Review

Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence

Nikolaj Blom[1], Thomas Sicheritz-Pontén[1], Ramneek Gupta[1], Steen Gammeltoft[2] and Søren Brunak[1]

[1]Center for Biological Sequence Analysis, The Technical University of Denmark, Lyngby, Denmark
[2]Department of Clinical Biochemistry, Glostrup Hospital, Glostrup, Denmark

Post-translational modifications (PTMs) occur on almost all proteins analyzed to date. The function of a modified protein is often strongly affected by these modifications and therefore increased knowledge about the potential PTMs of a target protein may increase our understanding of the molecular processes in which it takes part. High-throughput methods for the identification of PTMs are being developed, in particular within the fields of proteomics and mass spectrometry. However, these methods are still in their early stages, and it is indeed advantageous to cut down on the number of experimental steps by integrating computational approaches into the validation procedures. Many advanced methods for the prediction of PTMs exist and many are made publicly available. We describe our experiences with the development of prediction methods for phosphorylation and glycosylation sites and the development of PTM-specific databases. In addition, we discuss novel ideas for PTM visualization (exemplified by kinase landscapes) and improvements for prediction specificity (by using ESS – evolutionary stable sites). As an example, we present a new method for kinase-specific prediction of phosphorylation sites, NetPhosK, which extends our earlier and more general tool, NetPhos. The new server, NetPhosK, is made publicly available at the URL http://www.cbs.dtu.dk/services/NetPhosK/. The issues of underestimation, over-prediction and strategies for improving prediction specificity are also discussed.

Keywords: Glycosylation / Kinase-specific / Neural networks / Phosphorylation / Prediction / Review

Received	2/12/03
Revised	23/1/04
Accepted	2/2/04

Contents

Correspondence: Dr. Nikolaj Blom, Center for Biological Sequence Analysis, BioCentrum-DTU, Building 208, The Technical University of Denmark, DK-2800 Lyngby, Denmark
E-mail: nikob@cbs.dtu.dk
Fax: +45-4593-1585

Abbreviations: ANN, artificial neural network; **CC**, correlation coefficient; **cdc2**, cell division cycle 2 kinase; **HMM**, hidden Markov model; **PKA**, cyclic AMP dependent protein kinase; **PKC**, protein kinase C; **PKG**, cyclic GMP-dependent protein kinase

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

1 Introduction

1.1 From sequence to function

In the early days of molecular biology, the function of a protein was typically known before the sequence of amino acids encoded by the gene was revealed. The advent of genomic sequencing has stood on its head this situation, and almost as long as sequences have been accumulating in the databases, pattern recognition methods for their functional analysis have been designed. Bioinformatics in the form of biological sequence analysis is not a new field, as it is often stated. There is a long tradition for creation of motif and consensus sequence methods for localization of features, which may provide hints to the local or global functional mechanisms of a protein.

Currently the number of experimentally validated examples of post-translational modifications (PTMs) grows tremendously. Not only examples of covalent modification, such as glycosylation or phosphorylation, but also for many different cases of protease processing and other types of signals that control protein sorting, protein half-life or folding. This type of data has led to an explosion in new types of prediction tools, many of which belong to the category of "machine learning techniques", where data can be converted into algorithms capable of predicting and locating such signals in biological sequences. This review focuses on this development and also addresses the link between protein feature prediction and overall functional assignment of proteins.

In the conventional picture, one needs to consider the protein as a three-dimensional structure with active sites on its surface that can bind to other proteins, or perform catalytic activity, in order to understand its function. The paradigm has been that protein sequence determines its structure, and knowing the structure yields the functional information [1, 2]. A somewhat complementary approach is currently being developed with bioinformatics, namely the idea of making use of protein features and then assign function using the features in an integrated fashion [3, 4]. Such features include global properties such as molecular weight, isoelectric point, localization signals and information about potential PTMs. This novel approach of including PTMs in function prediction is based on the assumption that overall structure and function of a mature protein in a cell is influenced significantly by PTMs and that a common cellular machinery processes all proteins within a cell, even those which are "alone" in sequence space (where alignment cannot provide functional hints). The aim is then to use a number of correlated features to predict the functional category of a protein.

1.2 Post-translational modifications

Most proteins do not perform their molecular function as unmodified folded polypeptides. In most cases, proteins need to acquire permanent or transient molecular features in order to function as they should. Post-translational modifications (PTMs) of proteins most often come in the form of proteolytic cleavage events or covalent modifications at specific amino acid residues. Proteolytic cleavage is of course an irreversible modification, while covalent modifications may be reversible, *e.g.* in the case of protein phosphorylation. In that case, modifications performed by the protein kinase may be reversed by the action of a complementary acting enzyme, the phosphatase, by removing the attached phosphate moiety. Virtually any of the 20 natural amino acids may be modified by some type of PTM as evidenced by the many examples shown in the RESID database [5].

After being synthesized (translated), the polypeptide chain is subject to many different types of post-translational processing in different cellular compartments, including the nucleus, cytosol, endoplasmic reticulum and Golgi apparatus. This happens during or after folding, and involves enzymatic processing including removal of one or more amino acids from the amino terminus, proteolytic cleavage, or addition of acetyl, phosphoryl, glycosyl, methyl or other groups to certain amino acid residues. These modifications may confer various structural and functional properties to the affected proteins.

A database of protein post-translational modifications with descriptive, chemical, structural and bibliographic information is available: RESID (http://pir.georgetown.edu/pirwww/dbinfo/resid.html) [5]. Certain types of modifications affect only specific amino acid residues. For example, phosphorylation mainly occurs on serine, threonine and tyrosine residues, while covalent glycosylation mainly occurs on asparagine, serine and threonine resi-

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

dues. Still, not all of these specific residues in a protein are actually modified. Often, the transferase involved (for enzymatic PTMs) recognizes sequence patterns (acceptor motifs) around the concerned amino acid. Such patterns are for example listed in the PROSITE database (http://www.expasy.org/prosite/) [6].

An acceptor motif does not necessarily mean a consensus sequence though. Some motifs are not defined well enough to produce a 'consensus', or in some cases a linear consensus is not possible due to the correlations between sequence positions within the context. The tyrosine sulfation site, for instance, has been characterized [7–9] as the presence of an acidic (Glu or Asp) amino acid within two residues of the tyrosine (typically at -1); the presence of at least three acidic residues from -5 to $+5$; no more than one basic residue and three hydrophobic residues from -5 to $+5$; presence of turn-inducing amino acids; absence of disulfide-bonded cysteine residues from -7 to $+7$; absence of *N*-linked glycans near the tyrosine. Even with these restrictions, the tyrosine sulfation acceptor site is still not adequately defined for a consensus sequence [10, 11]. Thus, the prediction of PTMs is not at all a trivial task.

1.3 Glycosylation

Many proteins in eukaryotic cells are glycoproteins as they contain oligosaccharide chains covalently linked to certain amino acids. With the notable exception of glycation [12] which is a nonenzymatic reaction involving the addition of (mostly) glucose or fructose to lysine residues, glycosylation occurs enzymatically in biological systems. Glycosylation is known to affect protein folding, localization and trafficking, protein solubility, antigenicity, biological activity and half-life, as well as cell-cell interactions [13–15]. Incidents of prokaryotic glycosylation have been reported in the literature (for a review, see [16]), but shall not be discussed in detail in this review.

Protein glycosylation can be divided into four main categories mainly depending on the linkage between the amino acid and the sugar. These are *N*-linked glycosylation, *O*-linked glycosylation, *C*-mannosylation and glycophosphatidlyinositol (GPI) anchor attachments. *N*-glycosylation is characterized by the addition of a sugar to the amino group (NH_2) of an asparagine. In *O*-glycosylation, a sugar is attached to the hydroxyl group (OH^-) of a serine or threonine residue being modified.

GPI anchors refer to glycophosphatidyl-inositol groups attached near the *C*-terminal of a protein chain, that anchor the protein to the cell membrane. Recently, a method has been made available for predicting these sites (http://mendel.imp.univie.ac.at/gpi/) [17].

C-mannosylation is the attachment of an α-mannopyranosyl residue to the indole C2 of tryptophan *via* a C-C link [18], and occurs on the first tryptophan in the motif W-X-X-W (or in some cases, W-X-X-C and W-X-X-F) [19, 20]. So far, there has been little experimentally verified site-mapped data [19–21] for this type of modification, but the W-X-X-W motif does occur in 2917 mammalian proteins in Swiss-Prot V. 42.6 [20].

