MS experiment can identify only their subset. This can be further compounded by high false-positive error rate as a result of nonspecific protein co-purification (in one comparative study of protein interactions identified by various high-throughput techniques only a small subset was found common for all techniques [65]). Several approaches can be used to decrease the number of false-positive interactions. First, protein covalent cross-linking is used to stabilize weak protein binding [66]. In another method, called protein correlation profiling, the normalized square deviation between ion current profiles of centrifugation fractions with previously characterized centrosomal proteins and novel ones was used to find new putative centrosomal proteins among a background of nonspecifically co-purified proteins [67]. This method estimated relative ion current ratios among different separation

fractions as an indication of possible protein localization. It can be used to characterize protein composition for organelles with well-established purification protocols.

Compared to the approaches described above, the quantitative proteomics methods resolve problems of nonspecific binding in a different way by employing the following three-step strategy. First, the protein, whose interaction partners are being searched for (bait) is encoded with stable isotopes and is isolated either by specific antibody or by prior tagging along with its interacting partners (prey). The discrimination between genuine and nonspecific binding is achieved at the second stage, where the same protein is isolated by nonspecific antibody or through a mutated/deleted version of the protein's gene that is incapable of forming protein complexes. The purification and quantification methods in the third step are similar in all quantitative proteomics approaches and lead to MS spectra where true interactions have higher protein abundance ratios than nonspecific ones (nonspecific binders have a 1 : 1 ratio for both isotopic forms). Such protocol has been successfully used to characterize formation of core polymerase II transcription complex [57] and phosphorylation-dependent complex formation upon stimulation with epidermal growth factor (EGF) [56]. In one study ICAT reagent was used to differentiate between components of polymerase II complex and nonspecific interactions in cells expressing temperature-sensitive mutated gene of TBP (TATA binding protein) component essential for complex formation. Proteins were affinity-purified with template containing the TATA box region plus upstream promoters, and most of the genuine polymerase II complex components were identified by analyzing relative peptide abundance ratios. The second study used SILAC reagent to identify proteins in a complex formed at the EGFR (epidermal growth factor receptor) intracellular domain upon its stimulation.

One biologically important type of protein interactions is regulated by post-translational modifications, and quantitative proteomics can provide new approaches for their identification. In the case of EGF-dependent phosphorylation SILAC was used to differentially label proteins in EGF-stimulated versus unstimulated cells. Combined cell lysates were then affinity-purified over the SH2 domain of the adapter protein Grb2 (GST-SH2 fusion protein) that specifically binds phosphorylated EGFR. Many signaling molecules were found to specifically form complexes with the activated EGFR, as well as plectin, epiplakin, cytokeratin networks, and histone H3 among other molecules [56].

Proteomics of organelles. Organellar proteomics is concerned with identification and characterization of protein composition in individual organelles [68]. Akin to identification of protein complexes, quantitative proteomics can be employed in organellar proteomics to study temporal changes in organellar protein composition. *Prior to* the advent of stable isotope labeling techniques most of the reported mass spectrometry studies in this field were descriptive, providing a list of proteins found in one or another organelle. It should be remembered that proteins perform their functions in cells by constantly circulating between different organelles and various compartments within organelles (as well as organelle-like structures such as centrosome and nucleolus). Therefore, elucidation of dynamic changes in organellar protein composition can provide a new dimension for systems biology modeling and can contribute to our understanding of changes in protein transport triggered by a cell in response to various signals. Application of quantitative proteomics has been proposed for investigating dynamic changes in protein composition in the nucleolus. The nucleolus is a key organelle that coordinates the synthesis and assembly of ribosomal subunits and forms in the nucleus around the repeated ribosomal gene clusters. Using mass spectrometry-based organellar proteomics and stable isotope labeling, a flux of 489 endogenous nucleolar proteins in response to three different metabolic inhibitors affecting nucleolar morphology was characterized [69]. The relative levels of all these 489 factors were quantified by SILAC in two large-scale experiments, measured at five or nine separate time points after inhibiting transcription with actinomycin D. The steady-state levels of many nucleolar proteins decreased to various extents, including ribosomal proteins, RNA processing factors, exosome components, and RNA polymerase I. Remarkably, the level of some proteins increased up to tenfold. This result suggests that the nucleolus is not a simple ribosome synthesis machine that progressively breaks down in the absence of transcription. Instead, transcription inhibition leads to a more subtle redistribution of nuclear proteins, and the partition of proteins between the nucleolus and nucleoplasm could therefore reflect the general physiological status of the cell.

