Russian Chemical Reviews **83** (2) 143 – 154 (2014)

DOI 10.1070/RC2014v083n02ABEH004377

Prediction of protein post-translational modifications: main trends and methods

B N Sobolev, A V Veselovsky, V V Poroikov

V N Orekhovich Institute of Biomedical Chemistry, Russian Academy of Medical Science
ul. Pogodinskaya 10, 119121 Moscow, Russian Federation

The review summarizes main trends in the development of methods for the prediction of protein post-translational modifications (PTMs) by considering the three most common types of PTMs — phosphorylation, acetylation and glycosylation. Considerable attention is given to general characteristics of regulatory interactions associated with PTMs. Different approaches to the prediction of PTMs are analyzed. Most of the methods are based only on the analysis of the neighbouring environment of modification sites. The related software is characterized by relatively low accuracy of PTM predictions, which may be due both to the incompleteness of training data and the features of PTM regulation. Advantages and limitations of the phylogenetic approach are considered. The prediction of PTMs using data on regulatory interactions, including the modular organization of interacting proteins, is a promising field, provided that a more carefully selected training data will be used.
The bibliography includes 145 references.

Contents

I. Introduction

As a result of numerous genome sequencing projects, a huge number of protein-coding regions and the related amino acid sequences were identified.[1] These data are widely used in studies of molecular mechanisms providing the biological functions. The human genome contains from 20 to 25 thousands of protein-coding DNA sequences.[2] According to estimates,[3] the number of transcripts read from these regions should exceed 100 thousands due to such factors as alternative splicing, more than one translation start codons, alternative polyadenylation and RNA editing. An even greater protein diversity is due to chemical modifications performed by specialized enzymes after the synthesis of a polypeptide chain on a ribosome. In genomes of higher eukaryotes, genes encoding enzymes responsible for these post-translational modifications (PTMs) account for approximately 5% of their total amount.[4] More than 300 different types of PTMs are distinguished,[5] due to which the number of protein variants may exceed a million.[6, 7]

There are the following types of PTMs.[4]

1. Limited or controlled proteolysis, which provides the formation of the functionally active variant of proteins and determines their subcellular localization. The polypeptide chain cleavage is performed by external proteases or through autocatalysis.

2. The covalent attachment of chemical groups to the terminal amino or carboxyl group of the polypeptide chain.

B N Sobolev Candidate of Biological Sciences, leading researcher of the IBMC RAMS. Tel. (7-499) 246 09 20, e-mail: boris.sobolev@ibmc.msk.ru
Current research interests: systematic biology, bioinformatics, protein post-translational modifications, functional protein annotation, protein functional mapping, amino acid sequence analysis, identification of functionally important sites in protein molecules.
A V Veselovsky Doctor of Biological Sciences, head of the Laboratory at the IBMC RAMS. Tel. (7-499) 245 07 68, e-mail: veselov@ibmh.msk.su
Current research interests: structural, functional and evolutionary genomics, systematic biology, bioinformatics, three-dimensional protein modelling, protein post-translational modifications.
V V Poroikov Doctor of Biological Sciences, Professor, head of the Department at the IBMC RAMS. Tel. (7-499) 246 09 20, e-mail: vladimir.poroikov@ibmc.msk.ru
Current research interests: molecular modelling, pattern recognition, bioinformatics, comparative genomics, pharmacology, pharmaceutical and medicinal chemistry, computational chemistry, systematic biology, chemoinformatics, computational drug design, post-genomic technologies, protein structure and functions, signalling regulatory networks analysis, toxicology.

Received 11 February 2013
Uspekhi Khimii **83** (2) 143 – 154 (2014); translated by T N Safonova

3. The attachment of chemical groups to amino acid side chains. This type of modifications were founds for 15 canonical amino acid residues.[4,8]

Residue	Type of modification
Arg	methylation
Asn	N-linked glycosylation
Asp	phosphorylation
Cys	phosphorylation, hydroxylation, formation of disulfide bridges
Gln	protein cross-linking by glutaminase
Glu	methylation, carboxylation
Gly	hydroxylation followed by C-terminal amide formation
His	phosphorylation
Lys	methylation, acetylation, ubiquitination
Met	oxidation to sulfoxide
Pro	hydroxylation
Ser	phosphorylation, O-linked glycosylation
Thr	phosphorylation, O-linked glycosylation
Trp	C-linked glycosylation (mannosylation)
Tyr	phosphorylation

According to calculations routinely performed at the PTM Statistics Curator web server[9] such modifications as phosphorylation, acetylation and glycosylation are most abundant among the experimental data presented in the UniProt[1] knowledgebase (Fig. 1). These modifications result in the attachment of the phosphate group, the acetyl group or a carbohydrate moiety, respectively, to an amino acid side chain (Fig. 2).

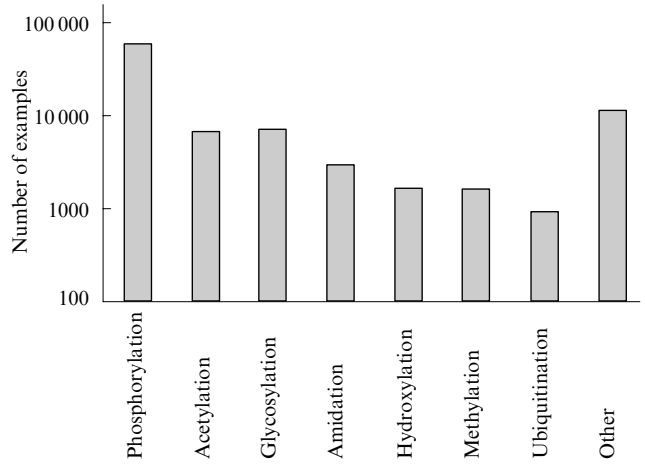

Figure 1. Representation of different PTM types in the UniProt knowledgebase (January, 2013).

The attachment of chemical groups leads to structural rearrangements in the protein molecule accompanied by changes in its functional characteristics, such as enzymatic activity, subcellular localization, *etc*. This affects the protein folding, protein–protein interactions, protein degradation, signal transduction from the receptor to the gene, regulation of the cell cycle, apoptosis, intercellular adhesion and so on.[10–15]

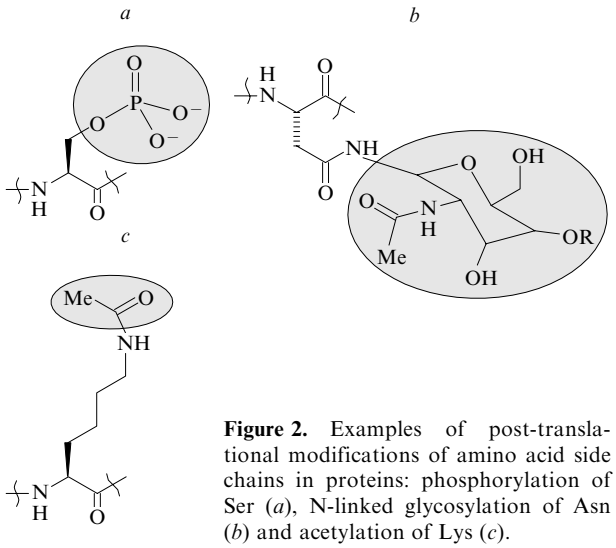

Figure 2. Examples of post-translational modifications of amino acid side chains in proteins: phosphorylation of Ser (*a*), N-linked glycosylation of Asn (*b*) and acetylation of Lys (*c*).

Data on protein post-translational modifications are necessary for the detailed analysis of key molecular biological processes. About 5% of mutations associated with pathologies were revealed in known PTM sites, whereas neutral mutations account for only 2%.[16] Disorders of PTMs were found in many diseases including cancer.[17–20]

A complex of experimental methods is used to identify modified proteins and localize PTM sites.[21] However, even the application of high-throughput mass-spectrometry methods does not allow researchers to determine the totality of PTMs, which can occur in living systems. In addition, the results of such investigations should be interpreted with caution. Data obtained *in vitro* depend to a large extent on the experimental conditions, therefore not all identified modifications can occur *in vivo*. In cells, reversible modifications of proteins are often short-term events, which take place in the course of dynamic processes and are followed by the recovery of the initial forms. Hence, when a material obtained *in vivo* is used, the sample can contain only a small part of potentially possible modifications.

At the present time, the prediction of PTMs based on amino acid sequences is one of the rapidly developing fields of bioinformatics. Several reviews on this subject were published in recent years.[5,22–28] These reviews cover methods based on the analysis of the neighbouring environment of modification sites in amino acid sequences. They deal with problems associated with the generation of initial data and the application of various approaches for the recognition of modification sites. Factors responsible for the accuracy of the prediction, such as the size of the training set and the classification of data according to modifying enzymes, are analyzed.