1.3.1 *N*-linked glycosylation

Oligosaccharides attached to Asn residues of secreted or membrane bound proteins are described as *N*-linked. *N*-linked glycoforms fall into three main categories: high mannose, hybrid and complex. All of these are derived from a precursor oligosaccharide comprising $GlcNAc_2$ Man_9Glc_3. The Dol-P-$GlcNAc_2Man_9Glc_3$ precursor is added cotranslationally *en bloc* to the amide group of an asparagine residue. The process occurs in the endoplasmic reticulum (ER) and is known to influence protein folding. The sequence motif Asn-Xaa-Ser/Thr (Xaa is any amino acid except Pro) has been defined as a prerequisite for *N*-glycosylation [22]. Although rare, the sequence motif Asn-Xaa-Cys has also been shown to act as an acceptor site [23]. Some studies propose that the acceptor motif Asn-Xaa-Ser is less well utilized compared to Asn-Xaa-Thr [24]. The addition of the *N*-linked precursor is catalyzed by an oligosaccharyltransferase in the ER [25] and is a conserved process through eukaryote evolution. However, the ability to form hybrid or complex *N*-glycans varies in different eukaryotic systems and during development within a given system or cell type. The sequence motif described above is not sufficient to act as a glycosylation site, though it does appear to be a prerequisite.

1.3.2 *O*-linked glycosylation

O-linked glycosylation reactions may happen at two cellular locations in the cell. Those taking place in the Golgi are initiated by the addition of various reducing terminal linkages such as *N*-acetylgalactosamine, *N*-acetylglucosamine, mannose, fucose, phosphodiester linked *N*-acetylglucosamine, glucose, galactose or xylose to hydroxyl amino acids (usually serine or threonine). These often lead to branched oligosaccharide structures which influence many 'sticky' properties on the secreted and membrane glycoproteins. Recently, however, *O*-glycosylation has also been shown to occur in the nucleus and cytoplasm of cells [26, 27]: this is characterized by the attachment of a monosaccharide, *N*-acetylglucosamine, to a serine or threonine residue.

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

O-glycosylation of secreted and membrane bound proteins is a post-translational event, taking place in the *cis*-Golgi compartment [28] after *N*-glycosylation and folding of the protein [29]. *O*-glycans are built up in a stepwise fashion with sugars added one at a time to a growing branch (except for cytoplasmic glycosylation where only a monosaccharide is attached).

There is no acceptor motif defined for *O*-linked glycosylation. The only common characteristic among most *O*-glycosylation sites is that they occur on serine and threonine residues in close proximity to proline residues, and that the acceptor site is usually in a beta-conformation.

1.3.3 Mammalian mucin type (*O*-GalNAc) glycosylation

The best known form of *O*-glycosylation in mammals is the addition of GalNAc linked to serine or threonine residues of secreted and cell surface proteins, and further addition of Gal, GalNAc or GlcNAc moieties [15]. Also known as mucin type glycosylation [79, 80], this reaction is catalyzed by a family of UDP-*N*-acetylgalactosamine: polypeptide *N*-acetylgalactosaminyltransferases (GalNAc-transferases). These enzymes are differentially expressed in different cells and organs and have different substrate specificities [30].

1.3.4 *O*-α-GlcNAc glycosylation in simple eukaryotes

O-GlcNAc residues are found in two conformations attached to the polypeptide backbone – an alpha anomeric configuration, and a beta anomeric configuration. This leads to the terminology *O*-α-GlcNAc residues (found on membrane and secreted proteins) and *O*-β-GlcNAc (found on cytoplasmic and nuclear proteins). The process, especially for *O*-β-GlcNAc attachment, is sometimes referred to as *O*-GlcNAcylation. A predictor for *O*-α-GlcNAc sites has been developed for *Dictyostelium discoideum*.

1.3.5 *O*-β-GlcNAc glycosylation on cytoplasmic/nuclear proteins

In 1984 [31] it was found that *O*-glycosylation is not only restricted to proteins which enter the ER cotranslationally, but it is a modification that also occurs on nuclear and cytoplasmic proteins. Nuclear and cytoplasmic glycoproteins are modified on multiple sites with a single *N*-acetylglucosamine residue (*O*-β-GlcNAc) [32]. In contrast to *O*-glycosylation on membrane/secreted protein, the attachment in cytoplasmic/nuclear proteins is *via* a beta anomeric linkage to produce *O*-β-GlcNAc sites. A high concentration of *O*-GlcNAcylated proteins is found in the nuclear pore complex [33]. *O*-GlcNAcylation has been described as a dynamic process with turn over rates much higher than the protein backbones to which it is attached [34, 35].

As many *O*-linked GlcNAc containing proteins examined to date are also phosphoproteins, and in some instances Ser(Thr)-*O*-GlcNAc and Ser(Thr)-*O*-phosphate appear to reciprocally occupy the same hydroxyl groups [36, 37], it has been proposed that *O*-GlcNAcylation is a regulatory protein modification [26, 38, 39].

Acceptor sites for *O*-GlcNAcylation are serines and threonines close to proline and usually in a beta conformation. The sequence context does not display a consensus, but exhibits possible correlations between sequence positions close to the acceptor site. A combined prediction approach for these yin-yang sites (positions capable of both being phosphorylated and glycosylated), where the *O*-β-GlcNAc glycosylation prediction is combined with a phosphorylation site prediction [40] has been developed for this case (Gupta, R., manuscript in preparation).

1.4 Phosphorylation

Protein phosphorylation is the primary means of switching the activity of a cellular protein rapidly from one state to another. Thus, protein phosphorylation is considered as being a key event in many signal transduction pathways of biological systems. Phosphorylation of substrate sites at serine, threonine or tyrosine residues is performed by members of the protein kinase family, the second largest family in the human genome.

Reversible protein phosphorylation is a fundamental regulatory cellular mechanism. Biochemically, this includes a transfer of a phosphate moiety from adenosine triphosphate (ATP) to the acceptor residue, thereby generating adenosine diphosphate (ADP). It is a post-translational event which normally occurs in either the cytosol or the nucleus of the cell. Protein kinases catalyze the phosphorylation events that are essential for the regulation of cellular processes like metabolism, proliferation, differentiation and apoptosis [41–46]. This very large family of enzymes share homologous catalytic domains and the mechanism of substrate recognition may be similar despite large variation in sequence. Crystallization studies indicate that a region, between seven and twelve residues in size, surrounding the acceptor residue contacts the kinase active site [47].

The specificity of protein kinases is dominated by acidic, basic or hydrophobic residues adjacent to the phosphorylated residue, but the large variation makes it difficult to manually inspect protein sequences and predict the position of biologically active sites. A wide range of algorithms have been used to implement prediction strategies. These range from simple motif searches (regular expressions), to more complex methods like neural networks where sequence correlations can be taken into account, when discriminating between potential phosphorylation sites and those which appear never to be modified.

Sequence motifs for specific kinases have, for example, been used by the PROSITE database [6]. The drawback of these patterns has been the fact that they often were based on very limited data. Thus, the sensitivity of using these patterns is quite low, as documented earlier [40]. Since determinants of phosphorylation sites probably are no longer than about ten residues, most local sequence alignment tools, such as BLAST and FASTA, will not be useful for detecting phosphorylation sites due to a large number of irrelevant hits in the protein databases, even to nonphosphorylated proteins. In contrast, many machine learning techniques are capable of classifying even highly complex and nonlinear biological sequence patterns, where correlations between positions are important. Artificial neural networks (ANN) is one such technique that has been extensively used in biological sequence analysis [48, 49], and also for phosphorylation site prediction [40].

For the human genome and the known role of phosphorylation in many human diseases, it is clear that there is a need for novel methods capable of characterizing the human phosphoproteome. Experimentally, the identification of phosphoproteins and the determination of individual phosphorylation sites occurring on phosphoproteins *in vivo* is a difficult and time consuming step. It is also clear that experimental methods are not always optimized for analyzing phosphorylated residues in peptide fragments. For example, traditional experimental MS methods have been shown to disfavor the identification of phosphate-modified residues, leading to underestimation of the extent of phosphorylation present *in vivo* [50].

Estimates of phosphorylation in the eukaryotic cell range from 30% of all proteins [51] to the "majority of human proteins may be phosphorylated at multiple sites (total $>$ 100 000 sites)" [52]. Consequently, faster and more efficient screening methods for tracing phosphorylation sites in protein sequences are needed. The computational approaches that have been constructed for prediction of kinase-specific phosphorylation sites in novel proteins differ in relation to the type of data used to construct them. Some methods are based on experimentally verified phosphorylation sites as reported in the literature and in curated databases such as the Swiss-Prot database. The NetPhos method [40] is one such example (see Section 3 for the kinase-specific version of this method, NetPhosK). Other approaches use data from *in vitro* experiments on incubations of a randomized peptide library with a given kinase. The modified peptides are separated and sequenced and a profile of the amino acid preference of the given kinase at each position relative to the acceptor site can be tabulated. A prediction method, Scansite, using peptide library experimental results has been implemented [53].

2 PTM resources – data repositories and data-driven prediction tools

2.1 Data curation and the sequence space

The key repository for PTMs has been the Swiss-Prot database [54], but other PTM specific databases have also emerged, including O-GlycBase [55], PhosphoBase [56], ELM [57] and PhosphoSite (http://www.phosphosite.org/). These resources rely heavily on human annotator specialists who plough through large amounts of scientific literature and evaluate experimental evidence for PTMs relating to a particular protein sequence. Often a semantic analysis is required to be able to correctly identify what the authors conclude. For example, the following sentence would make a totally opposite conclusion if the first part (before "PKA") was left out: "Based on these mutation studies, it is doubtful that the actions of PKA led to phosphorylation of Serine-123 in protein X".