In addition to SILAC, protein correlation profiling has recently been used to study protein composition and dynamic changes in protein abundances within and between organelles [70]. In mouse liver, subcellular locations of 1502 proteins have been mapped. Ten major clusters were found that corresponded to well-characterized cellular compartments. Protein correlation profiles validated genuine organellar components and enabled assessing the specificity of previously published organellar proteomic inventories. About 41% of all organellar proteins were found in more than one location. The authors [70] also integrated the proteomic data with atlases of RNA abundance and genome sequence to identify networks of co-expressed genes, *cis*-regulatory motifs, and putative transcriptional regulators involved in organelle biogenesis. This is another example of contributions that quantitative proteomics data can make for systems biology research.

Quantitative phosphoproteomics. Protein phosphorylation is one of the key mechanisms whereby signals at the cell membrane are transmitted into the cytoplasm to trigger immediate protein-mediated and transcriptional responses. Blagoev et al. described a new quantitative phosphoproteomics technique that allows comparison of three protein samples and demonstrated its potential with a study of the temporal dynamics of EGFR tyrosine kinase signaling in HeLa cells [60]. Three differentially labeled cell lysates encoding five time points were mixed, immunoprecipitated with anti-phosphotyrosine antibodies, and analyzed by LC-MS/MS. Proteins isolated at each time point were quantified as fold change over the basal, zero time point. Upon EGFR stimulation, 81 proteins showed a more than 1.5-fold change in their level. Many of these proteins were already known to be involved in signal transmission from activated EGFR, but some newly implicated proteins were also identified. The fact that most of the proteins known to be involved in EGF signaling were found bodes well for the usefulness of the method. The feature of this method is that it does not distinguish between isolated tyrosine-phosphorylated proteins and proteins associated with tyrosine-phosphorylated proteins. However, phosphorylation-dependent protein interactions may be just as important as the phosphorylations themselves. It should be remembered that a protein detected with this technique is not necessarily a direct substrate of any tyrosine kinase [71]. Nor does the technique identify sites of phosphorylation, although simultaneous use of bioinformatics and standard experimental techniques may allow quick identification. The protocol could also be modified to isolate only phosphorylated peptides: the samples could be digested before immunoprecipitation with anti-phosphotyrosine antibodies that will allow temporal profiling of individual phosphorylation sites. A similar protocol has been used to study temporal profiles of tyrosine phosphorylation in insulin-induced brown adipocytes [72].

Recently quantitative phosphoproteomics applications have been taken even further. The number of distinct intracellular signaling pathways is believed to be smaller than the number of signals cells receive at any given time and the differential responses it generates. Several models responsible for diversification of signaling pathways exist; a significant contribution to them results from combinatorial encoding of individual signal transduction pathways and pathway motifs [73]. This could be one of the major reasons why closely related signals often lead to very different cellular outcomes. Therefore, it is essential to have methods that would permit determination of which signaling pathways are activated in response to a given stimulus. Using a three-state SILAC labeling approach, Kratchmarova et al. were able to discriminate between signaling pathways operating in mesenchymal stem cells (hMSC) in response to stimulation by EGF and PDGF (platelet derived growth factor) [74]. Differentiation of hMSC into osteoblasts is stimulated by EGF but not PDGF. Using quantitative phosphoproteomics, more than 90% of the signaling proteins were found to be used by both ligands, whereas the phosphatidylinositol 3-kinase (PI3K) pathway was exclusively activated by PDGF, implicating it as a possible control point. Chemical inhibition of the phosphatidylinositol pathway with PI3K-specific inhibitor in PDGF-stimulated cells removed the differential effect of the two growth factors, suggesting that PI3K pathway is responsible for inhibition of osteoblast differentiation in PDGF-stimulated stem cells. It is therefore possible, with a combination of proteomics and chemical biology, to elucidate pathways that influence cell fate.

The greatest shortcoming of the approaches discussed above is that they consider only tyrosine phosphorylation. Similar analysis of Ser/Thr phosphorylation presents a considerably greater technical challenge, but it would be very useful if these techniques could apply to all phosphorylation. This is especially important since most transcription factors are regulated through Ser/Thr phosphorylation. Overcoming this challenge would allow integration of intracellular signaling events and transcriptional responses that they trigger (easily detectable by DNA microarrays), thus producing invaluable data for analysis of entire signaling cascades from effectors to downstream transcriptional responses.