The available prediction methods give a large number of false-positive results.[22] Besides, in many cases it is impossible to clearly distinguish between true- and false-positive results because of the incompleteness of available data. Due to the stochastic character of PTM processes and the reversible character of modifications of many types, predicted PTMs can be absent in particular experimentally studied samples, which does not exclude their possible appearance in other situations.

In the present review, we focus on the multicomponent character of PTM-related processes, consider the molecular

mechanisms of protein modifications and discuss the possibility of increasing the efficiency of the prediction with the use of particular methods utilizing data on phylogenetic relationships and regulatory interactions. The following aspects will be highlighted:

— regulatory interactions in post-translational modifications;

— structural features of the neighbouring environment of modification sites;

— information resources that can be used for the generation of training sets for the prediction of PTMs;

— computational methods for the determination of modifications of different types in the proteins under study.

These aspects will be discussed for the three most common PTMs: phosphorylation, acetylation and glycosylation.

II. Interactions responsible for the specificity of post-translational modifications

Post-translational modifications occur as a result of interactions between two or more proteins. For these interactions, it is common that the structural domain of one protein molecule binds to a short peptide sequence of another protein.[29–33] All sequences specific for a particular domain can be described using a pattern or a linear motif. In this pattern, each position can be occupied by one of several residues having particular physicochemical properties. Short modules of protein molecules are referred to in the literature as SLiM (Short Linear Motifs),[34] minimotifs [35] or ELM (Eucaryotic Linear Motifs).[36, 37]

The linear motif is determined by a small number of amino acid positions so that the corresponding regions can be identified in amino acid sequences with a rather high probability. It should be noted that only some of the identified regions are functioning modules of this type.[38]

The region containing a site of modification (modified residue) and a small number of residues surrounding this site is also considered as a short module, which can interact with the catalytic domain of the modifying enzyme. Due to a small length of the motifs describing only the neighbouring environment of modification sites and their highly degenerate character, the presence of appropriate regions is not a conditions sufficient for the modification to occur. In addition, not all possible modification sites undergo changes in each particular case. On the other hand, not all experimentally identified sites are important for particular functionally significant events.

The specificity, as well as the possibility of performing PTMs, is determined not only by the structure of enzyme active site and neighboring environment of the modification site. The interaction and proper orientation of the protein with respect to the modifying enzyme can be provided by domain structures of the enzymes and short regions of substrate proteins, the so-called docking sites, located separately from PTM sites (Fig. 3).

In the case of mitogen-activated protein kinases, docking sites of the substrate bind to pockets in the catalytic domain located outside the active site (see Fig. 3 *a*).[39–42] Tyrosine-specific protein kinases of the Src group contain separate domains (SH2 and SH3) recognizing docking sites in substrate proteins (see Fig. 3 *b*).[43] Adaptor and scaffold proteins can play an important role in the determination of the phosphorylation specificity.[44] An adaptor protein binds

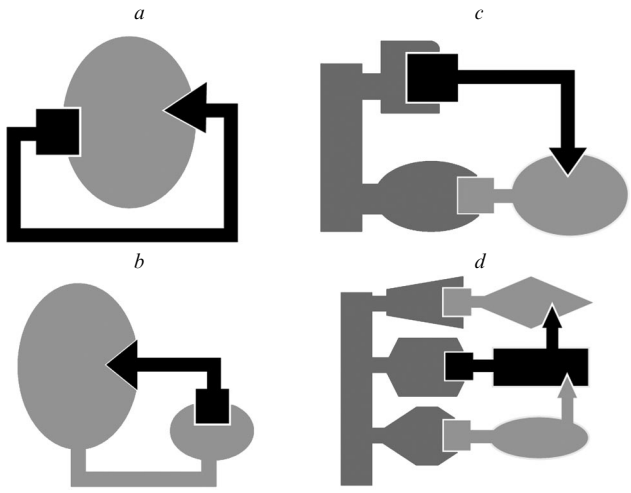

Figure 3. Interactions of proteins involved in the regulatory signal transduction through phosphorylation.
(*a*) Substrate recognition by single-domain protein kinase; (*b*) substrate recognition by the non-enzymatic domain of protein kinase; (*c*) recognition of enzymes and substrates by an adaptor protein; (*d*) the cascade of post-translational modifications coordinated by a scaffold protein. Modification sites of the substrate bound in the active site of the enzyme are indicated by triangles. Docking sites of the substrate and enzyme interacting with recognition sites are indicated by squares.

the enzyme and substrate, thus determining their specific interaction (see Fig. 3 *c*). Multidomain scaffold proteins mediate space-time organization of cell regulation pathways,[45] including the physical contact between the proteins involved in phosphorylation cascades (see Fig. 3 *d*). The specificity of modifying enzymes can also be affected by their interactions with other regulatory proteins. Let us mention cyclins as a classical example. These proteins form complexes with protein kinases of respective types, thus activating these enzymes and determining the set of substrate proteins.[46]

Therefore, the specificity of PTMs is determined not only by the neighbouring environment of the modification site in the amino acid sequence but also by numerous other factors. An analysis of methods for the prediction of PTMs (these methods are considered in detail below) shows that the more reasonable approaches are those which take into account these additional factors rather than only modification motifs.

III. Structural features of the neighbouring environment of modification sites

Regions containing modification sites, like all types of short modules, are usually identified in parts of protein molecules with less ordered spatial structure. The results obtained by NMR spectroscopy for non-modified proteins were reported in a study [47] where the authors analyzed the non-modified regions, which corresponded to the previously identified PTM sites. It was shown that the sites specific for particular types of PTMs are characterized by a higher degree of conformational flexibility. Thus, potentially phosphorylated regions are distinguished by a substantially

higher flexibility compared to non-modified regions. It should be noted that Lys acetylation sites differ only slightly in this parameter from Lys-containing regions for which the acetylation was not observed. Apparently, the calculated structurization/disorder characteristics should be taken into account in different ways for different types of modifications.

In the case of N-linked glycosylation, which occurs already during the translation of the polypeptide chain, the structure of modified sites is formed during the attachment of oligosaccharide chains. The regions containing glycosylation sites differ in composition from other parts of amino acid sequences.[48] For positions located at close distances from the modification site, the following preferences were mentioned: the sites -2 and $+4$ are characterized by a higher frequency of nonpolar residues, and the sites $+3$ and $+5$ are characterized by a higher frequency of hydrophobic residues with a bulky side chain.

IV. Information resources

Vast data sets stored in information resources (available *via* the Internet) contain various data on different types of PTMs. The number of such resources is rather large. Despite the diversity of the used formats, a typical record in the database contains the following fields: 1) the protein identifier; 2) the sequence position number of the modified residue; 3) the type of the amino acid residue; 4) the type of modification; 5) the attached group; 6) the modifying enzyme; 7) citation information (bibliographical references); 8) characteristics of the experiment.

Some fields are often missed, including the type of the modifying enzyme. Incomplete data pose serious problems in the generation of training sets, when it is necessary to relate a set of modification sites to the class of the enzyme.

The databases on PTMs have been summarized in sufficient detail.[5] Here, we briefly consider the most well-known resources.

The UniProt knowledgebase[1] is the main international resource of data on amino acid sequences and functional characteristics of the corresponding proteins. The data on the functional mapping (including data on the predicted and experimentally identified PTM sites) in the UniProt are collected from other information resources and bibliographical sources.

A large number of resources are specially designed for data on PTMs. For instance, dbPTM[49] is an information repository of PTMs retrieved from several databases, including UniProt,[1] Phospho.Elm,[50] UbiProt,[51] O-GLYCBASE,[52] HPRD[53] and PhosphoSitePlus,[54] as well as data extracted directly from publications. Such databases as Human Protein Reference Database (HPRD)[53] and Phospho.Elm,[50] are manually curated based on expert knowledge of appropriate publications.

Information sets available in different databases overlap with each other to a substantial extent. Some databases (for example, PhosphoSitePlus[54] and SysPTMs[55]) integrate data on regulatory interactions from the STRING,[56] KEGG,[57] PFAM[58] and Gene Ontology[59] information resources, thus providing better insight into the relationship between particular PTMs and biological processes.

The Phospho.Elm resource[50] is one of the largest freely available collection of information on phosphorylation. In addition to phosphorylation sites, this resource includes data on sequences that are recognized by phosphopeptide-binding domains. The Phospho3D database[60] includes projections of the identified phosphorylation sites on the three-dimensional structures of the corresponding proteins. Since most of the known three-dimensional structures refer to non-phosphorylated proteins, the database contains mainly information on the structures of potentially modified rather than actually modified sites.