For methods which are based on experimental data, the prediction accuracy is strongly limited by the amount and redundancy of the underlying data. High quality annotated data is key for the development of PTM classifiers, but unfortunately most of the data in the databases is annotated based on similarity, and not on first hand experimental evidence. The sequence space of possible acceptor site contexts, say for five amino acids up- and downstream of the acceptor residue, is very large, 20^{10}, compared to the amount of experimental data available, which typically for a given kinase is less than 500 examples. The amount of data required for obtaining a high quality prediction obviously depends on the diversity of the acceptor motifs for a given kinase, and it is difficult to make a reliable estimate of the "sufficient" amount of data. Another related issue is how well a particular algorithm is able to construct a model of the sequence space from a limited set of known sites. Neural networks have been quite successful for this task, and one contributing factor is that they are quite good at suppressing "out-

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

liers", highly atypical examples, which, if not handled correctly, may lead to severe overprediction and many false positives.

Our own experiences with PTM databases includes the early integration of the glycosylation site database O-GlycBase into the Sequence Retrieval System (SRS). SRS [58] was an early attempt to associate biological databases which mainly existed as large flat files (*i.e.* not in a relational database format). One of the very early parsers built into SRS was that of O-GlycBase (personal communication). This was perhaps due to the free and easy availability of the entire database as a plain text file or indeed due to the challenge posed in parsing plain text, curated mainly by a molecular biology specialist.

O-GlycBase [55] is a database of *O*-glycosylated proteins where the *O*-glycosylation sites have been experimentally mapped. This database was initially curated in concert with the construction of the NetOGlyc predictor [59, 60], which relied on a substantial training set of sequence windows from glycosylated and nonglycosylated serines and threonines. The aim of this database was to collect site-mapped data with adequate reference so the data could be authenticated and glycosylation prediction reproduced if needed, but more importantly as a resource for other biologists.

A valuable source of data was the Swiss-Prot database [54]. Swiss-Prot is an excellent resource for post-translational modifications and since 2002, it categorises glycosylation sites into the different types of attachments/linkages. Before inclusion in O-GlycBase, each Swiss-Prot entry with confirmed glycosylation sites is cross-checked against the original references. We also mined Medline abstracts for data which Swiss-Prot did not contain, and did communicate with the authors in ambiguous cases. Only proteins with experimentally verified glycosylation sites qualify for inclusion in O-GlycBase. Data found from Medline, but not in Swiss-Prot, was usually communicated to Swiss-Prot for inclusion there.

The most significant problem when compiling such a database is that there is almost never any negative data that can be included. In other words, to know that a specific serine or threonine is *not* glycosylated can be extremely useful when making prediction methods. Unfortunately, such information is very rarely published. To conclusively prove that a site is negative under all conditions is impossible, but to know that it is negative even in some contexts would be useful. Another problem while working with protein sequences from the literature is that database accession numbers are very rarely mentioned. This leaves a lot of work for the database curator to track down the original sequence and figure out which organism the protein belongs to. Even when a database accession is found, the site numbering used in the literature does not always correspond to the database entry, due to for instance, signal peptides that are cleaved in the literature entry (and thus not considered while counting amino acid position numbering in a sequence).

2.2 PTM web resources

In this section we list a nonexclusive list of valuable, publicly available resources for PTM prediction and annotation (Table 1). Naturally, more global resources, like Swiss-Prot and other databases, also contain a lot of information on PTMs, but will not be mentioned in this section. The first part of Table 1 lists a number of resources dealing with many types of PTMs. Some of them (ELM and PROSITE) include search tools for scanning a query protein with a certain motif or pattern. The Human Protein Reference Database (HPRD) contains high quality annotation for a large number of human disease-related proteins. This includes information on experimentally validated PTMs for individual proteins and links to original references. One special feature of the RESID database is that it contains information on many unusual PTM types and also shows the structure of the chemical groups added.

The next part of Table 1 lists a number of resources related to protein phosphorylation. Several of the servers (Scansite, PREDIKIN, NetPhos/NetPhosK) provide prediction of phosphorylation sites using different methods. The PREDIKIN method has the unique feature of predicting potential acceptor sites based on the primary sequence of the protein kinase catalytic domain. Scansite and NetPhos/NetPhosK are based on data sets of observed acceptor sites, either by peptide library scans or extracted from the literature, respectively. Resources that mainly deal with information about individual phosphorylation sites include the curated databases of PhosphoSite and PhosphoRase.

Early work in the glycosylation field included prediction of *O*-glycosylation site specificity using frequency based methods [61, 62]. For example, Elhammer's group [63, 64] proposed a prediction model for *O*-glycosylation based on the occurrence of amino acid residues positioned at positions ±4 relative to identified *O*-glycosylation sites.

Much of this work inspired later developments in this field and many of the resources mentioned in Table 1. This includes the many prediction servers developed in our group: NetOGlyc, a method for *O*-GalNAc glycosylation on mammalian proteins and the YinOYang server for predictions of *O*-β-GlcNAc attachment sites in eukaryotic protein sequences. This server can also use the NetPhos

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. Publicly available PTM web resources. Databases and classification/prediction servers

Resource	Web URL	Classification method/Database	Reference	Info
General PTM-related				
ELM	http://elm.eu.org/	Consensus patterns	(57)	Predicts Eukaryotic Linear Motifs (ELMs) based on consensus patterns. Applies context-based rules and logical filters
PROSITE	http://www.expasy.org/prosite/	Consensus patterns	(6)	Curated database of consensus patterns for many types of PTMs, including phosphorylation sites. MotifScan feature allows for scanning of query sequence
HPRD	http://www.hprd.org/	Database	(74)	Human Protein Reference Database (HPRD). Highly curated database of disease-related proteins and their PTMs.
RESID	http://pir.georgetown. edu/pir-www/dbinfo/resid.html	Database	(5)	Part of the PIR protein database. Comprehensive collection of annotations and structures for PTMs
Phosphorylation				
Scansite	http://scansite.mit.edu/	Weight matrix	(53)	Based on peptide library studies. Predicts kinase-specific motifs and other types of motifs involved in signal transduction, *e.g.* SH2 domain binding
PREDIKIN	http://www.biosci.uq.edu.au/kinsub/home.htm	Expert system	(75)	The program produces a prediction of substrates for S/T protein kinases based on the primary sequence of a protein kinase catalytic domain
NetPhos	http://www.cbs.dtu.dk/services/NetPhos/	Neural network	(40)	Predicts general phosphorylation status based on sets of experimentally validated S, T and Y phosphorylation sites
NetPhosK	http://www.cbs.dtu.dk/services/NetPhosK/	Neural network	This paper	Predicts kinase-specific phosphorylation sites based on sets of experimentally validated S, T and Y phosphorylation sites
PhosphoBase	http://www.cbs.dtu.dk/databases/PhosphoBase/	Database	(56)	Curated database of validated phosphorylation sites
Phosphosite	http://www.phosphosite.org/	Database	Unpublished	Curated database of *in vivo* validated phosphorylation sites
Glycosylation				
bigPI	http://mendel.imp. univie.ac.at/gpi/gpi_server.html	Weight matrix	(17)	Predicts GPI-modification sites using a composite prediction function including a weight matrix and physical model
GlycoMod	http://www.expasy.org/tools/glycomod/	Look-up table	(76)	Predicts glycan structure from its experimentally determined mass
NetOGlyc	http://www.cbs.dtu.dk/services/NetOGlyc/	Neural network	(60)	Predicts mucin type CalNAc *O*-glycosylation sites in mammalian proteins
NetNGlyc	http://www.cbs.dtu.dk/services/NetNGlyc/	Neural network	(Gupta, R el al., 2003, in prep.)	Predicts *N*-glycosylation sites in human proteins by examining the sequence context of Asn-Xaa-Ser/Thr sequons
DictyOGlyc	http://www.cbs.dtu.dk/services/DictyOGlyc/	Neural network	(77)	Predicts GlcNAc *O*-glycosylation sites in *Dictyostelium discoideum* proteins
YinOYang	http://www.cbs.dtu.dk/services/YinOYang/	Neural network	Unpublished	Predicts *O*-β-GlcNAc attachment sites in eukaryotic protein sequences.
O-GlycBase	http://www.cbs.dtu.dk/databases/OGLYCBASE/	Database	(55)	A curated database of *O*- and *C*-glycosylated proteins
Sulfation				
Sulfinator	http://www.expasy.org/tools/sulfinator/	HMM	(78)	Predicts tyrosine sulfation sites using a combination of HMM models

server, to mark possible phosphorylated sites and hence identify the so-called "Yin-Yang" sites. These are serine or threonine sites which may be reciprocally modified by transfer of either *O*-β-GlcNAc or phosphate moieties.