Mass spectrometry based quantitative proteomics holds high potential as a technique of choice for a large number of systems biology applications. However, quantitative proteomics, as every other high-throughput technique, has its limitations. It is still difficult to identify and quantify all the low-abundance proteins, especially in the presence of highly abundant ones. Furthermore, as in DNA microarray technology, owing to the large amount of data generated, the results produced by quantitative proteomics are often "noisy" and it can be difficult to distil functional and mechanistic hypotheses from such global experiments. It is also difficult to perform proteome wide quantitative experiments due to difficulties incurred in handling complex protein samples. This can be one of the most serious limitations of quantitative proteomics to model building in systems biology at present, since the later requires replicated data, which is difficult to obtain for complex samples. However, the detection limits and dynamic range of mass spectrometers are rapidly improving by development of new hardware and software.

In this review we have provided various examples of how quantitative proteomics can be used to study protein interactions, temporal profiles of cellular signaling events, and dynamic changes in the protein composition of organelles. Such data sets can be an invaluable asset for model building and model verification in systems biology. Further insights into the complex biological network are

achieved by integrating proteomics studies with data on gene expression from DNA microarray experiments, high-throughput mutagenesis studies, and bioinformatics applications [75].

Knowledge about key components of the system can be used for development of highly specific inhibitors for drug targets with minimal side effects [76, 77]. Early proteomic diagnostic of human diseases on the system level can become the basis for personalized medicine of the future [77].

The authors are grateful to Dr. Douglas Armstrong for comments on drafts of this paper and useful discussions.