Information on the acetylation of Lys residues is collected in the Compendium of Protein Lysine Acetylation (CPLA) resource.[61] This resource also integrates data on interactions between the modified proteins and other proteins.

Resources such as O-GLYCBASE[52] and UniPep,[62] are databases of glycosylation sites in amino acid sequences. The GlycoSuiteDB web server[63] provides information on the chemical structures of polysaccharide groups (glycans) involved in glycoproteins, as well as on the binding sites of these groups in amino acid sequences. The GlycomeDB metadatabase[64] integrates structural data on glycans from main databases of carbohydrate structures. The dbOGAP database[65] is the first specialized resource containing information on the O-linked glycosylation of proteins with attached *N*-acetylglycosamine.

V. Methods for the prediction of PTMs

1. General aspects

When developing new methods for the prediction of PTMs, it is necessary to evaluate the efficiency of original approaches compared to the already available methods. It is convenient to estimate the accuracy of prediction based on calculations of such parameters, as sensitivity and specificity. The sensitivity is determined as the percentage of the correctly predicted positive examples out of the total number of positive examples. The specificity is determined as the percentage of the correctly predicted negative examples of the total number of negative examples. Developers of new software strive to provide a high level of sensitivity at a rather high level of specificity.

An important problem faced by researchers when generating a training set that is used for the design of an algorithm for the prediction, as well as when estimating the prediction accuracy, is that it is impossible to determine unambiguously the positive and negative examples. Recall that the reactions resulting in PTMs have a stochastic character. In many studies, modifications are determined under different conditions (for example, in the selection of biological samples in different stages of the cell cycle). Hence, available experimental data not necessarily contain information on all PTMs occurring in living cells. This leaves place for the alternative interpretation of negative patterns: either these patterns actually attest to the absence of PTMs in particular regions of proteins or researchers failed to find experimental conditions under which these modifications occur. On the other hand, modifications identified *in vitro* not necessarily occur in living cells, *i.e.*, positive examples can be in fact negative.

In many cases, a comparison of the published data on a particular type of PTMs in one and the same protein shows that data on all known modification sites are available in none of the publications. This is due to different aims of particular studies, as well as because of the use of different

methods. The collection of experimental data bit by bit is a necessary step in the information work with PTMs.

The extensive list of references on programmes for the prediction of PTMs is available in the Cuckoo Group website (http://www.biocuckoo.org).

The largest number of methods was developed for the prediction of phosphorylation at Ser, Thr and Tyr residues. This is due to the abundance of available experimental data and the better understanding of the reversible phosphorylation, particularly in relation to the signal transduction and cell cycle regulation. Hence, the main approaches to the prediction of PTMs will be considered in more detail for this type of modifications.

2. Prediction of phosphorylation

The phosphorylation of amino acid side chains is a common way to regulate protein activity utilized in cell signal transduction.[66–69] It is believed that up to one-half of eukaryotic proteins undergo phosphorylation.[70] Data on this type of modifications are actively used in the modeling of cell signaling networks.[71] The phosphorylation at Ser, Thr and Tyr residues is most well studied.[72] Striking examples of the role of PTMs in regulated cellular signal transduction are hierarchical phosphorylation cascades,[73] in which one and the same protein kinase acts both as a modifying enzyme and a substrate.

In all methods, training sets include amino acid sequence regions, each containing several residues on both sides of the modification site. Various approaches are used to derive prognostic markers from these training sets.

In the GPS (Group-Based Prediction System) method,[74, 75] the estimate for the probable phosphorylation region is calculated based on its similarity with the already identified phosphorylation regions specific for a particular group of protein kinases. The estimates of similarity are calculated based on substitution matrices optimized for each group of protein kinases.

Generalized patterns (or motifs) of modified regions are often used for the prediction of phosphorylation.

The most pictorial representation of PTM motifs is based on the use of regular expressions. The latter are proposed for the first time in the PROSITE database.[76] Each position is described by a set of residues, which can present in this position. Despite the apparent simplicity of this representation of expressions, the procedure for the identification of motifs is rather complicated and is different in different approaches.

The Motif-x algorithm[77] takes a single-character pattern as an input, in which one position is occupied by only one symbol designating one of the twenty amino acid residues; the 21th additional symbol is used when the position can be occupied by any residue. The iterative procedure involves the successive selection of statistically significant motifs. The F-Motif method[78] employs the same type of pattern with the difference that the initial data set is divided into clusters, and motifs are determined for each particular cluster. This allows the identification of specific patterns that are not recognized by other similar algorithms. The MoDL programme[79] initially generates a set of single-character patterns based on phosphopeptides. In subsequent iterations, the motifs are combined thus creating classical regular expressions, in which the conserved position can be occupied by one of several letters. The coverage of one set by multiple motifs makes it possible to take into account, in implicit form, the relationships between individual positions. The match between the region under consideration and the motif can be estimated taking into account the scores of amino acid residues obtained for individual positions of the pattern.[80]

The position-specific scoring matrix (PSSM), where the scores for each of 20 amino acid residues are assigned to each position of the pattern, is employed in the Scansite programme.[81] Position-specific scoring matrices are used also in the Predikin programme.[82] The Predikin web server allows predicting phosphorylation sites also for new types of protein kinases. Users input amino acid sequences of protein kinases and expected substrates. The type of the modifying enzyme is determined based on the match with substrate-binding regions of known protein kinases.

Machine-learning methods have found wide application. The NetPhos[83] and NetPhosK programmes[23] employ artificial neural networks (ANNs). This approach allows not only weighting the positional scores but also taking into account the relationships between the positions. Such ANNs are trained based on phosphorylated regions. The NetPhos programme enables prediction of phosphorylation sites without taking into account the enzyme specificity. The more recent version of this method, NetPhosK, was developed for the prediction of sites specific for particular types of protein kinases. Artificial neural networks are employed also in the AMS 4.0 programme.[84] Different versions of support vector machines are used in the PredPhospho[85] and KinasePhos programmes,[86] as well as in the programme available *via* the PHOSIDA web server.[87] The PPSP programme[88] allows predicting the phosphorylation sites specific for different types of protein kinases based on Bayesian decision theory.

The DISPHOS (DISorder-enhanced PHOSphorylation predictor) method[89] implements the algorithm based on logistic regression. Input data include frequencies of amino acid residues, predicted disorder and secondary structure, estimates of conformational flexibility, hydrophobicity and other physicochemical parameters.

The pkaPS method was developed for the prediction of sites specific for protein kinase A (PKA).[90] When developing this method, regions containing modification sites were studied in detail. Neuberger *et al.*[90] also employed data on the three-dimensional structures of PKA in complex with a peptide, resulting in the development of a simplified model of interactions. This model enabled the prediction of phosphorylation sites with a rather high accuracy.

Let us mention the Phos3D web server,[91] which takes as input three-dimensional structures of proteins (in the PDB format). Support vector machines are used for the prediction of modification sites. Durek *et al.*[91] showed that the accuracy of the prediction steadily increases, although slightly, compared to other methods utilizing information only on amino acid sequences.

A comparative trial of several above-mentioned methods showed that all these programmes predict phosphorylation sites with rather low accuracy.[92] Combining several methods with weighing of their results provides some improvement of the prediction accuracy.[93]

The low efficiency of the above-mentioned methods may be associated with the following factors. The use of a large number of parameters is not justified in the case of a small training set, besides containing some percentage of false negative examples.[22] Approaches that should take into

account interpositional relationships and frequency characteristics of the positions in the pattern do not provide substantial advantages due probably to an insufficient size of the training sets and small length of specific regions. The low specificity of the above-mentioned methods is in part explained by the fact that the negative control in these methods is usually accumulated from regions surrounding Ser, Thr or Tyr residues, in which the phosphorylation was not identified experimentally, *i.e.*, from the same amino acid sequences for which the control is positive. It is difficult to propose an alternative solution. However, it should be remembered that negative examples can contain modification sites that are still not identified.

The degeneration of phosphorylation motifs is responsible for low specificity of the recognition of interacting molecules. Hence, in order to get a clear understanding of the mechanisms of phosphorylation, it is necessary to take into account all complex regulatory interaction networks, where phosphorylation events can be considered as individual steps.[68]

In order to increase the accuracy of phosphorylation site prediction, developers of new methods utilize information on evolutionary relationships and regulatory interactions.