Similarly, the DictyOGlyc server produces predictions for *O*-α-GlcNAc glycosylation in *Dictyostelium discoideum* proteins and other simple eukaryotes and the NetNGlyc server predicts *N*-glycosylation sites in human proteins using artificial neural networks (ANN) that examine the sequence context of Asn-Xaa-Ser/Thr sequons. All of the glycosylation data used for training of the *O*-glycosylation prediction methods may be accessed at the O-GlycBase database.

Other useful tools in the glycosylation PTM field includes a predictor for GPI-anchors (bigPI) and the GlycoMod tool for prediction of glycan structures from mass data (see Table 1). In the last part of the table, we mention a useful tool for the prediction of tyrosine sulfation sites, a modification which is important for many hormone-receptor interactions. This method, Sulfinator, represents, in contrast to many of the other methods mentioned, one that is based on a hidden Markov model (HMM).

2.3 Data-driven prediction methods

Since the appearance of the first DNA sequence, scientists have sought for patterns in biological sequences to explain and predict molecular biological phenomena. Early observations of simple conserved sequence patterns led to the definition of consensus patterns. These patterns were normaly quite specific but also inflexible, not allowing for sequence substitutions. The next step up in complexity included the use of weight matrices, where each position in an alignment of a known set of sequence patterns, is weighted by the occurrence of a given sequence symbol in that position. This allowed for much more diverse patterns and provided the opportunity for a scoring scheme to rank potential hits. Going even further in complexity, machine learning approaches were introduced into the field of sequence analysis. These included the hidden Markov models (HMMs) and artificial neural networks (ANNs), which allowed for the classification of complex motifs containing positional correlations. An example of a correlation within a functional site might be a potential serine phosphorylation acceptor site, which requires an arginine at either position -2 or at position -3 but not at both. Such cases cannot be handled correctly by weight matrix methods.

Although the more sophisticated methods normally are better suited at classifying highly complex sequence patterns, this improvement comes at a price: it becomes much more difficult to "decode" the decisions behind the classification. While it is quite easy to infer from a weight matrix the most important determinants for a functional site, this becomes increasingly difficult when moving on to an HMM or neural network.

2.3.1 Neural networks

Artificial neural networks (ANNs) are capable of classifying highly complex and nonlinear biological sequence patterns, where correlations between positions are important [65]. Not only does the network recognize the patterns seen during training, but it also retains the ability to generalize and recognize similar, though not identical patterns. Artificial neural network algorithms have been extensively used in biological sequence analysis [48, 49].

The basic idea with ANNs is to use a network of neurons, each node (or neuron) having multiple inputs and a single output based on the weights (or strengths) associated with the various inputs. Neurons are organized in layers, each neuron typically connected to every neuron in the next layer. The connections are weighted. By presenting sequence windows during training, exhibiting particular features, repetitively, randomly initialized weights can be adjusted iteratively to classify the pattern correctly. With a minimum of three layers (an input layer of neurons, a hidden layer and an output layer), nonlinear features, such as correlations between sequence positions, can be learnt.

For example, training a neural network to recognize glycosylated sites involves presenting the network with sequence windows surrounding known glycosylation sites and known nonglycosylation sites. For each glycosylated site presented, the weights would be gradually adjusted to produce a network output tending to a positive prediction (most often encoded as the value 1.0). For nonglycosylated sites, the same weights would be adjusted to produce an output of zero.

2.3.2 Evaluation strategy

In order to evaluate a prediction method, two parameters are of utmost importance: *sensitivity* and *specificity*. We want to identify as many true sites as possible (sensitivity), while on the other hand, ensure that those sites predicted as positives, in fact are true (specificity). With a lot of biologically relevant problems, especially for predicting PTM sites experimental site mapped data are scarce and precious. This makes it hard to reserve a large part of data solely for network testing. Sometimes this could mean sacrificing essential diversity in the training data. One way to tackle this is to use a procedure called *cross-validation*, where the data is divided into a number of subsets (more or less equal in size). One subset is marked for testing, and the others are used for training (Fig. 1). The prediction performance is recorded, and the process is repeated for another subset as test set (while the first test set is now part of the training data). The process is iterated until all shuffling is complete, *i.e.* all subsets have been tested exactly once. The performances, recorded for each test set, are compiled and presented as the cross-validated performance of the neural network.

In the two category problem of predicting whether a particular site is modified or not, incorrect predictions are as important as the correct predictions for both categories. Thus, four quantities are defined for test set performance: 'True positives' (*Tp*) – experimentally verified modified sites that are also predicted to be modified; 'True negatives' (*Tn*) – experimentally verified unmodified sites that are also predicted to be unmodified; 'False positives' (*Fp*) – experimentally verified unmodified sites that are predicted (incorrectly) to be modified; and 'False negatives' (*Fn*) – Experimentally verified modified sites that are predicted (incorrectly) to be unmodified.

The sensitivity (S_n) of a method is defined as the proportion of positive sites that the method can correctly identify. Specificity (S_p^{med}) has been defined in medical texts as the proportion of negative sites correctly identified. However, another measure of specificity widespread in statistical (S_p^{PPV}) texts, is also known as the Positive Predictive Value (PPV) of the method. A similar measure for negative sites, is the Negative Predictive Value (NPV). These measures are defined as:

$$S_n = \frac{Tp}{Tp + Fn} \qquad S_p^{med} = \frac{Tn}{Tn + Fp} \tag{1}$$

$$PPV = \frac{Tp}{Tp + Fp} \qquad NPV = \frac{Tn}{Tn + Fn} \tag{2}$$

As an unbiased overall performance estimator, Matthews' coefficient of correlation CC [66] is widely used:

$$\mathrm{CC} = \frac{TpTn - FpFn}{\sqrt{(Tn + Fn)(Tn + Fp)(Tp + Fn)(Tp + Fp)}} \tag{3}$$

The correlation coefficient balances positive predictions equally with negative predictions. For example, a method which predicts every residue to be positive, is 100% correct in detecting positive sites but would practically be of no use and the correlation coefficient would reflect this (in this case it would be zero). For a perfect prediction, the correlation coefficient, CC would equal unity, while for a random prediction, it would be zero. In the rare event of a precisely wrong prediction, CC would be −1.

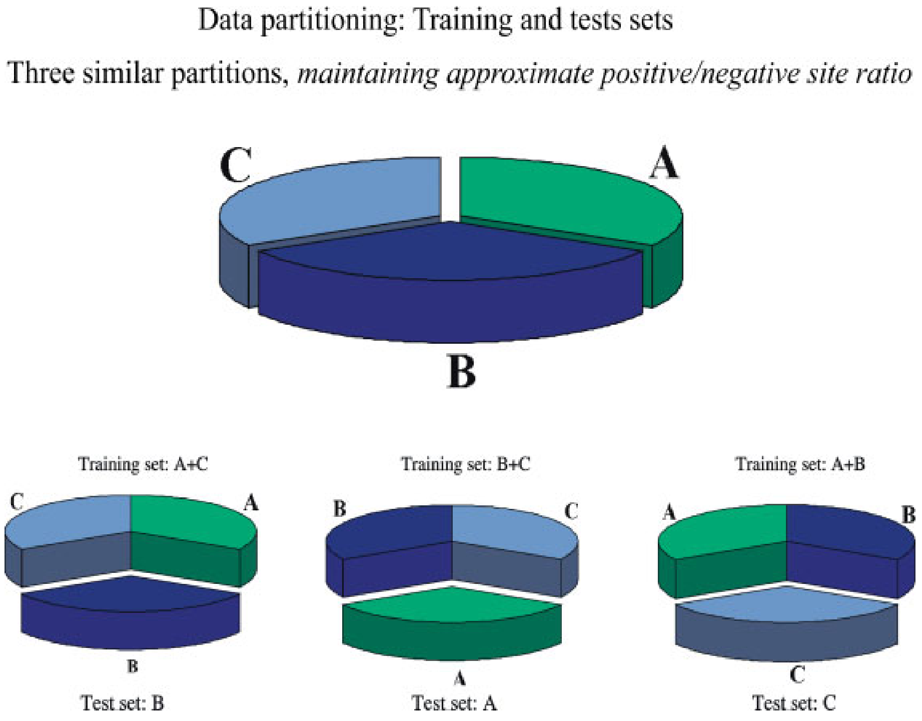

Data partitioning: Training and tests sets
Three similar partitions, *maintaining approximate positive/negative site ratio*
C
A
B
Training set: A+C
C
A
B
Test set: B
Training set: B+C
B
C
A
Test set: A
Training set: A+B
A
B
C
Test set: C

Figure 1. A cross-validation strategy for network evaluation. The data, in this example, is divided into three sets with low inter-set sequence similarity. Networks are always evaluated on a test set independent of the training set. All three combinations of using a test set (as illustrated) are employed, and performance summed.