REFERENCES

1. Kitano, H. (2002) *Science*, **295**, 1662-1664.
2. Aggarwal, K., and Lee, K. H. (2003) *Brief Funct. Genomic Proteomic*, **2**, 175-184.
3. Butcher, E. C., Berg, E. L., and Kunkel, E. J. (2004) *Nat. Biotechnol.*, **22**, 1253-1259.
4. Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dia, H., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C., and Friend, S. H. (2000) *Science*, **287**, 873-880.
5. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Lo, H., Downing, J. R., and Caligiuri, M. A. (1999) *Science*, **286**, 531-537.
6. DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997) *Science*, **278**, 680-686.
7. Mann, M. (1999) *Nature Biotechnol.*, **17**, 954-956.
8. Aebersold, R., and Mann, M. (2003) *Nature*, **422**, 198-207.
9. Steen, H., and Mann, M. (2004) *Nat. Rev. Mol. Cell. Biol.*, **5**, 699-711.
10. Mann, M., Hendrickson, R. C., and Pandey, A. (2001) *Annu. Rev. Biochem.*, **70**, 437-473.
11. De Hoog, C. L., and Mann, M. (2004) *Annu. Rev. Genomics Hum. Genet.*, **5**, 267-293.
12. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000) *Proc. Natl. Acad. Sci. USA*, **97**, 9390-9395.
13. Ivakhno, S., and Kornelyuk, A. (2005) *Mol. Cell. Proteomics*, **4**, (Suppl. 1), HUPO 4th Ann. World Congr., S23.
14. Sechi, S., and Oda, Y. (2003) *Curr. Opin. Chem. Biol.*, **7**, 70-77.
15. Li, X. J., Yi, E. C., Kemp, C. J., Zhang, H., and Aebersold, R. (2005) *Mol. Cell. Proteomics*, **9**, 1328-1340.
16. Kuster, B., Schirle, M., Mallick, P., and Aebersold, R. (2005) *Nat. Rev. Mol. Cell. Biol.*, **7**, 577-583.
17. Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 6591-6596.
18. Washburn, M. P., Ulaszek, R., Deciu, C., Schieltz, D. M., and Yates, J. R., 3rd. (2002) *Analyt. Chem.*, **74**, 1650-1657.
19. MacCoss, M. J., Wu, C. C., Liu, H., Sadygov, R., and Yates, J. R., 3rd. (2003) *Analyt. Chem.*, **75**, 6912-6921.
20. Krijgsveld, J., Ketting, R. F., Mahmoudi, T., Johansen, J., and Artal-Sanz, M. (2003) *Nat. Biotechnol.*, **21**, 927-931.
21. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) *Mol. Cell. Proteomics*, **1**, 376-386.
22. Ong, S. E., Kratchmarova, I., and Mann, M. (2003) *J. Proteome Res.*, **2**, 173-181.
23. Ong, S. E., Foster, L. J., and Mann, M. (2003) *Methods*, **29**, 124-130.
24. Ishihama, Y., Sato, T., Tabata, T., Miyamoto, N., Sagane, K., Nagasu, T., and Oda, Y. (2005) *Nat. Biotechnol.*, **23**, 617-621.
25. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) *Nat. Biotechnol.*, **17**, 994-999.
26. Zhou, H., Ranish, J. A., Watts, J. D., and Aebersold, R. (2002) *Nat. Biotechnol.*, **20**, 512-515.
27. Kuyama, H., Watanabe, M., and Toda, C. (2003) *Rapid Commun. Mass Spectrom.*, **17**, 1642-1650.
28. Ibarrola, N., Molina, H., Iwahori, A., and Pandey, A. (2004) *J. Biol. Chem.*, **279**, 15805-15813.
29. Cagney, G., and Emili, A. (2002) *Nat. Biotechnol.*, **20**, 163-170.
30. Zhang, H., Yan, W., and Aebersold, R. (2004) *Curr. Opin. Chem. Biol.*, **8**, 66-75.
31. Zhang, X., Jin, Q. K., Carr, S. A., and Annan, R. S. (2002) *Rapid Commun. Mass Spectrom.*, **16**, 2325-2332.
32. Mirgorodskaya, O. A., Kozmin, Y. P., Titov, M. I., Korner, R., Sonksen, C. P., and Roepstorff, P. (2000) *Rapid Commun. Mass Spectrom.*, **14**, 1226-1232.
33. Schutz, F., Kapp, E. A., Simpson, R. J., and Speed, T. P. (2003) *Biochem. Soc. Trans.*, **31**, 1479-1483.
34. Sadygov, R. G., Cociorva, D., and Yates, J. R., 3rd. (2004) *Nat. Meth.*, **1**, 195-202.
35. Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) *Mol. Cell. Proteomics*, **3**, 1154.
36. Standing, K. G. (2003) *Curr. Opin. Struct. Biol.*, **13**, 595-601.
37. Yates, J. R. (1998) *Electrophoresis*, **19**, 893-900.
38. Bern, M., Goldberg, D., McDonald, W. H., and Yates, J. R., III. (2004) *Bioinformatics*, **20**, I49-I54.
39. Eng, J. K., McCormack, A. L., and Yates, J. R., III. (1994) *J. Am. Soc. Mass Spectrom.*, **5**, 976-989.
40. Mann, M., and Wilm, M. (1994) *Analyt. Chem.*, **66**, 4390-4399.
41. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) *Electrophoresis*, **20**, 3551-3567.
42. Hart, S. R., and Riba-Garcia, I. (2004) *Drug Discov. Today*, **9**, 391-392.
43. Qian, W. J., Liu, T., Monroe, M. E., Strittmatter, E. F., Jacobs, J. M., Kangas, L. J., Petritis, K., Camp, D. G., 2nd, and Smith, R. D. (2005) *J. Proteome Res.*, **4**, 53-62.
44. Ewing, B., and Green, P. (1998) *Genome Res.*, **8**, 186-194.
45. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) *Analyt. Chem.*, **74**, 5383-5392.
46. Rappsilber, J., and Mann, M. (2002) *Trends Biochem. Sci.*, **27**, 74-78.
47. Nesvizhskii, A. I., and Aebersold, R. (2004) *Drug Discov. Today*, **9**, 173-181.
48. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) *Analyt. Chem.*, **75**, 4646-4658.
49. Li, X. J., Zhang, H., Ranish, J. A., and Aebersold, R. (2003) *Analyt. Chem.*, **75**, 6648-6657.

50. Schulze, W. X., and Mann, M. (2004) *J. Biol. Chem.*, **279**, 10756-10764.
51. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001) *Nat. Genet.*, **29**, 365-371.
52. Taylor, C. F., Paton, N. W., Garwood, K. L., Kirby, P. D., Stead, D. A., Yin, Z., Deutsch, E. W., Selway, L., Walker, J., Riba-Garcia, I., Mohammed, S., Deery, M. J., Howard, J. A., Dunkley, T., Aebersold, R., Kell, D. B., Lilley, K. S., Roepstorff, P., Yates, J. R., 3rd, Brass, A., Brown, A. J., Cash, P., Gaskel, S. J., Hubbard, S. J., and Oliver, S. G. (2003) *Nat. Biotechnol.*, **21**, 247-254.
53. Orchard, S., Hermjakob, H., Julian, R. K., Jr., Runte, K., Sherman, D., Wojcik, J., Zhu, W., and Apweiler, R. (2004) *Proteomics*, **4**, 490-491.
54. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) *Nat. Biotechnol.*, **22**, 1459-1466.
55. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001) *Nat. Biotechnol.*, **19**, 946-951.
56. Blagoev, B., Kratchmarova, I., Ong, S. E., Nielsen, M., Foster, L. J., and Mann, M. (2003) *Nat. Biotechnol.*, **21**, 315-318.
57. Ranish, J. A., Yi, E. C., Leslie, D. M., Purvine, S. O., Goodlett, D. R., Eng, J., and Aebersold, R. (2003) *Nat. Genet.*, **33**, 349-355.
58. Ranish, J. A., Hahn, S., Lu, Y., Yi, E. C., Li, X. J., Eng, J., and Aebersold, R. (2004) *Nat. Genet.*, **36**, 707-713.
59. Andersen, J. S., Lam, Y. W., Leung, A. K., Ong, S. E., Lyon, C. E., Lamond, A. I., and Mann, M. (2005) *Nature*, **433**, 77-83.
60. Blagoev, B., Ong, S. E., Kratchmarova, I., and Mann, M. (2004) *Nat. Biotechnol.*, **22**, 1139-1145.
61. Washburn, M. P., Koller, A., Oshiro, G., Ulaszek, R. R., Plouffe, D., Deciu, C., Winzeler, E., and Yates, J. R. (2003) *Proc. Natl. Acad. Sci. USA*, **100**, 3107-3112.
62. Mootha, V. K., Bunkenborg, J., Olsen, J. V., Hjerrild, M., Wisniewski, J. R., Stahl, E., Bolouri, M. S., Ray, H. N., Sihag, S., Kamal, M., Patterson, N., Lander, E. S., and Mann, M. (2003) *Cell*, **115**, 629-640.
63. MacKay, V. L., Li, X., Flory, M. R., Turcott, E., Law, G. L., Serikawa, K. A., Xu, X. L., Lee, H., Goodlett, D. R., Aebersold, R., Zhao, L. P., and Morris, D. R. (2004) *Mol. Cell. Proteomics*, **3**, 478-489.
64. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002) *Nature*, **415**, 180-183.
65. Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002) *Nature*, **417**, 399-403.
66. Rappsilber, J., Siniossoglou, S., Hurt, E. C., and Mann, M. (2000) *Analyt. Chem.*, **72**, 267-275.
67. Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Erich, A., Nigg, E. A., and Mann, M. (2003) *Nature*, **426**, 570-574.
68. Taylor, S. W., Fahy, E., and Ghosh, S. S. (2003) *Trends Biotechnol.*, **21**, 82-88.
69. Johnson, S. A., and Hunter, T. (2004) *Nat. Biotechnol.*, **22**, 1093-1094.
70. Foster, L., De Hoog, C., Xie, X., Mootha, V., and Mann, M. A. (2005) *Mol. Cell. Proteomics*, **4** (Suppl. 1), HUPO 4th Ann. World Congr., S6.
71. Johnson, S. A., and Hunter, T. (2004) *Nat. Biotechnol.*, **22**, 1093-1094.
72. Krueger, M., Kratchmarova, I., Blagoev, B., Kahn, R., and Mann, M. (2005) *Mol. Cell. Proteomics*, **4** (Suppl. 1), HUPO 4th Ann. World Congr., S38.
73. Pasquale, E. B. (2005) *Nat. Rev. Mol. Cell. Biol.*, **6**, 462-475.
74. Kratchmarova, I., Blagoev, B., Haack-Sorensen, M., Kassem, M., and Mann, M. (2005) *Science*, **308**, 1472-1477.
75. Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001) *Science*, **292**, 929-934.
76. Souchelnytskyi, S. (2005) *Proteomics*, **16**, 4123-4137.
77. Govorun, V. M., and Archakov, A. I. (2002) *Biochemistry (Moscow)*, **67**, 1109-1123.