Evolutionary relationships of related proteins are studied to make a more accurate prediction of phosphorylation. The presence of phosphorylation sites in conserved regions of homologous proteins identified by the alignment of amino acid sequences is considered as the support of their involvement in similar signalling pathways in related organisms.[94] The phosphorylation site conservation compared to non-phosphorylated Ser, Thr and Tyr residues was mentioned by other authors as well.[95–98] Methods using multiple alignment were developed for the phosphorylation site prediction based on such observations. The iterative construction of position-specific scoring matrices with the PSI-Blast method[99] allows taking into account the evolutionary stability of modification sites and neighbouring positions for their prediction. This approach was implemented in the PPRED[100] and PostMod programmes.[101]

Methods employing the alignment have limitations. For example, mammalian orthologous proteins can differ by certain phosphorylation sites, even in highly conserved regions.[102] Examples were found in which phosphorylation sites are shifted in the aligned sequences of orthologous proteins.[103]

The modification site conservation is estimated using the CS (Conservation Score) method based on the alignment of the sequence under study and its homologues.[104] The degree of conservation of the region is determined by taking into account phylogenetic relationships between the given protein and its homologues. The high conservation score is considered as an additional support that the predicted site is actually functional. The conservation scores obtained using the CS algorithm are included in records of the Phospho.ELM database.

Phylogenetic models are implemented in the phyloHMM method. The latter was used for the identification of short conserved sequences in disordered regions of proteins from the yeast proteome.[105] Nguyen Ba *et al.*[105] identified the known functional modules (including phosphorylation sites) and the previously unknown conserved regions. Short conserved regions are often present in disordered areas of hub proteins (proteins involved in many regulatory pathways).

Disordered regions of proteins are characterized by a specific amino acid composition. Due to the limited set of the types of amino acids used and a small length of linear motifs, the alignment programmes often cannot choose the best variant of sequence matching. The possibility of coupled mutations cannot be ruled out as well.

Apparently, methods based on the alignment allow users to identify conserved phosphorylation motifs in functionally related homologous proteins. This approach is less suitable for the identification of regions in more distantly related proteins.[106]

The phosphorylation site conservation can be due to the fact that the respective proteins play a similar role in the same regulatory pathways. Let us consider SMAD proteins involved in the signal transduction from the receptor of the transforming growth factor β (TGF-beta).[107] The SMAD2 and SMAD3 proteins belong to the same functional R-SMAD group, whereas SMAD4 proteins are considered as the separate coSMAD group. The alignment of the corresponding amino acid sequences (Fig. 4) shows that in most cases the phosphorylation sites in SMAD2 and SMAD3 either coincide or are located in the same alignment region.

Five phosphorylation sites were experimentally found for the SMAD4 protein. Three of these sites are located in the more conserved N-terminal region and match upon the alignment with the sites of SMAD2 and SMAD3. Two other sites are present in the insertions, and are associated with the functional differences of the proteins under consideration.

This example shows that functionally diverged homologues can have different phosphorylation sites. Such web servers, as dbPTM,[49] PhosphoSitePlus,[54] Phospho.Elm[50] and PHOSIDA,[87] allow users to align phosphorylation sites with amino acid sequences retrieved from the commonly available resources (for example, UniProt). Hence, users can score the degree of conservation/variability of modification sites in homologous proteins.

As mentioned above, the enzymatic reactions generating PTMs are involved in complex processes, in the course of which the interaction between the modifying enzyme and the substrate protein is determined by the space-time arrangement of intermolecular interactions;[108] the latter can be both direct and mediated.[109]

The specificity of protein kinases depends on such factors as the involvement of adaptors and scaffold proteins, coexpression, subcellular colocalization and so on. In order to take into account all these factors, it is necessary to consider the results of the site prediction combined with the data from different resources of protein–protein interactions, regulatory network models, *etc.* Many commonly used web servers for the prediction of phosphorylation sites contains references, which allow users to extract additional information from available sources.

The automated integration of data on phosphorylation, complex protein–protein interactions and regulatory pathways was implemented for the first time in the NetWorKin programme.[110, 111] The proposed approach utilizes information on experimentally identified and predicted protein–protein regulatory interactions; this information can be extracted from the String database.[56] The NetWorKin method was developed for the prediction of protein kinase types responsible for modification of the known sites; the data on sites are extracted from Phospho.Elm[50] and PhosphoSitePlus.[54] This problem is very important because the

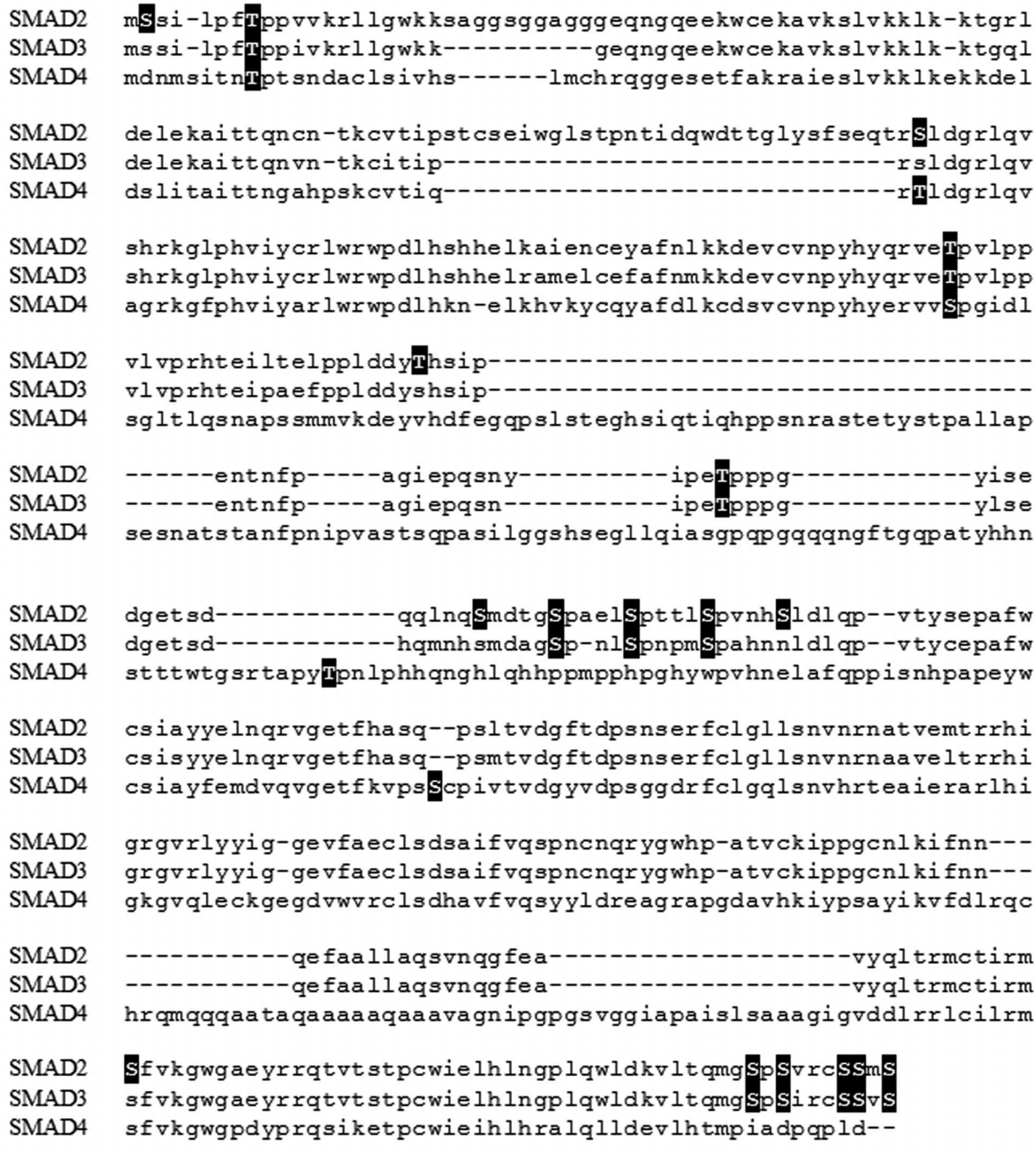

Figure 4. Alignment of SMAD proteins. Phosphorylation sites are highlighted.

information resources do not contain data on the types of enzymes for two-thirds of phosphorylation sites. The algorithm includes a two-step procedure. In the first step, motifs specific for different subfamilies of protein kinases are determined for specified modification sites using the Scansite and NetPhosK programmes. From each group, an enzyme is chosen with the smallest distance between the corresponding node and the node of the expected substrate in the network extracted from the String database.[56] The aim is to estimate the probability of the interaction and exclude less probable variants. The NetWorKin resource containing data on phosphorylation sites from PhosphoElm allows users to study network interactions, which include phosphorylation events.[111] The NetWorKin algorithm is implemented also in the PhosphoSiteAnalyzer bioinformatic platform[112] designed in order to explore large phosphoproteomic data sets.