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

2.3.3 Comparison of prediction methods – the need for golden standards

Evaluation of classifications is largely dependent on high quality data sets of both positive and negative examples of the modification being classified. In computational gene prediction, highly curated data sets of eukaryotic exon structures are available and accepted as a golden standard by most researchers in this field [67]. Unfortunately, similar golden standard data sets are not yet available for PTM predictions.

A key problem is of course that any data set quickly becomes outdated as more data accumulate in the public domain. It is therefore also common to agree on standards for redundancy reduction instead of agreeing on a static golden standard data set. Such intra-set similarity thresholds have been put in common use within the protein structure prediction area [68] and also for signal peptide prediction [69], but so far in the area of PTM prediction the principles for data set preparation have not been uniform.

2.4 Improving specificity

Many PTM-classifiers tend to overpredict. In particular, if methods are trained to find short linear motifs of around 10–15 residues, similar motifs will occur at random in nonmodified proteins. These seemingly false predictions might actually be verifiable by *in vitro* experiments involving the modifying enzyme and the motif peptide. However, in the *in vivo* situation, the motif may be inaccessible either structurally in the protein or the protein may be in a compartment not accessible to the modifying enzyme. Overprediction leads to lower prediction specificity, *i.e.*, a lower fraction of the predicted sites that are true. To overcome this, several strategies can be pursued. These involve contextual filters, evolutionary conservation of PTM sites and visualization.

Contextual filters are best illustrated by the ELM project which aims at predicting Eukaryotic Linear Motifs [57]. The predicted sites are filtered if they occur in regions believed to be devoid of PTMs, such as signal peptide regions and globular domains. Since some functional sites can occur in loops of globular domains, some true predictions are removed by this filter, and the users are encouraged to inspect the unfiltered results as well.

Other filter strategies include the use of taxonomic information, based on the notion that if predicted PTMs are conserved in proteins from a related organism they are more likely to be physiologically relevant. Also, the application of a cellular compartment filter may be relevant if known for the query protein. Since many functional sites are restricted to one or several cellular compartments, this knowledge may be used to reduce the number of false predictions. For example, phosphorylation very rarely takes place on secreted or extracellular parts of proteins (exceptions include some types of milk proteins).

Visualization of predicted PTMs is also being developed at many PTM resources including the ELM and the Scansite servers. In particular, for single protein analysis, the visualization of several predicted PTMs and other protein features is extremely important for the end-user to evaluate the predictions provided.

2.4.1 Evolutionary stable sites: Improving specificity

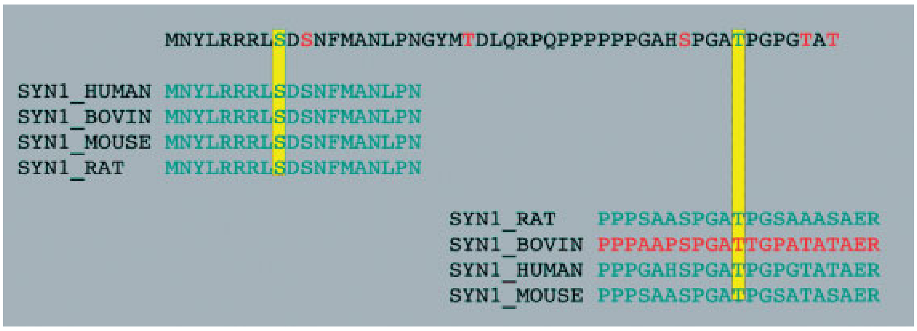
One important strategy for evaluating the validity of predicted PTM sites has become more feasible as more and more complete proteomes become available. The idea is to compare predicted sites in the query protein to homologous sites in closely related species. If the acceptor residue is conserved and also predicted as being the acceptor of the same PTM type, the confidence of the prediction is increased. In particular, if the site is not conserved, the validity of the prediction should be questioned.

In our experience, these observations seem to hold true for phosphorylation sites, but less so for glycosylation sites. This may be due to the fact that glycosylation is more of a bulk effect, where the exact acceptor site location is less important (Karin Julenius *et al.*, unpublished). We have introduced the concept of Evolutionary stable sites (ESS) in our most recent predictor for kinase-specific phosphorylation sites, NetPhosK, which is described in detail in Section 3. A simplified step-by-step explanation of the ESS procedure following the initial neural network prediction is outlined in Fig. 2.

2.5 PTM landscapes

Visualization often aids the human mind in understanding complex phenomena. We have for example developed means for increasing the understanding of PTM dynamics in processes which have temporal aspects, such as the cell cycle or disease progression, *e.g.*, cancer [70]. To obtain an overview of the potential sites of modification in a given protein, a graphical representation may be helpful. We have introduced the concept of “kinase landscapes”, which may easily be transferred to other types of PTMs. Consider the kinase specificity prediction along an amino acid sequence only at serine, threonine or tyrosine (STY) sites: if the prediction for different kinases are represented by different letters of heights proportional to

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim



```
          MNYLRRRLSDSNFMANLPNGYMTDLQRPQPPPPPPGAHSPGATPGPGTAT

SYN1_HUMAN MNYLRRRLSDSNFMANLPN
SYN1_BOVIN MNYLRRRLSDSNFMANLPN
SYN1_MOUSE MNYLRRRLSDSNFMANLPN
SYN1_RAT   MNYLRRRLSDSNFMANLPN

                    SYN1_RAT   PPPSAASPGATPGSAAASAER
                    SYN1_BOVIN PPPAAPSPGATTGPATATAER
                    SYN1_HUMAN PPPGAHSPGATPGPGTATAER
                    SYN1_MOUSE PPPSAASPGATPGSATASAER
```

Figure 2. A simplified step-by-step explanation of the ESS procedure. (1) Predict kinase-specific acceptor sites for full length protein; (2) for each predicted site P → get 21mer = P + 10 residues *N*- and *C*-terminally; (3) compare 21mer with nonredundant protein database; (4) extract near-identical matches; (5) collect homologues from higher eukaryotes; (6) for all homologues run the kinase-specific predictor on the potential acceptor site; and (7) tabulate the number of homologous sites in related species that receive same type of kinase-specific prediction.

the prediction scores and plotted along the sequence position, one can recognize a kinase landscape, a potential region of the sliding kinase motif.

Prediction of kinase specificity along an amino acid sequence is normally done at serine, threonine or tyrosine residues. However, if the kinase specificity is also predicted at non-STY sites, one can assess whether the context fits a good acceptor site although the acceptor residue does not conform. We propose that it may be possible to identify "hot spots" in protein sequences – regions, in which mutations or variations, *e.g.*, coding single nucleotide polymorphisms (SNPs), may change a given nonacceptor residue into an acceptor residue (S, T or Y). This change may introduce novel and potentially disruptive phosphorylation sites.

As an example, we examined the sequence of the human p53 protein, which plays a major role in regulating the cell cycle. More than 10 000 tumour-associated mutations in p53 have been discovered, in organisms ranging from humans to clams [71]. Using the novel kinase-specific phosphorylation site predictor, NetPhosK, described below, we investigated the consequences of mutating non-STY residues in p53 into either serine or threonine residues. The potential novel phosphorylation sites, which would arise due to a substitution were then predicted.

A proline residue is normally found at position 151 in the human p53 protein (Swiss-Prot identifier p53_HUMAN). In some observed cancer types, substitutions of this Pro-151 to either serine or threonine has been observed. Based on this, we analyzed whether the observed substitutions could lead to the generation of novel phosphorylation sites.

In Fig. 3 (middle panel) we show a fragment of the wild-type p53 protein sequence including a proline at position 151. A cell division cycle 2 kinase (cdc2) phosphorylation

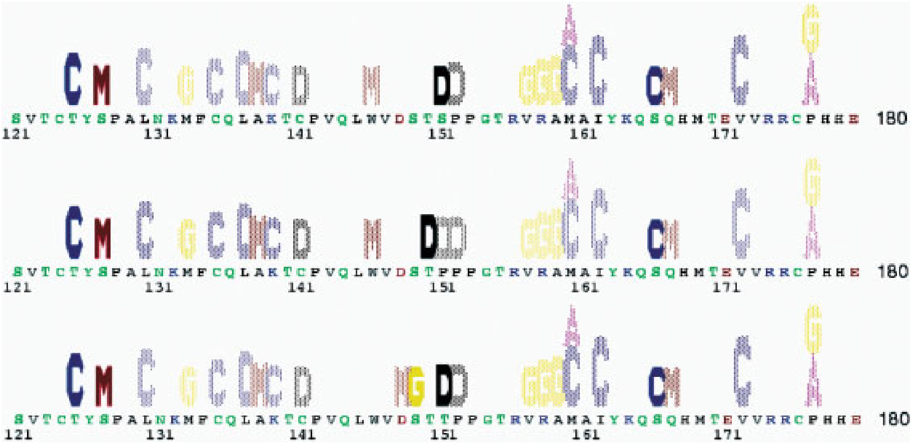


Figure 3. Kinase landscape of partial wild-type and mutant human p53 sequences. Mutant sequences display different kinase landscapes in the vicinity of the substituted proline-151. Upper panel p53 Pro-151→Ser-151 mutant sequence; Middle panel p53 wild-type sequence; Lower panel p53 Pro-151→Thr-151 mutant sequence. Kinase predictions with scores above 0.5 are shown by uppercase letters proportional to the score. Color, coding and kinase symbols: PKA (A) = magenta, PKC (C) = blue, PKG (G) = yellow, CKII (K) = green, cdc2 (D) = black, CaM-II (M) = red. Kinase predictions for non-STY sites are drawn in a stippled font.