In the iGPS programme, information on the experimentally determined and predicted protein – protein interactions is taken into account.[113]

Another group of researchers[114] proposed to consider the factors responsible for the specificity of modifications using characteristics of phosphorylated proteins compiled from various resources: KEGG[57] (involvement in regula-

tory pathways), Gene Ontology[59] (localization, biological process, molecular function), String[56] (protein–protein interactions), Pfam[58] and InterPro[115] (domain composition). Based on these data, a set of features was generated and the score was developed to estimate the probability that the protein under consideration serves as a substrate for this type of protein kinases. The study of the feature space using the support vector machine (SVM) revealed the factors most significant for determining the specificity of phosphorylation sites and separate proteins for the particular enzyme types. The influence of various factors was different for different types of protein kinases. However, it was mentioned that, in addition to the amino acids surrounding the modification site, the specificity of the enzyme substantially depends also on the cellular localization and protein–protein interactions.

The identification of local similarity regions in amino acid sequences can indicate the presence of modules responsible for interactions with the same or similar proteins. The approach is implemented in the PAAS (Projection of Amino Acid Sequences) method, which allows the prediction of substrate proteins for particular types of protein kinases without identification of modification sites.[116]

The approach based on investigation of the domain composition of protein kinases and their substrates was proposed by Liu and Tozeren.[117] The authors determined the domain profiles in amino acid sequences and then found statistically significant domain pairs in partner proteins. They succeeded in distinguishing stable domain complexes (domain strings), which characterize the specificity of interactions between different types of protein kinases and their partners.

The use of information on complex interactions for the prediction of PTMs seems to be quite justified. However, the efficiency of methods depends on the completeness and reliability of these data. The involvement of information on predicted interactions can reduce the advantages of the approach under consideration almost to zero. The experimental data on interactions collected in the respective resources should be qualified with different degrees of confidence depending on the methods used for their obtaining.[118]

To compare the efficiency of various methods, we examined several programmes used for the prediction of phosphorylation sites in the SMAD2 protein at Ser and Thr residues. Table 1 summarizes the results of prediction of the experimentally determined phosphorylation sites (dbPTM[49] database) without taking into account the enzyme specificity. The Ser or Thr residues, which are not marked as phosphorylation sites, are considered as negative examples.

The amino acid sequence of the human SMAD2 protein contains 75 Ser and Thr residues. The GPS programme recognized the largest number of known sites (16 of 17), but predicted phosphorylation sites in most of positions occupied by Ser or Thr residues, *i.e.*, showed very low specificity. The KinasePhos method, although being slightly less sensitive, also showed very low specificity. The NetPhosK method allowed us to slightly increase the specificity due to a decrease in the number of incorrectly predicted negative samples. The programme available *via* the PHOSIDA web server provided a even larger increase in the specificity but at the expense of a considerable decrease in the sensitivity. The use of a filter in the DISPHOS programme to exclude conformationally ordered regions led to a sharp decrease in both the total number of predicted sites and the number of correctly predicted sites. In all cases, the balanced accuracy (the half-sum of specificity and sensitivity) was low and reached the maximum value (0.65) with the use of the iGPS programme. In the latter case, the filter based on the

Table 1. Prediction of phosphorylation at Ser and Thr residues in the known sites of the SMAD2 protein. Positive results are represented by shaded rectangles.

Site	KinasePhos2	GPS	NetPhosK	Phosida	DISPHOS	iGPS
S2		■	■			■
T8	■	■	■			■
S110	■	■		■		
T172	■	■	■			■
T197	■					■
T220	■	■	■	■	■	■
S240	■	■	■			
S245	■	■	■	■		■
S250	■	■	■			■
S255	■	■				■
S260	■	■	■			
S417	■	■	■			
S458	■	■	■	■		■
S460	■	■	■	■		
S464		■	■			■
S465		■	■		■	■
S467		■			■	■
Correctly predicted	13	16	13	5	3	12
All predicted	70	63	43	24	5	35
Sensitivity	0.76	0.94	0.76	0.29	0.18	0.71
Specificity	0.02	0.19	0.48	0.67	0.97	0.60
Balanced accuracy	0.39	0.57	0.62	0.48	0.57	0.65

experimentally determined protein–protein interactions was applied. This made it possible to achieve the most reasonable ratio of sensitivity and specificity; this result can be attributed to a good knowledge of regulatory interactions in the SMAD2 protein.

3. Prediction of acetylation at lysine residues

The reversible acetylation of Lys residues was revealed for the first time for histone molecules. The switching of their functional state by means of this modification plays an important role in the gene transcription regulation.[119, 120] Later on it was shown that the acetylation of lysine residues in proteins of other classes is necessary for the regulation of cell signal transduction at different levels.[121]

The human genome encodes more than twenty lysine acetyltransferases exhibiting different substrate specificity. Their classification is a difficult problem because of high variability of amino acid sequences. Nine lysine acetyltransferases are grouped into three families: CBP/p300, GCN5/PCAF and MYST.[122] In most of the methods used, Lys acetylation sites are predicted without taking into account the specificity of acetyltransferase based solely on the data on the amino acid composition of positions in the direct vicinity of the acetylated Lys residue. The experimentally determined regions containing acetylated and non-acetylated Lys residues are used for the generation of training sets, which are processed using various algorithms.

The PAIL (Prediction of Acetylation of Internal Lysine) algorithm [123] implements Bayesian discriminant analysis. The probability scores for acetylation are calculated based on the analysis of the frequencies of amino acid occurrence in positions around modified and non-modified Lys residues. In the programme designed by Gnad *et al.*,[124] regions of amino acid sequence are represented in the feature space. Each dimension of the space is determined by the sequence position number and the type of the residue. The support vector machine method is applied to recognize modified regions. The same approach is implemented in the LysAcet programme.[125] A support vector machine ensemble is used in the method proposed in the study.[126]

The very simple predMod algorithm [127] employs the clusterization of modified and non-modified regions, which are pre-aligned based on the central Lys residue. A comparison of the amino acid sequence regions under study with the regions from different clusters allows the prediction of acetylation sites in histones and cytoplasmic proteins.

Regular expressions constructed with the use of the above-mentioned Motif-x method are used also for the prediction of acetylation sites *via* the scan-x web site.[128] The prediction of Lys acetylation sites based solely on sequences of site-surrounding regions has a relatively low accuracy, particularly in the case of a comparative trials of several methods using the same test set. Too optimistic estimates of the quality of the predictions often reported in the original publications may be associated with the fact that the samples under test, which are used to estimate the efficiency of the prediction, are included in the training set employed for the optimization of the algorithm.[129]

It would be expected that the generation of training sets for Lys acetylation sites taking into account the type of modifying enzymes will lead to an increase in the efficiency of the prediction. The ASEB programme predicts acetylation sites taking into account the enzyme specificity.[122] The estimate of the probability of the presence of the potential acetylation site is calculated based on the similarity of amino acid sequences.[130] In addition, data providing evidence for the interaction (either direct or mediated) between the enzyme and the substrate protein can be utilized. These relationships are determined from the analysis of regulatory networks (retrived from the STRING [56] and PINA [131] databases) by calculating the shortest distance between enzyme and the substrate in the graph of relationships.

4. Prediction of glycosylation

The term 'glycosylation' includes several different PTMs, which lead to changes in the structure and function of modified proteins.[132] The N-linked glycosylation results in the attachment of oligosaccharides to the NH_2 group of asparagine. This modification is important for the folding of membrane and secretory proteins, protein subcellular localization, the formation of the immune response and so on.[133] The O-linked glycosylation occurs through the attachment of sugar residues to the oxygen atoms of Ser or Thr side chains.[134] The mucin-type O-linked glycosylation was studied in most detail. This process results in the binding of *N*-acetylgalactosamine to which, in turn, a branched polysaccharide structure is attached. These proteins are involved in cell adhesion and the formation of the intracellular matrix.[135] The O-linked glycosylation, which leads to the reversible binding of *N*-acetylglucosamine, is shown for nuclear and cytoplasmic proteins involved in the transcription, ubiquitination, cell cycle, stress responses and so on.[136] The C-linked glycosylation resulting in the attachment of mannose to the C(2) atom of the indole ring of tryptophan was described in the literature.[137, 138]

For N-linked glycosylation sites, the characteristic regular expression N–{P}–[ST]–{P} was determined. According to this expression, the first position can be occupied only by Asn; the second and fourth positions, by any residue except for Pro; the third position, by Ser or Thr.[15] For C-linked glycosylation sites, the specific pattern W–X–X–[WFYC] was determined, according to which the first position should be occupied by Trp; the fourth position, by Trp, Phe, Tyr or Cys; the second and third positions, by any residue.[138] It is believed that O-linked glycosylation patterns are too vague to be represented as a regular expression. It was reported [139] that particular positions in the neighbouring environment of mucin-type O-linked glycosylation sites can be responsible for the specificity for different types of modifying enzymes.