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

site is predicted for Thr-150 as indicated graphically in Fig. 3. Now, substituting the wild-type Pro-151 for a serine residue (Fig. 3, upper panel) abolishes the predicted cdc2 phosphorylation site (which requires a proline at P+1) at Thr-150 and introduces a novel, potential cdc2 phosphorylation site at Ser-151 (denoted by uppercase D in Fig. 3). Substituting Pro-151 for a threonine residue (Fig. 3, lower panel) also changes the predicted cdc2 site from Thr-150 to Thr-151, but at the same time introduces a predicted site for cyclic GMP-dependent protein kinase (PKG) at Ser-149 (denoted by uppercase G in Fig. 3).

These observed substitutions may lead to a disrupted "phosphorylation status" of p53 and thereby lead to malignant changes. Naturally, the specific hypothesis presented here is highly speculative, but might lead to the generation of novel PTM-associated hypotheses about the (abnormal) function of a given protein.

3 NetPhosK – a kinase-specific phosphorylation site predictor: A case story

Our first attempt at developing a predictor for phosphorylation sites, NetPhos, was somewhat hindered by the fact that much of the annotated data did not include information about the nature of the active kinase [40]. Since we did not have access to laboratory facilities for generating kinase-specific data, we had to settle for general predictors for each of the phosphoacceptor residues, serine, threonine and tyrosine. We were aware that this represented a potential drawback to our method since many users would like to know the nature of the most likely kinase to interact with their protein of interest. On the other hand, for prediction-aided interpretation of MS data it is the potential site which is of interest and to a smaller degree the associated kinase.

As more kinase-specific data became available in the literature and in databases, we decided to train individual kinase-specific predictors for the few kinases, where many different acceptor sites had been reported. The NetPhosK predictor represents a kinase-specific extension to the general NetPhos method.

3.1 Data set and protein kinases analyzed

To capture as much information as possible about the substrate specificity of a given kinase, it is essential that the data set of known sites is as large as possible. We selected six serine/threonine kinases for which we had the largest number of known acceptor sites annotated in the current version of Phosphobase, a database of phosphorylation sites [56]. This included also a scan of novel sites from the scientific literature and from a recent version of the Swiss-Prot database.

The selected kinases were PKA (cyclic AMP dependent protein kinase; cAPK), PKC (protein kinase C), PKG (cyclic GMP-dependent protein kinase), cdc2 (cell division cycle 2 kinase, cdk1), CK-2 (casein kinase 2) and CaM-II (Ca^{2+}/calmodulin-dependent protein kinase). For each of the kinases between 22 and 258 different substrate sites were known (defined as being unique within a nine-residue window centered on the acceptor residue). Sequence logos were generated based on the set of known, unique acceptor sites for each kinase (see logo figures embedded in Table 2).

In general, the sequence logos conform to the accepted consensus sequence motifs, but also show that some sequence variations around the acceptor sites do occur. This is also reflected by the fact that simple consensus patterns do not match all of the known acceptor sites for a given kinase. Matches range from 37% for cdc2 to 82% for PKA (Table 2). These logos also indicate why kinase-unspecific phosphorylation site prediction is a nonlinear prediction problem. The most significant features in the logo are related to charge, either positive or negative in the acceptor residue context. Typically, an acceptor site has positive or negative charge in the context, but not both negative and positive charged residues simultaneously.

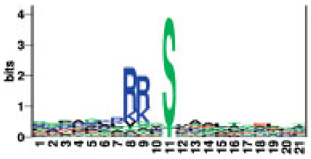
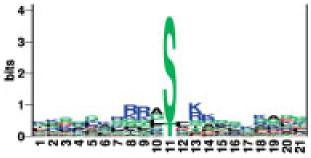
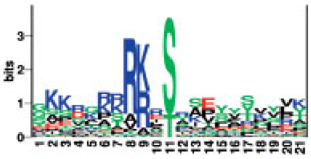
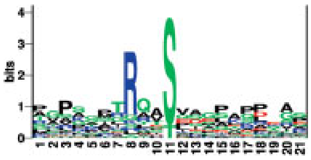
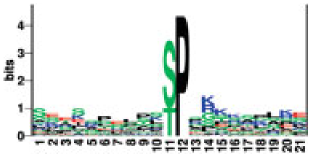
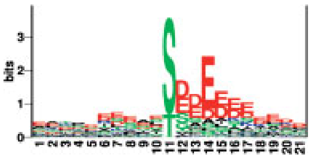
3.2 Neural network training and predictive performance of kinase-specific prediction

Neural networks were trained according to the procedure described earlier [72, 40]. For each of the six kinases analyzed an ensemble of either three or four neural networks were optimized for performance on an independent test data set. The performance of the network ensembles is best evaluated by calculating a correlation coefficient based on the cross-validated test procedure described above in Section 2.3.2. The values obtained for each kinase type are shown in Table 2 and fall in the range of 0.22–0.61.

When evaluating the performance of prediction methods and also selecting data for classifiers such as neural networks, one is often faced with the problem of defining a set of true negative examples of the modification being evaluated. In the case of *e.g.*, signal peptide prediction, the issue is relatively simple: either the preprotein is cleaved at position X or at position Y. The cleavage at one site precludes the cleavage at the other. This is not so in the case of phosphorylation sites, which are non-

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2. Kinase types included in current NetPhosK. Sequence logos, PROSITE/Consensus patterns and neural network performance values are shown. PROSITE format: *e.g.* $(R/K)_{2-3}$ xS/Tx means 2–3 R or K residues followed by any residue followed by S or T. Max.CC: best Matthews correlation coefficient achieved on cross-validated test set. Pos: total number of positive sites in data set. TP: Percentage true positives predicted at best performance. Neg: total number of negative sites in data net. TN: Percentage true negatives predicted at best performance. Win: input window size at best performance. Percentage of Pos (positive sites) matching PROSITE pattern.

Kinase		Logo	PROSITE	Max.CC	Pos	TP	Neg	TN	Win	Match
PKA	cyclic AMP dependent protein kinase (cAPK)		$Rx_{1-2}S/Tx$	0.61	258	79	1309	89	17	82
PKC	protein kinase C		xS/TxR/K	0.48	193	74	1374	86	15	62
PKG	cyclic GMP-dependent protein kinase		$(R/K)_{2-3}xS/Tx$	0.46	31	71	1536	97	15	71
CaM-II	Ca2+/cal-modulin-dependent protein kinase		RxxS/Tx	0.22	26	50	1541	94	17	73
cdc2	cyclin dependent kinase (cdk1)		xS/TPxR/K	0.36	22	55	1539	98	15	37
CKII	Casein Kinase 2 (CK2)		xS/TxxD/R	0.52	85	81	1482	92	17	75

exclusive. Phosphorylation at position X does not preclude that this protein is also phosphorylated at position Y.

Most often, a given protein has been mapped with regard to phosphorylation sites at only selected sites and not at all possible acceptor sites within the protein. Also, in general, only positive identifications of phosphorylation sites are reported in the scientific literature. It is very rare to report that position X in protein P is *not* phosphorylated as discussed above. Thus, it is nontrivial to define a true set of negative examples in the case of phosphorylation by a given kinase. In the case of our neural network training sessions described in this and previous work, we adopted an approach, where we eliminated conflicting sites from our data set. Nonconflicting sites, not annotated as phosphorylation sites are then defined as negative sites.

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

3.3 Comparison of predictive performance

One of the popular methods currently available for predicting kinase specific phosphorylation sites is the Scansite method [53], which is based on peptide library data. This method includes acceptor site predictions, not only for a range of protein kinases, but also for a number of phosphobinding domains (*e.g.*, the SH2 domain). In order to compare our method to Scansite, we have focused on the prediction performance of a method trained on PKA phosphorylation sites.

In order to do a fair and unbiased comparison of Scansite and NetPhosK, it was important to define a reliable set of positive and negative phosphorylation sites. The positive set used in training our method was based on experimental evidence and may therefore be regarded as representing the *in vivo* or *in vitro* situation. In contrast, the negative set of phosphorylation sites used to develop our method was not based exclusively on experimental verification as described above and therefore might be considered biased towards the NetPhosK method. Some of the negative sites might in fact be false negatives, *i.e.*, true acceptor sites that have not been tested experimentally.

To define an unbiased and independent set of negative phosphorylation sites we adopted a strategy in which we selected sites that are truly buried within protein structures. Presumably, these sites are not sterically available to the potential kinase and are thus never phosphorylated. We chose to extract from the Protein Data Bank (http://www.rcsb.org/pdb/) a set of serine and threonine sites known to be truly buried in a set of protein sequences for which the protein structure had been determined. The calculated exposure of these sites was below 20 square angstroms (calculated using a simulated water probe of 1.5 angstrom diameter) indicating low surface accessibility. Since PKA has a preference for basic residues, in particular arginine residues, on the *N*-terminal side of the acceptor residue, we extracted the top 220 sites containing the highest number of arginine or lysine residues in a region from position −6 to −1 relative to the acceptor residue. These sites represent the "hardest" cases from the set of buried potential acceptor sites. Predictions for PKA phosphorylation sites were performed using both the Scansite method at different stringency levels and the NetPhosK method at different threshold values on the positive and negative sets described above. Results are summarized in Table 3.

It is evident from the results that none of the methods are able to identify all of the positive sites at any of the thresholds examined. It is also evident, that lowering the threshold increases the number of true positives predicted, but also increases the number of false positive predictions in the negative set. For each threshold, the Matthews correlation [66] was calculated. The maximum values, indicating best overall prediction performance, were obtained for 'medium stringency' for the Scansite method (value = 0.28) and at threshold 0.6 for the NetPhosK method (value = 0.59), see Table 3. Generally, the lower performance of Scansite on this particular data set seems to be caused by a higher number of false positive predictions (negative sites predicted as positive) at a given rate of true positives. One contributing factor may be that the algorithm behind Scansite is not able to take sequence correlations into account. This means that "evidence" is summed up position by position, with no possibility to subtract "evidence" if the presence of a particular amino acid pair has negative influence on the substrate fitness.

Table 3. Comparison of Scansite *vs.* NetPhosK PKA-specific ESS prediction of 220 negative (buried and basic) and 142 positive (known PKA) sites. The number of true and false positive predictions and the calculated Matthews correlation coefficient (CC) is shown for a range of confidence levels (high, medium and low for Scansite) and threshold values (0.5–0.8 for NetPhosK). Best overall performance values are shown in italics.

	False Positives	True Positives	Matthews CC
Scansite			
high	14 (6%)	20 (14%)	0.13
medium	*34* (16%)	*58* (41%)	*0.28*
low	81 (37%)	90 (63%)	0.26
NetPhosK			
0.8	21 (10%)	50 (35%)	0.32
0.75	35 (16%)	73 (51%)	0.38
0.7	38 (17%)	95 (67%)	0.50
0.6	*52* (24%)	*119* (84%)	*0.59*
0.5	75 (34%)	132 (93%)	0.58

As shown by the comparative study of NetPhosK-PKA *vs.* Scansite, we show that for a given specificity level, NetPhosK-PKA achieves a higher sensitivity rate. In other words, for a given rate of false positives, more true positives are predicted by NetPhosK-PKA. However, blindly trusting a prediction for a given protein is not recommended. The predictions are quite likely to hold true for the *in vitro* situation where a particular kinase and predicted acceptor site peptide are incubated and assayed for phosphate transfer, but for the *in vivo* situation other factors need to be considered. These include the subcellular localization of the kinase and substrate protein

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

and also the exact location of the acceptor site within the substrate protein. These issues were also discussed in Section 2.4.

3.4 PTMs and protein function

Even several years after the publication of the human genome, the largest functional category of the predicted and known genes, is the one labelled "function unknown". Classification of these orphan proteins which have no homology to known proteins represents a gigantic experimental task. Prediction methods may aid in solving this task. For example, it has been shown that some proteins, which are related functionally, but not related at the sequence or structure levels, share some of the same PTM features. This has been incorporated into the ProtFun server which produces *ab initio* predictions of protein function from sequence. The method queries a large number of other feature prediction servers to obtain information on various post-translational and localizational aspects of the protein, which are then integrated into final predictions of the cellular role, enzyme class (if any), and selected Gene Ontology categories of the submitted sequence [3, 4].

Many of the functional, cellular role categories used to classify entire proteomes, are so broad that the correlation with structure is quite weak. It is also well known that many structurally similar proteins, even from the same superfamily, may represent widely different biochemical function. In the ProtFun method, we have attempted to predict protein function based on predicted properties of proteins, such as physicochemical properties, predicted PTMs and subcellular localization signals [3, 73]. Although predicted from sequence, they are more conserved among orthologs than paralogs given the same degree of sequence conservation. This is in contrast to three-dimensional structure, which is conserved for paralogs as well as orthologs [4]. It was also demonstrated that the sequence derived protein properties, characterize proteins of different cellular roles in ways that are conserved not only within Eukarya, but in several cases within all three domains of life: Eukarya, Archaea and Bacteria. These discoveries have been made through a cross-species analysis of the performance of the ProtFun prediction method [3] for a wide variety of organisms covering mammals, invertebrates, plants and fungi as well as crenarchaeota, euryarchaeota and eubacteria.

PTMs are therefore of interest much beyond the individual sequence, but also for understanding evolutionary pressures that go beyond maintaining protein structure. It might be speculated that for some proteins the ability to become phosphorylated is more important than to preserve its three-dimensional structure. It is not unlikely that the understanding of protein function in the coming years will involve PTMs in a much more prominent role, and in that sense balance the picture which so far has mostly been based on protein structure.

4 Conclusion

It is clear that simple pattern searches for PTM modification sites in proteins either tend to underestimate (pattern is too restricted) or overpredict (pattern is too broad). In our experience, more complex classification approaches, such as statistical or machine learning methods are better suited for these tasks. However, even the most sophisticated prediction methods are not perfect and predicted PTM sites need to be carefully evaluated by the user. We have proposed several strategies for improving the likelihood that the predicted modification makes sense in a biologically relevant situation. This includes taking the predicted subcellular localization, protein domain structure or evolutionary conservation into account. It is our goal that an integrated environment for PTM predictions should include these secondary types of evidence and warn the user of any potential conflicts, *e.g.*, that phosphorylation of a transmembrane region is unlikely.

The annotation of complete proteomes based only on predicted modifications is still premature. However, rapid developments of high-throughput mass spectrometry methods will, in the near future, generate large amounts of PTM data sets which will guide the future development and improvement of PTM predictors. In combination with the expanding knowledge of the physiological protein-protein interactions in the cell, this will, hopefully, allow us to reliably simulate, *e.g.*, signaling cascades and predict the effects on the level of gene regulation. Future drug testing may very well rely as much on *in silico* simulations as they now do on animal experiments.

5 References

[1] Bork, P., Dandekar T., Diaz-Lazcoz, Y., Eisenhaber, F. *et al.*, *J. Mol. Biol.* 1998, *283*, 707–725.

[2] Attwood, T., *Int. J. Biochem. Cell Biol.* 2000, *32*, 139–155.

[3] Jensen, L., Gupta, R., Blom, N., Devos, D. *et al.*, *J. Mol. Biol.* 2002, *319*, 1257–1265.

[4] Jensen, L., Ussery, D., Brunak, S., *Genome Res.* 2003, *13*, 2444–2449.

[5] Garavelli, J., *Nucleic Acids Res.* 2003, *31*, 499–501.

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[6] Sigrist, C., Cerutti, L., Hulo, N., Gattiker, A. *et al.*, *Brief Bioinform.* 2002, *3*, 265–274.

[7] Huttner, W., *Annu. Rev. Physiol.* 1988, *50*, 363–376.

[8] Rosenquist, G., Nicholas, H., *Protein Sci.* 1993, *2*, 215–222.

[9] Hofmann, K., Bucher, P., Falquet, L., Bairoc, A., *Nucleic Acids Res.* 1999, *27*, 215–219.

[10] Nicholas, H., Chan, S., Rosenquist, G., *Endocrine* 1999, *11*, 285–292.

[11] Kehoe, J., Bertozzi, C., *Chem. Biol.* 2000, *7*, R57–61.

[12] Vasan, S., Zhang, X., Zhang, X., Kapurniotu, A. *et al.*, *Nature* 1996, *382*, 275–278.

[13] Stanley, P., *Glycobiology* 1992, *2*, 99–107.

[14] Varki, A., *Glycobiology* 1993, *3*, 97–130.

[15] Hounsell, E. F., Davies, M. J., Renouf, D. V., *Glycoconj. J.* 1996, *13*, 19–26.

[16] Messner, P., *Glycoconj. J.* 1997, *14*, 3–11.

[17] Eisenhaber, B., Bork, P., Eisenhaber, F., *J. Mol. Biol.* 1999, *292*, 741–758.

[18] Hofsteenge, J., Muller, D., de Beer, T., Loffler, A. *et al.*, *Biochemistry* 1994, *33*, 13524–13530.

[19] Doucey, M., Hess, D., Cacan, R., Hofsteenge, J., *Mol. Biol. Cell* 1998, *9*, 291–300.