In the NetNGlyc,[23] NetCGlyc [140] and NetOGlyc algorithms,[141, 142] neural networks are used for the recognition of N-, C- and O-linked (mucin type) glycosylation sites. Both amino acid sequence regions and predicted structural characteristics (surface localization, secondary structure, *etc.*) can be used as training data. To more clearly identify regions of the protein under consideration that can be actually glycosylated, the sequence under study should be analyzed for the presence of transmembrane regions and cellular localization signals, thus revealing extracellular regions.

Different approaches were implemented in such programmes as EnsembleGly.[143] The latter employs support vector machine ensembles, which are trained using positive and negative examples of modification regions, for the prediction of N-, O- and C-linked glycosylation sites. The support vector machine method is applied also for the prediction of O-linked glycosylation sites with attached

N-acetylglucosamine[65] and is implemented in the CKSAAP_OGlySite programme[144] for the recognition of mucin-type O-linked glycosylation sites. The GPP (Glycosylation Prediction Programme) method[145] iutilizes a random forest algorithm using frequency characteristics of amino acid residues in positions surrounding the modified residue.

VI. Conclusion

Post-translational modifications are events that are very important steps in complex intra- and intercellular processes of functioning of biological systems. Thus, phosphorylation reactions are involved in cell signal transduction, the acetylation is considered as mechanism of gene activity regulation and the glycosylation provides the transport of membrane and secretory proteins. Despite a great progress in the development of high-throughput experimental methods of isolation, expression and analysis of proteins, these methods do not allow researchers to found all possible modifications. Hence the *in silico* prediction of PTMs is important both for the design of experimental research and the comprehensive analysis of the regulatory mechanisms of the functioning of living systems.

Numerous methods were developed for the recognition of potential PTM sites in amino acid sequences of proteins. However, the analysis of the results obtained by these methods shows that the use of only short modification regions in amino acid sequences of substrate proteins for the prediction of PTMs does not provide sufficient accuracy even with the use of complex statistical methods. This is due to at least two factors.

First, post-translational modifications of a particular type are performed by a group of enzymes with different substrate specificity. It should be noted that the available experimental data are not always sufficient to generate representative training sets classified according to the specificity of modifying enzymes. At the present time, data on the substrate specificity are most completely used in methods for the prediction of phosphorylation sites.

Second, the above-mentioned methods do not take into account regulatory interactions involving the modifying enzyme and the substrate protein. Short linear regions cannot by themselves completely determine PTMs of a particular type. Short modules containing modification sites can ensure the specificity only when they are combined with other regions in the substrate protein molecule responsible for the interaction with the enzyme or the third protein, thus providing the required arrangement of the enzyme and substrate molecules. The use of data on direct and mediated interactions between modifying enzymes and substrate proteins seems to be most promising for the solution of the problem under consideration. However, special attention should be given to the selection of experimental data used as a base for the recognition of PTM patterns.

The identification of protein post-translational modifications are necessary for investigation of complex regulatory networks in living systems both in physiological and pathological states. Taking into account the above-mentioned limitations of experimental methods, the application and development of bioinformatic approaches for predicting PTMs is a promising field of systematic biology.

This review has been written with the financial support of the Ministry of Education and Science of the Russian Federation (Agreement No. 8274).

References

1. The UniProt Consortium *Nucleic Acids Res.* **40** (Database Issue) D71 (2012)
2. International Human Genome Sequencing Consortium *Nature (London)* **431** 931 (2004)
3. T W Nilsen, B R Graveley *Nature (London)* **463** 457 (2010)
4. C T Walsh, S Garneau-Tsodikova, G J Gatto Jr *Angew. Chem., Int. Ed.* **44** 7342 (2005)
5. K S Kamath, M S Vasavada, S Srivastava *J. Proteomics* **75** 127 (2011)
6. O N Jensen *Curr. Opin. Chem. Biol.* **8** 33 (2004)
7. V G Zgoda, A T Kopylov, O V Tikhonova, A A Moisa, N V Pyndyk, T E Farafonova, S E Novikova, A V Lisitsa, E A Ponomarenko, E V Poverennaya, S P Radko, S A Khmeleva, L K Kurbatov, A D Filimonov, N A Bogolyubova, E V Ilgisonis, A L Chernobrovkin, A S Ivanov, A E Medvedev, Y V Mezentsev, S A Moshkovskii, S N Naryzhny, E N Ilina, E S Kostrjukova, D G Alexeev, A V Tyakht, V M Govorun, A I Archakov *J. Proteome Res.* **12** 123 (2013)
8. E Basle, N Joubert, M Pucheault *Chem. Biol.* **17** 213 (2010)
9. G A Khoury, R C Baliban, C A Floudas *Sci. Rep. (Nature)* **1** Art. No 90 (2011)
10. Y Xiong, K L Guan *J. Cell Biol.* **198** 155 (2012)
11. T Ravid, M Hochstrasser *Nat. Rev. Mol. Cell Biol.* **9** 679 (2008)
12. J Eswaran, S Knapp *Biochim. Biophys. Acta: Proteins Proteomics* **1804** 429 (2010)
13. H R Christofk, N Wu, L C Cantley, J M Asara *J. Proteome Res.* **10** 4158 (2011)
14. S Dwane, P A Kiely *Bioeng. Bugs* **2** 247 (2011)
15. F Schwarz, M Aebi *Curr. Opin. Struct. Biol.* **21** 576 (2011)
16. S Li, L M Iakoucheva, S D Mooney, P Radivojac, in *Proceedings of Pacific Symposium on Biocomputing, Fairmont Orchid, Hawaii 2010* p. 337
17. J Brognard, T Hunter *Curr. Opin. Genet. Dev.* **21** 4 (2011)
18. A Iyer, D P Fairlie, L Brown *Immunol. Cell Biol.* **90** 39 (2012)
19. A Linares, F Dalenc, P Balaguer, N Boulle, V Cavailles *J. Biomed. Biotechnol.* **2011** 856985 (2011)
20. K Ohtsubo, J D Marth *Cell* **126** 855 (2006)
21. A R Farley, A J Link, in *Guide to Protein Purification (Methods in Enzymology* Vol. 463) (2nd Ed.) (Eds R R Burgess, M P Deutscher) (Amsterdam: Elsevier, Boston: Academic Press, 2009) p. 725
22. B Eisenhaber, F Eisenhaber, in *Data Mining Techniques for the Life Sciences (Methods in Molecular Biology* Vol. 609) (Eds O Carugo, F Eisenhaber) (New York: Humana Press, 2010) p. 365
23. N Blom, T Sicheritz-Pontén, R Gupta, S Gammeltoft, S Brunak *Proteomics* **4** 1633 (2004)
24. M L Miller, N Blom, in *Phospho-Proteomics: Methods and Protocols (Methods in Molecular Biology* Vol. 527) (Ed. M de Graauw) (New York: Humana Press, 2009) p. 299
25. Y Xue, X Gao, J Cao, Z Liu, C Jin, L Wen, X Yao, J Ren *Curr. Protein Pept. Sci.* **11** 485 (2010)
26. C Liu, H Li, in *In Silico Tools for Gene Discovery (Methods in Molecular Biology* Vol. 760) (Eds B Yu, M Hinchcliffe) (New York: Humana Press, 2011) p. 325
27. B Trost, A Kusalik *Bioinformatics* **27** 2927 (2011)
28. M Frank, S Schloissnig *Cell. Mol. Life Sci.* **67** 2749 (2010)
29. A I Archakov, V M Govorun, A V Dubanov, Y D Ivanov, A V Veselovsky, P Lewi, P Janssen *Proteomics* **3** 380 (2003)

30. S Ren, G Yang, Y He, Y Wang, Y Li, Z Chen *BMC Genomics* **9** 452 (2008)
31. E Akiva, G Friedlander, Z Itzhaki, H Margalit *PLoS Comput. Biol.* **8** e1002341 (2012)
32. B A Liu, B W Engelmann, P D Nash *Proteomics* **12** 1527 (2012)
33. J Vyas, R J Nowling, M W Maciejewski, S Rajasekaran, M R Gryk, M R Schiller *BMC Genomics* **10** 360 (2009)
34. N E Davey, K Van Roey, R J Weatheritt, G Toedt, B Uyar, B Altenberg, A Budd, F Diella, H Dinkel, T J Gibson *Mol. Biosyst.* **8** 268 (2012)
35. J C Merlin, S Rajasekaran, T Mi, M R Schiller *PLoS One* **7** e32630 (2012)
36. H Dinkel, S Michael, R J Weatheritt, N E Davey, K Van Roey, B Altenberg, G Toedt, B Uyar, M Seiler, A Budd, L Jödicke, M A Dammert, C Schroeter, M Hammer, T Schmidt, P Jehl, C McGuigan, M Dymecka, C Chica, K Luck, A Via, A Chatr-Aryamontri, N Haslam, G Grebnev, R J Edwards, M O Steinmetz, H Meiselbach, F Diella, T J Gibson *Nucleic Acids Res.* **40** (Database Issue) D242 (2012)
37. R J Weatheritt, P Jehl, H Dinkel, T J Gibson *Nucleic Acids Res.* **40** (Web Server Issue) W364 (2012)
38. V Neduva, R B Russell *FEBS Lett.* **579** 3342 (2005)
39. P M Holland, J A Cooper *Curr. Biol.* **9** R329 (1999)
40. D L Sheridan, Y Kong, S A Parker, K N Dalby, B E Turk *J. Biol. Chem.* **283** 19511 (2008)
41. A J Bardwell, E Frankson, L Bardwell *J. Biol. Chem.* **284** 13165 (2009)
42. K A Burkhard, F Chen, P Shapiro *J. Biol. Chem.* **286** 2477 (2011)
43. W T Miller *Acc. Chem. Res.* **36** 393 (2003)
44. R P Bhattacharyya, A Remenyi, B J Yeh, W A Lim *Annu. Rev. Biochem.* **75** 655 (2006)
45. M C Good, J G Zalatan, W A Lim *Science* **332** 680 (2011)
46. J Bloom, F R Cross *Nat. Rev. Mol. Cell Biol.* **8** 149 (2007)
47. J Gao, D Xu, in *Proceedings of Pacific Symposium on Biocomputing Fairmont Orchid, Hawaii, 2012* p. 94
48. A-J Petrescu, A-L Milac, S M Petrescu, R A Dwek, M R Wormald *Glycobiology* **14** 103 (2004)
49. C T Lu, K Y Huang, M G Su, T Y Lee, N A BretaKha, W C Chang, Y J Chen, Y J Chen, H D Huang *Nucleic Acids Res.* **41** (D1) D295 (2013)
50. H Dinkel, C Chica, A Via, C M Gould, L J Jensen, T J Gibson, F Diella *Nucleic Acids Res.* **39** (Database Issue) D261 (2011)
51. A L Chernorudskiy, A Garcia, E V Eremin, A S Shorina, E V Kondratieva, M R Gainullin *BMC Bioinformatics* **8** 126 (2007)
52. R Gupta, H Birch, K Rapacki, S Brunak, J E Hansen *Nucleic Acids Res.* **27** 370 (1999)
53. R Goel, H C Harsha, A Pandey, T S Prasad *Mol. Biosyst.* **8** 453 (2012)
54. P V Hornbeck, J M Kornhauser, S Tkachev, B Zhang, E Skrzypek, B Murray, V Latham, M Sullivan *Nucleic Acids Res.* **40** (Database Issue) D261 (2012)
55. H Li, X Xing, G Ding, Q Li, C Wang, L Xie, R Zeng, Y Li *Mol. Cell. Proteomics* **8** 1839 (2009)
56. A Franceschini, D Szklarczyk, S Frankild, M Kuhn, M Simonovic, A Roth, J Lin, P Minguez, P Bork, C von Mering, L J Jensen *Nucleic Acids Res.* **41** (D1) D808 (2013)
57. M Kanehisa, S Goto, Y Sato, M Furumichi, M Tanabe *Nucleic Acids Res.* **40** (Database Issue) D109 (2012)
58. M Punta, P C Coggill, R Y Eberhardt, J Mistry, J Tate, C Boursnell, N Pang, K Forslund, G Ceric, J Clements, A Heger, L Holm, E L L Sonnhammer, S R Eddy, A Bateman, R D Finn *Nucleic Acids Res.* **40** (Database Issue) D290 (2012)
59. R Roslan, R M Othman, Z A Shah, S Kasim, H Asmuni, J Taliba, R Hassan, Z Zakaria *Comput. Biol. Med.* **40** 555 (2010)
60. A Zanzoni, D Carbajo, F Diella, P F Gherardini, A Tramontano, M Helmer-Citterich, A Via *Nucleic Acids Res.* **39** (Database Issue) D268 (2011)
61. Z Liu, J Cao, X Gao, Y Zhou, L Wen, X Yang, X Yao, J Ren, Y Xue *Nucleic Acids Res.* **39** (Database Issue) D1029 (2011)
62. H Zhang, P Loriaux, J Eng, D Campbell, A Keller, P Moss, R Bonneau, N Zhang, Y Zhou, B Wollscheid, K Cooke, E C Yi, H Lee, E R Peskind, J Zhang, R D Smith, R Aebersold *Genome Biol.* **7** R73 (2006)
63. C A Cooper, H J Joshi, M J Harrison, M R Wilkins, N H Packer *Nucleic Acids Res.* **31** 511 (2003)
64. R Ranzinger, S Herget, C W von der Lieth, M Frank *Nucleic Acids Res.* **39** (Database Issue) D373 (2011)
65. J Wang, M Torii, H Liu, G W Hart, Z-Z Hu *BMC Bioinformatics* **12** Art. No 91 (2011)
66. J Mok, X Zhu, M Snyder *Expert Rev. Proteomics* **8** 775 (2011)
67. W A Lim, T Pawson *Cell* **142** 661 (2010)
68. J Lin, Z Xie, H Zhu, J Qian *Brief. Funct. Genomics* **9** 32 (2010)
69. L Wang, L Hou, M Qian, M Deng *PLoS One* **7** e33160 (2012)
70. J V Olsen, M Vermeulen, A Santamaria, C Kumar, M L Miller, L J Jensen, F Gnad, J Cox, T S Jensen, E A Nigg, S Brunak, M Mann *Sci. Signal.* **3** ra3 (2010)
71. H Imamura, N Yachie, R Saito, Y Ishihama, M Tomita *BMC Bioinformatics* **11** 232 (2010)
72. J Jin, T Pawson *Philos. Trans. R. Soc. London, Ser. B: Biol. Sci.* **367** 2540 (2012)
73. A Plotnikov, E Zehorai, S Procaccia, R Seger *Biochim. Biophys. Acta: Mol. Cell Res.* **1813** 1619 (2011)
74. Y Xue, J Ren, X Gao, C Jin, L Wen, X Yao *Mol. Cell. Proteomics* **7** 1598 (2008)
75. Y Xue, Z Liu, J Cao, Q Ma, X Gao, Q Wang, C Jin, Y Zhou, L Wen, J Ren *Protein Eng. Des. Sel.* **24** 255 (2011)
76. C J Sigrist, L Cerutti, N Hulo, A Gattiker, L Falquet, M Pagni, A Bairoch, P Bucher *Brief. Bioinform.* **3** 265 (2002)
77. M F Chou, D Schwartz *Curr. Protoc. Bioinformatics* Chapter 13 Unit 13.15-24 (2011)
78. Y C Chen, K Aguan, C W Yang, Y T Wang, N R Pal, I F Chung *PLoS One* **6** e20025 (2011)
79. A Ritz, G Shakhnarovich, A R Salomon, B J Raphael *Bioinformatics* **25** 14 (2009)
80. D Schwartz, M F Chou, G M Church *Mol. Cell. Proteomics* **8** 365 (2009)
81. J C Obenauer, L C Cantley, M B Yaffe *Nucleic Acids Res.* **31** 3635 (2003)
82. J J Ellis, B Kobe *PLoS One* **6** e21169 (2011)
83. N Blom, S Gammeltoft, S Brunak *J. Mol. Biol.* **294** 1351 (1999)
84. D Plewczynski, S Basu, I Saha *Amino Acids* **43** 573 (2012)
85. J H Kim, J Lee, B Oh, K Kimm, I Koh *Bioinformatics* **20** 3179 (2004)
86. Y-H Wong, T-Y Lee, H-K Liang, C-M Huang, T-Y Wang, Y-H Yang, C-H Chu, H-D Huang, M-T Ko, J-K Hwang *Nucleic Acids Res.* **35** (Web Server Issue) W588 (2007)
87. F Gnad, J Gunawardena, M Mann *Nucleic Acids Res.* **39** Database Issue) D253 (2011)
88. Y Xue, A Li, L Wang, H Feng, X Yao *BMC Bioinformatics* **7** 163 (2006)
89. L M Iakoucheva, P Radivojac, C J Brown, T R O'Connor, J G Sikes, Z Obradovic, A K Dunker *Nucleic Acids Res.* **32** 1037 (2004)
90. G Neuberger, G Schneider, F Eisenhaber *Biol. Direct* **2** Art. No 1 (2007)
91. P Durek, C Schudoma, W Weckwerth, J Selbig, D Walther *BMC Bioinformatics* **10** 117 (2009)
92. S Que, Y Wang, P Chen, Y R Tang, Z Zhang, H He *Protein Peptide Lett.* **17** 64 (2010)

93. J Wan, S Kang, C Tang, J Yan, Y Ren, J Liu, X Gao, A Banerjee, L B Ellis, T Li *Nucleic Acids Res.* **36** e22 (2008)
94. Y V Budovskaya, J S Stephan, S J Deminoff, P K Herman *Proc. Natl. Acad. Sci. USA* **102** 13933 (2005)
95. J Boekhorst, B van Breukelen, A Heck Jr, B Snel *Genome Biol.* **9** R144 (2008)
96. B Macek, F Gnad, B Soufi, C Kumar, J V Olsen, I Mijakovic, M Mann *Mol. Cell. Proteomics* **7** 299 (2008)
97. R Malik, E A Nigg, R Körner *Bioinformatics* **24** 1426 (2008)
98. Y Z Kurmangaliyev, A Goland, M S Gelfand *Biol. Direct* **6** 8 (2011)
99. M Bhagwat, L Aravind, in *Comparative Genomics (Methods in Molecular Biology* Vol. 395) (Ed. N H Bergman) (Totowa, NJ: Humana Press, 2007) p. 177
100. A K Biswas, N Noman, A R Sikder *BMC Bioinformatics* **11** 273 (2010)
101. I Jung, A Matsuyama, M Yoshida, D Kim *BMC Bioinformatics* **11** (Suppl. 1) S10 (2010)
102. D S Kim, Y Hahn *Bioinformatics* **27** 2494 (2011)
103. L J Holt, B B Tuch, J Villén, A D Johnson, S P Gygi, D O Morgan *Science* **325** 1682 (2009)
104. C Chica, A Labarga, C M Gould, R López, T J Gibson *BMC Bioinformatics* **9** Art. No 229 (2008)
105. A N Nguyen Ba, B J Yeh, D van Dyk, A R Davidson, B J Andrews, E L Weiss, A M Moses *Sci. Signal* **5** (215) rs1 (2012)
106. E Perrodou, C Chica, O Poch, T J Gibson, J D Thompson *BMC Bioinformatics* **9** 213 (2008)
107. K H Wrighton, X Lin, X H Feng *Cell Res.* **19** 8 (2009)
108. B N Kholodenko *FEBS Lett.* **583** 4006 (2009)
109. J A Endicott, M E Noble, L N Johnson *Annu. Rev. Biochem.* **81** 587 (2012)
110. R Linding, L J Jensen, G J Ostheimer, M A T M van Vugt, C Jørgensen, I M Miron, F Diella, K Colwill, L Taylor, K Elder, P Metalnikov, V Nguyen, A Pasculescu, J Jin, J G Park, L D Samson, J R Woodgett, R B Russell, P Bork, M B Yaffe, T Pawson *Cell* **129** 1415 (2007)
111. R Linding, L J Jensen, A Pasculescu, M Olhovsky, K Colwill, P Bork, M B Yaffe, T Pawson *Nucleic Acids Res.* **36** (Database Issue) D695 (2008)
112. M V Bennetzen, J Cox, M Mann, J S Andersen *J. Proteome Res.* **11** 3480 (2012)
113. C Song, M Ye, Z Liu, H Cheng, X Jiang, G Han, Z Songyang, Y Tan, H Wang, J Ren, Y Xue, H Zou *Mol. Cell. Proteomics* **11** 1070 (2012)
114. T Li, P Du, N Xu *PLoS One* **5** e15411 (2010)
115. S Hunter, P Jones, A Mitchell, R Apweiler, T K Attwood, A Bateman, T Bernard, D Binns, P Bork, S Burge, E de Castro, P Coggill, M Corbett, U Das, L Daugherty, L Duquenne, R D Finn, M Fraser, J Gough, D Haft, N Hulo, D Kahn, E Kelly, I Letunic, D Lonsdale, R Lopez, M Madera, J Maslen, C McAnulla, J McDowall, C McMenamin, H Mi, P Mutowo-Muellenet, N Mulder, D Natale, C Orengo, S Pesseat, M Punta, A F Quinn, C Rivoire, A Sangrador-Vegas, J D Selengut, C J Sigrist, M Scheremetjew, J Tate, M Thimmajanarthanan, P D Thomas, C H Wu, C Yeats, S Y Yong *Nucleic Acids Res.* **40** (Database Issue) D306 (2012)
116. B Sobolev, D Filimonov, A Lagunin, A Zakharov, O Koborova, A Kel, V Poroikov *BMC Bioinformatics* **11** Art. No 313 (2010)
117. Y Liu, A Tozeren *BMC Bioinformatics* **11** Art. No 349 (2010)
118. M H Schaefer, J F Fontaine, A Vinayagam, P Porras, E E Wanker, M A Andrade-Navarro *PLoS One* **7** e31826 (2012)
119. V E MacDonald, L J Howe *Epigenetics* **4** 139 (2009)
120. I Scott *Essays Biochem.* **52** 13 (2012)
121. S Spange, T Wagner, T Heinzel, O H Krämer *Int. J. Biochem. Cell. Biol.* **41** 185 (2009)
122. L Wang, Y Du, M Lu, T Li *Nucleic Acids Res.* **40** (Web Server Issue) W376 (2012)
123. A Li, Y Xue, C Jin, M Wang, X Yao *Biochem. Biophys. Res. Commun.* **350** 818 (2006)
124. F Gnad, S Ren, C Choudhary, J Cox, M Mann *Bioinformatics* **26** 1666 (2010)
125. S Li, H Li, M Li, Y Shyr, L Xie, Y Li *Protein Peptide Lett.* **16** 977 (2009)
126. Y Xu, X-B Wang, J Ding, L-Y Wu, N-Y Deng *J. Theor. Biol.* **264** 130 (2010)
127. A Basu, K L Rose, J Zhang, R C Beavis, B Ueberheide, B A Garcia, B Chait, Y Zhao, D F Hunt, E Segal, C D Allis, S B Hake *Proc. Natl. Acad. Sci. USA* **106** 13785 (2009)
128. M F Chou, D Schwartz *Curr. Protoc. Bioinformatics* Chapter 13 Unit 13.16 (2011)
129. D Schwartz *Essays Biochem.* **52** 165 (2012)
130. T Li, Y Du, L Wang, L Huang, W Li, M Lu, X Zhang, W G Zhu *Mol. Cell. Proteomics* **11** (1) (2012); doi 10.1074/mcp.M111.011080
131. M J Cowley, M Pinese, K S Kassahn, N Waddell, J V Pearson, S M Grimmond, A V Biankin, S Hautaniemi, J Wu *Nucleic Acids Res.* **40** (Database Issue) D862 (2012)
132. K W Moremen, M Tiemeyer, A V Nairn *Nat. Rev. Mol. Cell Biol.* **13** 448 (2012)
133. A Larkin, B Imperiali *Biochemistry* **50** 4411 (2011)
134. J Peter-Katalinić, in *Mass Spectrometry: Modified Proteins and Glycoconjugates (Methods in Enzymology* Vol. 405 (Ed. A L Burlingame) (Amsterdam: Elsevier, Boston: Academic Press, 2005) p. 139
135. L Zhang, K G Ten Hagen *Biochem. Soc. Trans.* **39** 378 (2011)
136. C Butkinaree, K Park, G W Hart *Biochim. Biophys. Acta: General Subjects* **1800** 96 (2010)
137. M A Doucey, D Hess, R Cacan, J Hofsteenge *Mol. Biol. Cell* **9** 291 (1998)
138. A Furmanek, J Hofsteenge *Acta Biochim. Pol.* **47** 781 (2000)
139. T A Gerken, O Jamison, C L Perrine, J C Collette, H Moinova, L Ravi, S D Markowitz, W Shen, H Patel, L A Tabak *J. Biol. Chem.* **286** 14493 (2011)
140. K Julenius *Glycobiology* **17** 868 (2007)
141. J E Hansen, O Lund, N Tolstrup, A A Gooley, K L Williams, S Brunak *Glycoconjugate J.* **15** 115 (1998)
142. K Julenius, A Mølgaard, R Gupta, S Brunak *Glycobiology* **15** 153 (2005)
143. C Caragea, J Sinapov, A Silvescu, D Dobbs, V Honavar *BMC Bioinformatics* **8** 438 (2007)
144. Y Z Chen, Y R Tang, Z Y Sheng, Z Zhang *BMC Bioinformatics* **9** 101 (2008)
145. S E Hamby, J D Hirst *BMC Bioinformatics* **9** 500 (2008)