[20] Krieg, J., Hartmann, S., Vicentini, A., Glasner, W. *et al.*, *Mol. Biol. Cell* 1998, *9*, 301–309.

[21] Hartman, S., Hofsteenge, J., *J. Biol. Chem.* 2000, *275*, 28569–28574.

[22] Gavel, Y., von Heijne, G., *Protein Eng.* 1990, *3*, 433–442.

[23] Miletich, J., Broze Jr., G., *J. Biol. Chem.* 1990, *265*, 11397–11404.

[24] Kasturi, L., Eshleman, J., Wunner, W., Shakin-Eshleman, S., *J. Biol. Chem.* 1995, *270*, 14756–14761.

[25] Silberstein, S., Gilmore, R., *FASEB J.* 1996, *10*, 849–858.

[26] Hart, G. W., *Ann. Rev. Biochem.* 1997, *66*, 315–335.

[27] Snow, D. M., Hart, G. W., *Int. Rev. Cytol.* 1998, *181*, 43–74.

[28] Roth, J., Wang, Y., Eckhardt, A. E., Hill, R. L., *Proc. Natl. Acad. Sci. USA* 1994, *91*, 8935–8939.

[29] Asker, N., Baeckstrom, D., Axelsson M., Carlstedt, I., Hansson, G., *Biochem. J.* 1995, *308*, 873–880.

[30] Clausen, H., Bennett, E., *Glycobiology* 1996, *6*, 635–646.

[31] Torres, C., Hart, G., *J. Biol. Chem.* 1984, *259*, 3308–3317.

[32] Hart, G., Haltiwanger, R., Holt, G., Kelly, W., *Annu. Rev. Biochem.* 1989, *58*, 841–874.

[33] Holt, G., Snow, C., Senior, A., Haltiwanger, R. *et al.*, *J. Cell Biol.* 1987, *104*, 1157–1164.

[34] Haltiwanger, R. S., Blomberg, M. A., Hart, G. W., *J. Biol. Chem.* 1992, *267*, 9005–9013.

[35] Roquemore, E., Chevrier, M., Cotter, R., Hart, G., *Biochemistry* 1996, *35*, 3578–3586.

[36] Chou, T., Dang, C., Hart, G., *Proc. Natl. Acad. Sci. USA* 1995, *92*, 4417–4421.

[37] Chou, T., Hart, G., Dang, C., *J. Biol. Chem.* 1995, *270*, 18961–18965.

[38] Hart, G. W., Greis, K. D., Dong, L. Y., Blomberg, M. A. *et al.*, *Adv. Exp. Med. Biol.* 1995, *376*, 115–123.

[39] Haltiwanger, R., Busby, S., Grove, K., Li, S. *et al.*, *Biochem. Biophys. Res. Commun.* 1997, *231*, 237–242.

[40] Blom, N., Gammeltoft, S., Brunak, S., *J. Mol. Biol.* 1999, *294*, 1351–1362.

[41] Kolibaba, K., Druker, B., *Biochim. Biophys. Acta* 1997, *1333*, F217–F248.

[42] Hunter, T., *Philos. Trans. R. Soc. Lond. B.* 1998, *353*, 583–605.

[43] Johnson, L., Noble, M., Owen, D., *Cell* 1996, *85*, 149–158.

[44] Johnson, L., Lowe, E., Noble, M., Owen, D., *FEBS Lett.* 1998, *430*, 1–11.

[45] Pinna, L. A., Ruzzene, M., *Biochim. Biophys. Acta* 1996, *1314*, 191–225.

[46] Graves, L., Bornfeldt, K., Krebs, E., *Adv. Sec. Mess. Phos. Res.* 1997, *31*, 49–62.

[47] Songyang, Z., Blechner, S., Hoagland, N., Hoekstra, M. F. *et al.*, *Curr. Biol.* 1994, *4*, 973–982.

[48] Wu, C. H., *Comput. Chem.* 1997, *21*, 237–256.

[49] Baldi, P., Brunak, S., *Bioinformatics: The Machine Learning Approach*, 2nd edition, MIT Press, Cambridge, MA 2002.

[50] Mann, M., Ong, S., Gronborg, M., Steen, H. *et al.*, *Trends Biotechnol.* 2002, *20*, 261–268.

[51] Hubbard, M., Cohen, P., *Trends Biochem Sci.* 1993, *18*, 172–177.

[52] Zhang, H., Zha, X., Tan, Y., Hornbeck, P. *et al.*, *J. Biol. Chem.* 2002, *277*, 39379–39387.

[53] Yaffe, M., Leparc, G., Lai, J., Obata, T. *et al.*, *Nat. Biotechnol.* 2001, *19*, 348–353.

[54] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. *et al.*, *Nucleic Acids Res.* 2003, *31*, 365–370.

[55] Gupta, R., Birch, H., Rapacki, K., Brunak, S., Hansen, J., *Nucleic Acids Res.* 1999, *27*, 365–370.

[56] Kreegipuu, A., Blom, N., Brunak, S., *Nucleic Acids Res.* 1999, *27*, 237–239.

[57] Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S. *et al.*, *Nucleic Acids Res.* 2003, *31*, 3625–3630.

[58] Zdobnov, E., Lopez, R., Apweiler, R., Etzold, T., *Bioinformatics* 2002, *18*, 1149–1150.

[59] Hansen, J. E., Lund, O., Engelbrecht, J., Bohr, H. *et al.*, *Biochem. J.* 1995, *308*, 801–813.

[60] Hansen, J. E., Lund, O., Tolstrup, N., Gooley, A. A. *et al.*, *Glycoconj. J.* 1998, *15*, 115–130.

[61] Wilson, I. B. H., Gavel, Y., Heijne, G. V., *Biochem. J.* 1991, *275*, 529–534.

[62] Yoshida, M., Yokota, S., Ouchi, S., *Exp. Cell Res.* 1997, *230*, 393–398.

[63] Elhammer, A., Poorman, R., Brown, E., Maggiora, L. *et al.*, *J. Biol. Chem.* 1993, *268*, 10029–10038.

[64] Elhammer, A., Kezdy, F., Kurosaka, A., *Glycoconj. J.* 1999, *16*, 171–180.

[65] Presnell, S. R., Cohen, F. E., *Annu. Rev. Biophys. Biomol. Struct.* 1993, *22*, 283–298.

[66] Matthews, B., *Biochim. Biophys. Acta* 1975, *405*, 442–451.

[67] Guigo, R., Agarwal, P., Abril, J., Burset, M., Fickett, J., *Genome Res.* 2000, *10*, 1631–1642.

[68] Sander, C., Schneider, R., *Proteins* 1991, *9*, 56–68.

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[69] Nielsen, H., Engelbrecht, J., von Heijne, G., Brunak, S., *Proteins* 1996, *24*, 165–177.

[70] de Lichtenberg, U., Jensen, T., Jensen, L., Brunak, S., *J. Mol. Biol.* 2003, *329*, 663–674.

[71] Vogelstein, B., Lane, D., Levine, A., *Nature* 2000, *408*, 307–310.

[72] Blom, N., Hansen, J. E., Blaas, D., Brunak, S., *Protein Sci.* 1996, *5*, 2203–2216.

[73] Gupta, R., Jensen, L., Brunak, S., *Ernst Schering Res. Found. Workshop* 2002, *38*, 276–294.

[74] Peri, S., Navarro, J., Amanchy, R., Kristiansen, T. *et al.*, *Genome Res.* 2003, *13*, 2363–2371.

[75] Brinkworth, R., Breinl, R., Kobe, B., *Proc. Natl. Acad. USA* 2003, *100*, 74–79.

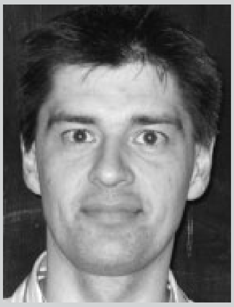
[76] Cooper, C., Gasteiger, E., Packer, N., *Proteomics* 2001, *1*, 340–349.

[77] Gupta, R., Jung, E., Gooley, A., Williams, K. *et al.*, *Glycobiology* 1999, *9*, 1009–1022.

[78] Monigatti, F., Gasteiger, E., Bairoch, A., Jung, E., *Bioinformatics* 2002, *18*, 769–770.

[79] Verma, M., Davidson, E., *Glycoconj. J.* 1994, *11*, 172–179.

[80] Carraway, K., Hull, S., *Glycobiology* 1991, *1*, 131–138.



Nikolaj Blom is an associate professor at the Center for Biological Sequence Analysis (CBS) at the Technical University of Denmark in Lyngby. He heads a group with special interest in protein post-translational modifications (PTMs) in particular the prediction of phosphorylation, glycosylation and proteolytic cleavage sites. He also works part-time for the Danish biotech company NsGene A/S coordinating the bioinformatics part of the company's gene discovery programme. He teaches Gene Finding and Gene Discovery courses at the university and is a board member of the Danish society for biochemistry and molecular biology. He has co-authored a number of articles on PTM prediction and is the co-inventor on several gene patents.

© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim