Biochimica et Biophysica Acta 1844 (2014) 316–322

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

HMMpTM: Improving transmembrane protein topology prediction using phosphorylation and glycosylation site prediction

Georgios N. Tsaousis [a], Pantelis G. Bagos [b], Stavros J. Hamodrakas [a,*]

[a] Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens 15701, Greece
[b] Department of Computer Science and Biomedical Informatics, University of Thessaly, Papasiopoulou 2–4, Lamia 35100, Greece

ARTICLE INFO

Article history:
Received 17 September 2013
Received in revised form 2 November 2013
Accepted 4 November 2013
Available online 10 November 2013

Keywords:
Transmembrane protein
Phosphorylation
Glycosylation
Topology
Prediction
Hidden Markov model

ABSTRACT

During the last two decades a large number of computational methods have been developed for predicting transmembrane protein topology. Current predictors rely on topogenic signals in the protein sequence, such as the distribution of positively charged residues in extra-membrane loops and the existence of N-terminal signals. However, phosphorylation and glycosylation are post-translational modifications (PTMs) that occur in a compartment-specific manner and therefore the presence of a phosphorylation or glycosylation site in a transmembrane protein provides topological information. We examine the combination of phosphorylation and glycosylation site prediction with transmembrane protein topology prediction. We report the development of a Hidden Markov Model based method, capable of predicting the topology of transmembrane proteins and the existence of kinase specific phosphorylation and N/O-linked glycosylation sites along the protein sequence. Our method integrates a novel feature in transmembrane protein topology prediction, which results in improved performance for topology prediction and reliable prediction of phosphorylation and glycosylation sites. The method is freely available at http://bioinformatics.biol.uoa.gr/HMMpTM.

1. Introduction

Transmembrane proteins constitute ~20 to 30% of fully sequenced proteomes and they are an important class of proteins, since they are crucial for a wide variety of cellular functions [1]. In order to understand their function we must acquire knowledge about their structure and topology in relation to the membrane. However, obtaining crystals of transmembrane proteins suitable for crystallographic studies is difficult and transmembrane proteins represent less than 2% of the structures in the Protein Data Bank [2]. Therefore, during the last two decades a large number of computational methods have been developed in order to predict the topology of transmembrane proteins [3]. By topology, we refer to the knowledge of the number and the exact localization of transmembrane segments, as well as their orientation with respect to the lipid bilayer. The first prediction methods made use of hydrophobicity scales in order to predict the location of transmembrane segments along the protein sequence [4]. Later, the positive inside rule was used for the prediction of the overall topology of a transmembrane protein by discriminating the regions facing the two sides of the membrane [5,6]. The evolution of transmembrane topology prediction methods involved the use of several algorithmic techniques including Statistical Analyses [7,8], Artificial Neural Networks (ANNs) [9,10], Hidden Markov Models (HMMs) [11–15], Support Vector Machines (SVMs) [16], Dynamic Bayesian Networks (DBNs) [17] and ensemble methods (e.g. Hidden Neural Networks, HNNs) [18,19]. Hidden Markov Models have been shown to outperform other techniques in topology prediction and are widely used [15,20,21]. In addition, there are a number of prediction methods (meta-predictors) that combine the results of several individual methods and produce a consensus prediction [22–25].

Transmembrane protein topology prediction methods predict the potential topology of a transmembrane protein from its protein sequence. In order to achieve this task, they use information 'hidden' in the protein sequence such as hydrophobicity, the distribution of charged residues [26], amino acid preferences, the existence of signal peptides [13,17,19,26–28] and evolutionary information derived from multiple sequence alignments [9,15,29–31]. Moreover, the use of domain assignments has been reported to be of benefit in topology prediction [32]. During the last few years ab initio topology prediction has been shown to be an attainable goal since it yields comparable performance [33]. Importantly, several methods developed during the last few years [1,13,14,18,21,24,33] allow the incorporation of topological information derived from biochemical studies (constrained prediction), which results in improved topology prediction performance. Such biochemical methods include gene fusion, using enzymes such as alkaline phosphatase, β-galactosidase, β-lactamase and various fluorescent proteins, detection of post-translational modifications such as glycosylation, phosphorylation and biotinylation, cysteine-scanning mutagenesis, proteolysis methods and epitope mapping techniques [34].

Phosphorylation and glycosylation are the most widespread post-translational modifications in eukaryotes [35,36] and occur in a compartment-specific manner in the cell. In eukaryotic cells,

* Corresponding author. Tel.: +30 210 727 4931; fax: +30 210 7274254.
E-mail address: shamodr@biol.uoa.gr (S.J. Hamodrakas).

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.11.001

glycosylation activity is found in the lumen of the endoplasmic reticulum (ER) and it is accomplished by the enzyme oligosaccharyl transferase (OST), which adds oligosaccharides to the amino group of Asparagine (Asn) residues of the consensus sequence Asn-X-Thr/Ser (N-linked glycosylation) [37]. It has been shown that the presence of Proline between Asn and Ser/Thr inhibits N-glycosylation [38] and about 50% of the sites that have a Proline C-terminal to Ser/Thr are not glycosylated [39]. In O-linked glycosylation the glycans are attached to either Serine (Ser) or Threonine (Thr) residues. In transmembrane proteins, glycosylation sites occur at parts of proteins facing the extracellular space and are located to a minimum distance away from the membrane surface [40]. It has been shown that, in some cases, glycosylation occurs only when the acceptor site (Asn residue) is located a minimum of 12 residues upstream or 14 residues downstream of a transmembrane segment ('12 + 14 rule') [40–42]. Therefore, in multi-spanning transmembrane proteins, glycosylated extracellular loops have a minimum length of approximately 30 residues [43]. These constraints are used to map the ends of transmembrane segments using N-glycosylation scanning mutagenesis [42,44].

Protein phosphorylation is the most important and well-studied post-translational modification in eukaryotes and is involved in the regulation of several cellular processes such as cell growth and differentiation, signal transduction and apoptosis [45–48]. The addition of a phosphate group usually occurs in Serine (Ser), Threonine (Thr), Tyrosine (Tyr) and Histidine (His) residues in eukaryotic proteins and approximately 30–50% of proteins are supposed to be phosphorylated at some point [49]. In transmembrane proteins, phosphorylation sites are located at the cytoplasmic regions. Therefore, both the existence of a phosphorylation or a glycosylation site along the sequence of a transmembrane protein provides valuable information about the orientation of the modified region with respect to the membrane [34].

However, phosphorylation and glycosylation prediction methods [50–52] predict modified sites along the whole sequence of a transmembrane protein, failing to distinguish between transmembrane segments, cytoplasmic regions and extracellular regions. One approach is to use a topology prediction algorithm and then filter phosphorylation or glycosylation site prediction results according to the predicted topology [53]. Another combined prediction approach is to use first a phosphorylation or glycosylation prediction method and then use the predicted sites as constraints to topology prediction. We compare these different approaches and discuss advantages and disadvantages of combining the two prediction problems.

We have designed a Hidden Markov Model with a novel architecture, which combines in a single model, topology prediction and phosphorylation and glycosylation site prediction. Finally, we use this model for the development of a novel computational method (HMM based) capable of predicting the topology of a transmembrane protein and the existence of kinase specific phosphorylation sites as well as N-linked and O-linked glycosylation sites. We show that the probability of the existence of a phosphorylation or glycosylation pattern along the protein sequence can be used by prediction algorithms in order to predict the orientation of a transmembrane protein more efficiently.

2. Methods

2.1. Transmembrane protein topology datasets

The training set that we used contains 72 α-helical transmembrane proteins with three dimensional structures determined at near atomic resolution, deposited in the Protein Data Bank (PDB) [2]. The dataset is the one used for the development of HMM-TM [12], and in all cases, the sequences used were obtained from Uniprot [54] after the removal of any signal peptides. For the construction of an independent test set we used PDBTM [55] in order to collect all the available high-resolution structures of eukaryotic α-helical TM proteins deposited in PDB until May 2013. We performed a redundancy check, using BLAST [56] and a non-redundant dataset was created by removing all chains for which a putative homologous entry was already in the set or the training set of 72 membrane proteins. The threshold was defined as <30% pairwise sequence similarity (in a length of more than 80 residues) in a BLAST alignment. For sequences shorter than 80 residues, which are frequent among single-spanning membrane proteins, we used the similarity of less than 50% as threshold in a length of more than 30 residues. The final set consists of 49 α-helical TM proteins (25 single spanning and 24 multi-spanning TM proteins). In order to access the prediction performance of HMMpTM compared against the other prediction methods a more appropriate but smaller dataset was created by removing any proteins sharing homology (using the same criteria mentioned above) with the training datasets of all the prediction methods under comparison. This dataset contains 21 transmembrane proteins. All datasets are available online at http://bioinformatics.biol.uoa.gr/HMMpTM/datasets.

2.2. Phosphorylation and glycosylation site datasets

For the collection of phosphorylation sites in eukaryotic TM proteins, UniProt Accession numbers and positions of phosphorylation sites were retrieved from PhosphoSitePlus [57] and Phospho.ELM version 9.0 [58]. Although PhosphoSitePlus and Phospho.ELM both include kinase specific and non-specific phosphorylation sites, we deliberately used kinase specific phosphorylation data, since they provide information about the catalytic kinase responsible for the modification. N-linked and O-linked glycosylation data were retrieved from UniProt [54] using the subsection of the 'Sequence annotation (Features)' section that specifies the position and type of each covalently attached glycan group (FT MOD RES). Modified sites annotated with non-experimental qualifiers (such as 'Potential', 'Probable', 'By similarity') were excluded. In addition, we used O-GlycBase version 6.0 [59] and ExTopoDB version 1.0 [60] in order to retrieve additional glycosylation data. We examined the location of phosphorylation and glycosylation sites in relation to the topology of transmembrane proteins. Topology information for phosphorylated and glycosylated transmembrane proteins was retrieved from Uniprot's sequence annotation, PDB if a three dimensional structure of the protein was available and from literature information present in ExTopoDB. All datasets are available online at http://bioinformatics.biol.uoa.gr/HMMpTM/datasets/.

2.3. Two stage approaches

We examined two different two-stage approaches for the combination of transmembrane protein topology prediction with phosphorylation and glycosylation site prediction. First, we used predicted phosphorylation and glycosylation sites as constraints to topology prediction. Predicted phosphorylation and glycosylation sites were used as constraints for cytoplasmic and extracellular localization respectively. A second approach was the use of topology prediction results as a filter for phosphorylation and glycosylation site prediction. More specifically, we evaluated as predicted phosphorylation sites, only sites residing in predicted cytoplasmic regions. In addition, predicted glycosylation sites in transmembrane and cytoplasmic regions were ignored. In all cases, HMM-TM was used for the topology prediction of transmembrane proteins. NetPhosK, NetNGlyc and NetOGlyc were used for phosphorylation, N-linked glycosylation and O-linked glycosylation site prediction, respectively.

2.4. The Hidden Markov Model

The Hidden Markov Model (HMM) that we used is quite similar to the one proposed by HMM-TM. It consists of five different submodels corresponding to the five desired labels to predict (Fig. 1),

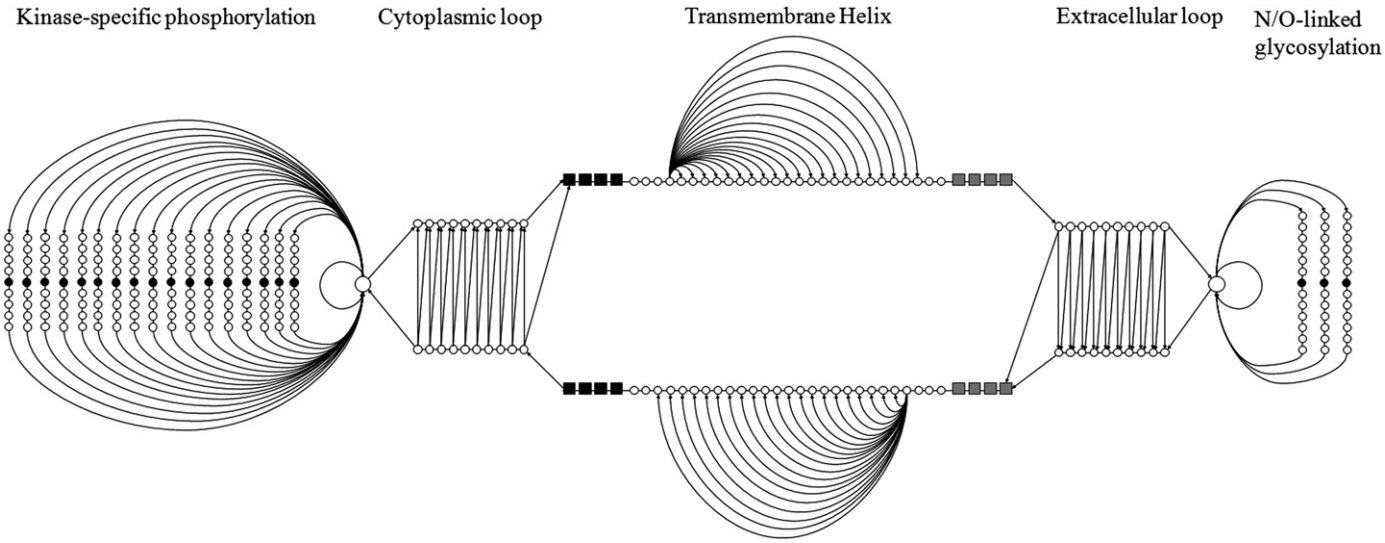

Fig. 1. A schematic representation of the model's architecture. The model consists of five sub-models denoted by the labels: Kinase-specific phosphorylation, Cytoplasmic loop, Transmembrane Helix and Extracellular loop, N/O-linked glycosylation. Within each sub-model, states with the same shape, size and color are sharing the same emission probabilities (parameter tying). Allowed transitions are indicated with arrows.

the Cytoplasmic Loop sub-model, the Transmembrane Helix sub-model, the Extracellular Loop sub-model, the Phosphorylation Site sub-model corresponding to kinase-specific phosphorylation sites and the Glycosylation Site sub-model used to model the existence of N/O-linked glycosylation sites.

The model is cyclic, consisting of 306 states, including begin (B) and end (E) states (Fig. 1) with 206 freely estimated transitions. On the other hand, the total number of freely estimated emission probabilities is 3036, yielding a total number of freely estimated parameters equal to 3242. All states are connected with the appropriate transition probabilities in order to be consistent with the known structures, that is, to ensure appropriate length distribution. The inner and outer loops are modeled with a "ladder" architecture. At both ends, there is a self transitioning state corresponding to residues too distant from the membrane; these cannot be modeled as loops, hence that state is named "globular". Due to the fact that phosphorylation and glycosylation sites are compartment specific, the Phosphorylation Site and Glycosylation Site sub-models are connected with the Cytoplasmic Loop and the Extracellular Loop sub-models, respectively. The Phosphorylation Site sub-model includes additional sub-models, each one corresponding to a kinase-specific phosphorylation pattern. Each phosphorylation pattern includes the phosphorylation site and 4 flanking residues at each side of the modified site $(-4, +4)$. For N-linked and O-linked glycosylation patterns, in the Glycosylation Site sub-model, we used 6 flanking residues at each side of the modified site $(-6, +6)$.

Due to the fact that there was not a dataset available where both the topologies of transmembrane proteins and the locations of phosphorylation (kinase-specific) and glycosylation sites were known, the model could not be trained as a whole. The Cytoplasmic Loop, the Transmembrane Helix and the Extracellular Loop sub-models correspond to the original Hidden Markov Model of HMM-TM and therefore we used the same emissions and transitions as described and estimated in HMM-TM, where the Baum–Welch algorithm for labeled sequences had been used [12]. On the other hand, emissions and transitions for the Phosphorylation Site and the Glycosylation Site sub-models were calculated using the 1022 and 1429 phosphorylation and glycosylation sites, respectively. In addition, the transitions from the Cytoplasmic Loop and the Extracellular Loop sub-models to the Phosphorylation Site and the Glycosylation Site sub-models were manually set. The decoding is performed using the Posterior–Viterbi algorithm [61].

2.5. Comparison to transmembrane protein topology prediction methods

To evaluate the accuracy of the developed method we used the independent test set of 49 TM proteins and compared our model against the performance of various prediction methods such as TMHMM [11], HMMTOP [14], PHOBIUS [13], PHILIUS [17], TOPCONS [24], TOPCONS-single [62], SCAMPI [33], OCTOPUS [18], MEMSAT3 [29] and MEMSAT_SVM [16]. In each case we used the Mathew's correlation coefficient (C), the percentage of correctly predicted residues (Q) [63], the segment overlap (SOV) measure [64] and the percentage of correctly predicted transmembrane segments and topologies.

In addition, following the approach of Tsirigos et al. [65], we used three additional large-scale datasets. A dataset of 546 membrane proteins with experimentally determined C-terminal locations in *Saccharomyces cerevisiae* [66] and the two datasets we compiled with 410 and 765 transmembrane proteins with experimentally verified phosphorylation and glycosylation sites respectively. In the *S. cerevisiae* dataset, we evaluated the prediction performance through the correct topology prediction of the experimentally determined C-terminal locations only. In the phosphorylation and glycosylation datasets, we examined the correct topology prediction of the modified sites only (cytoplasmic and extracellular respectively). All datasets are available online at http://bioinformatics.biol.uoa.gr/HMMpTM/datasets/.

2.6. Comparison to phosphorylation and glycosylation prediction methods

The prediction method we propose aims at predicting reliable topological models of transmembrane proteins by incorporating information about phosphorylation and glycosylation sites along the protein sequence rather than substituting available phosphorylation and glycosylation prediction methods. However, we evaluate the prediction performance of our method compared to available phosphorylation and glycosylation predictors. In the case of phosphorylation prediction we used NetPhosK 1.0 [50] whereas NetNGlyc 1.0 [51] and NetOGlyc 3.1 [52] were used for glycosylation prediction. In both cases, as measures of the prediction performance we used Sensitivity (Sn), Specificity (Sp), Accuracy (Acc) and the Mathew's correlation coefficient (C). True/false positives (TP, FP) and true/false negatives (TN, FN) for each method were counted on a per residue basis. Sensitivity is measured as $TP/(TP + FN)$,

Specificity is measured as TN/(TN + FP), Accuracy is calculated as (TP + TN)/(TP + TN + FP + FN) and Matthews Correlation Coefficient (C) is calculated as $(TP * TN - FP * FN)/\sqrt{((TN + FN) * (TN + FP) * (TP + FN) * (TP + FP))}$. Moreover, S/T/Y/N residues that have not been shown to be modified (phosphorylated or glycosylated) were used as negative data. However, we have to note that some of the negative data could be modified sites not experimentally studied yet.

3. Results

3.1. Analysis of phosphorylation and glycosylation sites in transmembrane proteins

PhosphoELM [58] contains more than 42,500 phosphorylation sites in 8718 phosphorylated eukaryotic proteins and PhosphoSitePlus [57] is comprised of approximately 208,928 phosphorylation sites in 31,642 proteins from different species. Using both databases we compiled a set of 32,667 unique phosphorylated proteins and further selected 29,925 entries mapped to Uniprot. According to Uniprot's annotation we selected 5970 transmembrane proteins having both kinase specific and non-specific phosphorylation data. Subsequently, we selected 502 transmembrane proteins with information about 107 kinases mediating the phosphorylation process. However, only 9 major protein kinase types (PKA, PKC, CAMKII, MAPK, CK1, CK2, GRKs, CDK1, and SRC) were used, where enough data were available for further analysis (Table 1). This resulted in 410 transmembrane proteins having 719, 161 and 142 phosphorylated Serine, Threonine and Tyrosine residues, respectively. Moreover, we collected 1313 N-linked and 116 O-linked experimentally verified glycosylation sites in 751 and 40 transmembrane proteins, respectively (Table 2).

First, the modified (phosphorylated and glycosylated) proteins were classified according to the number of their TM segments (Fig. S2) and their functions. As a result, 161 out of the 410 phosphorylated proteins were single spanning with 109 characterized as type I. In addition, among the 249 multi-spanning TM proteins there are 49 proteins with 7 TM segments that belong to the G-protein coupled receptor superfamily (GPCRs). Moreover, there are 111 protein channels involved in passive transport (facilitated diffusion) with 2–24 TM segments. Overall, there are 197 proteins with transporter activity, 145 proteins with receptor activity and 77 proteins with catalytic activity. As shown in Table 1, the 410 transmembrane proteins are modified by 9 protein kinase types. Most of the transmembrane proteins we collected are modified by kinases of the AGC group (PKA, PKC, GRKs), which contains many intracellular signaling kinases which are modulated by cyclic nucleotides (PKA) and phospholipids (PKC).

In all cases, the recognition motifs in the substrate transmembrane protein sequences were similar to the accepted consensus sequence motifs (Table S1). Sequence variations around the acceptor sites in some cases (e.g. CK1) occur as a result of the small number of observed phosphorylation sites in transmembrane proteins. As a result, we trained the phosphorylation and glycosylation site sub-models of the HMM using the phosphorylation (−4, +4) and glycosylation patterns (−6, +6) obtained from non-transmembrane sequences modified by the certain kinase category.

Table 1
Statistics for kinase-specific phosphorylation data in α-helical transmembrane proteins.

Catalytic kinase	Proteins	Sites	S	T	Y
Protein kinase A (PKA)	140	282	254	28	–
Protein kinase C (PKC)	179	336	273	63	–
Ca^{2+}/calmodulin-dependent protein kinase II (CAMKII)	51	81	62	19	–
Casein kinase 1 (CK1)	12	25	21	4	–
Casein kinase 2 (CK2)	47	93	80	13	–
Mitogen-activated protein kinases (MAPKs)	32	60	35	25	–
Cyclin-dependent kinase 1 (CDK1)	15	18	11	7	–
G protein-coupled receptor kinases (GRKs)	24	83	60	23	–
Proto-oncogene tyrosine-protein kinase Src (SRC)	73	142	–	–	142
Total	410	1022	719	161	142

Table 2
Statistics for glycosylation data in alpha-helical transmembrane proteins.

Glycosylation type	Proteins	Sites	S	T	N
N-linked	751	1313	–	–	1313
O-linked	40	116	61	55	–
Total	765	1429	61	55	1313

The majority of glycosylated transmembrane proteins are single-spanning (543 out of 765) (Fig. S2), mostly with receptor and signal transduction activity involved in cell communication. Among the glycosylated transmembrane proteins, there are 130 proteins with transporter activity, 199 proteins with receptor activity and 208 proteins with catalytic activity. As previously reported for N-linked glycosylation [67], 60% of glycosylation sites had the Asn-X-Thr motif. Notably, 95% of O-linked glycosylated TM proteins were single-spanning.

Next, we examined the location of phosphorylation and glycosylation sites in relation to the topology of TM proteins. As expected, phosphorylation sites are located at the cytoplasmic regions of TM proteins whereas N-linked and O-linked glycosylation sites are located at regions facing the extracellular space. In both cases, we observe that most phosphorylation and glycosylation sites are located at the terminal regions of transmembrane proteins [68] (Fig. S1). However, phosphorylation sites are mostly located at the C-terminal region, as opposed to glycosylation sites that are mostly located at the N-terminal region.

Specifically, 71% of glycosylation sites are located at the N-terminal region and only 15% of sites are located at the C-terminal region. Interestingly, in multi-spanning TM proteins, 95% of glycosylation sites are located at loop regions and 85% of them are located at the first extracellular loop or the first extracellular loop that is larger than approximately 30 residues. These findings are in agreement with previous studies reporting that, glycosylation sites are less frequent at the C-terminal end of a protein [39,69] and that when N-glycosylation sites are contained within more than one extracellular loop, only the first loop is modified [43]. However, in 40 multi-spanning TM proteins, glycosylation sites are not located at the first extracellular loop. Closer investigation of these cases showed that the first extracellular loop in these transmembrane proteins is smaller than approximately 30 residues and therefore these loops could not be glycosylated (according to the 12 + 14 rule) [40]. By contrast, 64% of phosphorylation sites are located at the C-terminal region of TM proteins. In multi-spanning TM proteins, only 26% of phosphorylation sites are located at loop regions and 16% of them are found in regions smaller than 30 residues.

3.2. Two stage approaches

As already discussed, one approach to combine topology prediction with phosphorylation and glycosylation site prediction is to use predicted topologies to filter the predicted modified sites across the transmembrane protein sequence. For this reason we used HMM-TM as a topology prediction algorithm and filtered the prediction results of NetPhosK, NetNGlyc and NetOGlyc on the datasets of 410 and 765 phosphorylated and glycosylated transmembrane proteins. The overall prediction accuracy of NetPhosK, NetNGlyc and NetOGlyc before and after topology prediction filtering is summarized in Table 3. We observe that topology filtering results in increased prediction specificity for phosphorylation and glycosylation prediction. Importantly, in kinase specific phosphorylation prediction using NetPhosK, specificity increased 29%. However, in all cases, prediction sensitivity significantly decreases since falsely predicted topologies produce incorrect filters for phosphorylation and glycosylation site prediction results.

Table 3
Sensitivity (Sn), Specificity (Sp) and Mathew's correlation coefficient (C) measures for phosphorylation and glycosylation site prediction before and after transmembrane protein topology filtering in the sets of 410 and 765 phosphorylated and glycosylated transmembrane proteins.

Method	Sn	Sp	C
Phosphorylation site prediction			
HMMpTM	0.64	0.76	0.23
NetPhosK	0.78	0.45	0.15
NetPhosK (after topology filtering)	0.63	0.74	0.11
N-linked glycosylation site prediction			
HMMpTM	0.64	0.85	0.28
NetNGlyc	0.80	0.87	0.40
NetNGlyc (after topology filtering)	0.56	0.92	0.34
O-linked glycosylation site prediction			
HMMpTM	0.22	0.95	0.14
NetOGlyc	0.70	0.77	0.23
NetOGlyc (after topology filtering)	0.42	0.92	0.24

Another approach we tested was the use of predicted phosphorylated and glycosylated sites as topological constraints for transmembrane protein topology prediction. To evaluate this methodology we used the non-redundant set of 49 alpha-helical transmembrane proteins. When all predicted phosphorylation and glycosylation sites are incorporated, the topology prediction procedure results in error, since neighboring residues are predicted as phosphorylated and glycosylated. In specific, this procedure produced conflicting topological constraints for 24 out of 49 transmembrane proteins. For the remaining 25 transmembrane proteins, only 5 (20%) were predicted with correct topologies. Thus, it is clear that this information cannot be used as topological information for constrained topology predictions. In an effort to recover from this error, we used as constraints only predicted phosphorylation and glycosylation sites with the highest probability. Again, this procedure produced conflicting topological constraints for 5 out of 49 transmembrane proteins. For the remaining 44 transmembrane proteins, only 18 (41%) were predicted with correct topologies. Therefore, we observe again that the combination of the two independent predictions (topology prediction and post-translational modification prediction) results in decrease of topology prediction accuracy.

3.3. Topology prediction performance

In Table 4, we compare the topology prediction performance of our HMM, on a test set of 49 eukaryotic transmembrane proteins. We have to note that the test set used has no homology with the training set of HMM-TM and HMMpTM. We observe 10.2% improvement in correctly predicted transmembrane segments and 6.1% in correctly predicted topologies compared to HMM-TM. It is evident that the incorporation of phosphorylation and glycosylation site prediction in topology prediction results in improved prediction performance. In order to access the influence of phosphorylation and glycosylation separately we developed two additional models. An HMM having only the Phosphorylation site sub-model (HMMpTM_phos) and one with only the Glycosylation site sub-model (HMMpTM_glyc). In both cases we observe an improvement in topology prediction compared to HMM-TM (Table 4). Interestingly, the incorporation of glycosylation site prediction (HMMpTM_glyc) results in better prediction of the number and the position of transmembrane segments compared to the use of phosphorylation site prediction only (HMMpTM_phos). However, the combination of both phosphorylation and glycosylation results (HMMpTM) in the highest topology prediction accuracy. Even though some of the proteins present in the test set were also included in the sets used for training the other predictors, we observe that HMMpTM performs better, compared to methods that use single sequences. As expected [15], prediction methods using multiple sequence alignments (MSA) outperform the methods using single sequence. In order to better access the prediction performance of HMMpTM compared to prediction methods that utilize multiple sequence alignments a more appropriate but smaller dataset (21 transmembrane proteins) was created by removing any proteins sharing homology with the training datasets of all compared prediction methods (Table 5). We observe again that HMMpTM outperforms HMM-TM and other single sequence based prediction methods or a consensus of them (TOPCONS-single). Compared to methods that use multiple sequence alignments we observe that HMMpTM has in most cases comparable or better performance. Only MEMSAT3 performs better in predicting the correct topology of proteins and MEMSAT-SVM in predicting the correct number of transmembrane segments in each protein.

In addition, we evaluated the prediction performance of HMMpTM in three additional large scale datasets. In all three cases, HMMpTM performs better than HMM-TM. In particular, HMMpTM predicts correctly 79% of the experimentally determined C-terminal locations in the *S. cerevisiae* dataset compared to 73% for HMM-TM (Table S2). As already discussed, phosphorylation and glycosylation sites in transmembrane proteins have cytoplasmic and extracellular topology respectively. Therefore, available topology prediction methods, although they cannot provide information about the existence of modified sites, they should efficiently predict the site's topology. As shown in Table S2, HMMpTM correctly predicts the topology for 91% of the phosphorylation sites and 78% of the glycosylation

Table 4
Topology prediction accuracy on an independent dataset of 49 eukaryotic transmembrane proteins with known three-dimensional structures.

Method	Q	C	SOV	Correctly predicted TM segments (%)	Correctly predicted topologies (%)
HMMpTM	0.88	0.70	0.91	81.6	75.5
HMMpTM_phos	0.88	0.72	0.91	75.5	71.4
HMMpTM_glyc	0.88	0.70	0.90	79.6	71.4
HMM-TM	0.86	0.66	0.87	71.4	69.4
TMHMM[b]	0.87	0.67	0.87	69.4	61.2
HMMTOP[b]	0.87	0.68	0.90	77.6	69.4
Phobius[b]	0.86	0.65	0.86	71.4	59.2
Philius[b]	0.87	0.68	0.89	75.5	69.4
SCAMPI[a,b]	0.88	0.69	0.90	83.7	75.5
SCAMPI-single[b]	0.87	0.66	0.89	79.6	71.4
OCTOPUS[a,b]	0.88	0.69	0.90	83.7	75.5
MEMSAT3[a,b]	0.88	0.70	0.92	87.8	83.7
MEMSAT-SVM[a, b]	0.91	0.78	0.96	95.9	85.7
TOPCONS[a,b]	0.89	0.71	0.92	85.7	77.6
TOPCONS-single[b]	0.87	0.66	0.88	73.5	63.3

[a] Methods using evolutionary information (through multiple sequence alignments).
[b] These predictors were trained on sets containing sequences homologous to the ones included in the test set we used here.

Table 5
Topology prediction accuracy on an independent dataset of 21 eukaryotic transmembrane proteins with known three-dimensional structures.

Method	Q	C	SOV	Correctly predicted TM segments (%)	Correctly predicted topologies (%)
HMMpTM	0.90	0.75	0.93	85.7	81.0
HMM-TM	0.90	0.74	0.91	76.2	76.2
TMHMM	0.90	0.75	0.90	71.4	61.9
HMMTOP	0.90	0.74	0.93	81.0	76.2
Phobius	0.90	0.75	0.91	71.4	61.9
Philius	0.91	0.77	0.89	71.4	71.4
SCAMPI[a]	0.90	0.74	0.95	85.7	76.2
SCAMPI-single	0.89	0.70	0.92	81.0	76.2
OCTOPUS[a]	0.90	0.76	0.95	85.7	76.2
MEMSAT3[a]	0.90	0.75	0.95	90.5	85.7
MEMSAT-SVM[a]	0.93	0.82	0.95	90.5	81.0
TOPCONS[a]	0.91	0.78	0.95	85.7	81.0
TOPCONS-single	0.91	0.76	0.94	81.0	76.2

[a] Methods using evolutionary information (through multiple sequence alignments).

sites, resulting in 12% and 7% improvement respectively, compared to HMM-TM.

3.4. Phosphorylation and glycosylation prediction performance

HMMpTM incorporates phosphorylation and glycosylation prediction in the topology prediction procedure, resulting in improved performance. In order to evaluate the prediction accuracy for phosphorylation and glycosylation sites we compared HMMpTM to NetPhosK 1.0, NetNGlyc 1.0 and NetOGlyc 3.1. In kinase specific phosphorylation prediction we observe that HMMpTM shows comparable and in some cases better performance than NetPhosK (Table S3).

However, we have to note that HMMpTM has been designed and optimized for accurate topology prediction of transmembrane proteins. Consequently, in the case of phosphorylation and glycosylation site prediction, high specificity was favored since a large number of false positives would lead the model to a wrong topology. Therefore, in phosphorylation and glycosylation prediction performance, sensitivity and specificity are not balanced. Notably, HMMpTM shows comparable performance with other available prediction tools (Tables S3, S4, S5 and S6). Overall, HMMpTM predicts phosphorylation sites in transmembrane proteins with 64% sensitivity and 76% specificity compared to 78% and 45% for NetPhosK. In N-linked glycosylation site prediction our method correctly predicts the location for 64% of sites with 85% specificity. On the other hand, in O-linked glycosylation site prediction, HMMpTM shows relative small sensitivity compared to NetOGlyc. Again we have to note that HMMpTM has been optimized for high specificity in the prediction of modified sites and it is not expected to be used as a dedicated PTM predictor. A large number of false positives would increase the probability of these regions to be predicted as cytoplasmic or extracellular resulting in less reliable topology prediction.

4. Discussion

We presented a method that integrates a novel feature in topology prediction. HMMpTM is not just a consensus of post-translational modification and topology prediction but integrates in a single Hidden Markov Model phosphorylation and glycosylation prediction in order to more accurately predict the orientation of transmembrane proteins in membranes. Therefore, we have shown that the accuracy in prediction of transmembrane topology increases, whereas at the same time, the model provides reliable (to some degree) predictions for glycosylation and phosphorylation. In addition, protein phosphorylation plays a fundamental role in most of the cellular regulatory pathways and protein glycosylation is important for protein folding and stability as well as cell–cell interactions. Consequently, prediction of phosphorylation and glycosylation events in transmembrane proteins provides important information about their function. N-linked glycosylation site prediction can also serve as a molecular ruler to define the ends of transmembrane segments [44]. Notably, SOV, which precisely measures the overlap of predicted with observed transmembrane segments, improves 4% compared to HMM-TM. Reliable phosphorylation and glycosylation sites in transmembrane proteins provide valuable topological information and can be additionally used for the evaluation of the performance of topology prediction methods for transmembrane proteins [65]. We provide evidence that the incorporation of phosphorylation and glycosylation probabilities in topology prediction improves the prediction performance and can be implemented in existing prediction methods. The clear improvement of HMMpTM over HMM-TM in all relevant measures of accuracy provides evidence that the same procedure may also increase the performance of the other predictors, although a smaller improvement is expected for top-scoring methods. Moreover, the same method can be used also for methods that utilize evolutionary information. There is evidence that evolutionary information in the form of multiple alignment can also increase the prediction accuracy of methods for predicting post-translational modifications [70–74], so this needs to be investigated in future studies.

Acknowledgements

The authors would like to thank the handling editor for properly handling this manuscript and the anonymous reviewers for their useful and constructive criticism.

Funding: This research has been co-financed by the European Union (European Social Fund — ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) — Research Funding Program 'Heracleitus II: Investing in knowledge society', through the European Social Fund.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbapap.2013.11.001.

References

[1] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a Hidden Markov Model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.
[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[3] M. Punta, L.R. Forrest, H. Bigelow, A. Kernytsky, J. Liu, B. Rost, Membrane protein prediction methods, Methods 41 (2007) 460–474.
[4] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[5] M.G. Claros, G. von Heijne, TopPred II: an improved software for membrane protein structure predictions, Comput. Appl. Biosci. 10 (1994) 685–686.
[6] L. Sipos, G. von Heijne, Predicting the topology of eukaryotic membrane proteins, Eur. J. Biochem. 213 (1993) 1333–1340.
[7] C. Pasquier, V.J. Promponas, G.A. Palaios, J.S. Hamodrakas, S.J. Hamodrakas, A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm, Protein Eng. Des. Sel. 12 (1999) 381–385.
[8] D.T. Jones, W.R. Taylor, J.M. Thornton, A model recognition approach to the prediction of all-helical membrane protein structure and topology, Biochemistry 33 (1994) 3038–3049.
[9] B. Rost, R. Casadio, P. Fariselli, C. Sander, Transmembrane helices predicted at 95% accuracy, Protein Sci. 4 (1995) 521–533.
[10] C. Pasquier, S.J. Hamodrakas, An hierarchical artificial neural network system for the classification of transmembrane proteins, Protein Eng. Des. Sel. 12 (1999) 631–634.
[11] E.L. Sonnhammer, G. von Heijne, A. Krogh, A Hidden Markov Model for predicting transmembrane helices in protein sequences, Proc. Int. Con.f Intell. Syst. Mol. Biol. 6 (1998) 175–182.
[12] P.G. Bagos, T.D. Liakopoulos, S.J. Hamodrakas, Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins, BMC Bioinformatics 7 (2006) 189.
[13] L. Kall, A. Krogh, E.L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method, J. Mol. Biol. 338 (2004) 1027–1036.
[14] G.E. Tusnady, I. Simon, The HMMTOP transmembrane topology prediction server, Bioinformatics 17 (2001) 849–850.
[15] H. Viklund, A. Elofsson, Best alpha-helical transmembrane protein topology predictions are achieved using Hidden Markov Models and evolutionary information, Protein Sci. 13 (2004) 1908–1917.
[16] T. Nugent, D.T. Jones, Transmembrane protein topology prediction using support vector machines, BMC Bioinformatics 10 (2009) 159.
[17] S.M. Reynolds, L. Kall, M.E. Riffle, J.A. Bilmes, W.S. Noble, Transmembrane topology and signal peptide prediction using Dynamic Bayesian Networks, PLoS Comput. Biol. 4 (2008) e1000213.
[18] H. Viklund, A. Elofsson, OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar, Bioinformatics 24 (2008) 1662–1668.
[19] H. Viklund, A. Bernsel, M. Skwark, A. Elofsson, SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology, Bioinformatics 24 (2008) 2928–2929.
[20] S. Moller, M.D. Croning, R. Apweiler, Evaluation of methods for the prediction of membrane spanning regions, Bioinformatics 17 (2001) 646–653.
[21] P.G. Bagos, T.D. Liakopoulos, S.J. Hamodrakas, Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method, BMC Bioinformatics 6 (2005) 7.
[22] V.J. Promponas, G.A. Palaios, C.M. Pasquier, J.S. Hamodrakas, S.J. Hamodrakas, CoPreTHi: a Web tool which combines transmembrane protein segment prediction methods, In Silico Biol. 1 (1999) 159–162.

[23] J. Nilsson, B. Persson, G. Von Heijne, Prediction of partial membrane protein topologies using a consensus approach, Protein Sci. 11 (2002) 2974–2980.
[24] A. Bernsel, H. Viklund, A. Hennerdal, A. Elofsson, TOPCONS: consensus prediction of membrane protein topology, Nucleic Acids Res. 37 (2009) W465–W468.
[25] M. Klammer, D.N. Messina, T. Schmitt, E.L. Sonnhammer, MetaTM — a consensus method for transmembrane protein topology prediction, BMC Bioinformatics 10 (2009) 314.
[26] G. von Heijne, Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule, J. Mol. Biol. 225 (1992) 487–494.
[27] L. Kall, A. Krogh, E.L. Sonnhammer, Advantages of combined transmembrane topology and signal peptide prediction — the Phobius web server, Nucleic Acids Res. 35 (2007) W429–W432.
[28] L. Kall, Prediction of transmembrane topology and signal peptide given a protein's amino acid sequence, Methods Mol. Biol. 673 (2010) 53–62.
[29] D.T. Jones, Improving the accuracy of transmembrane protein topology prediction using evolutionary information, Bioinformatics 23 (2007) 538–544.
[30] L. Kall, A. Krogh, E.L. Sonnhammer, An HMM posterior decoder for sequence feature prediction that includes homology information, Bioinformatics 21 (Suppl. 1) (2005) i251–i257.
[31] P.L. Martelli, P. Fariselli, R. Casadio, An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins, Bioinformatics 19 (Suppl. 1) (2003) i205–i211.
[32] A. Bernsel, G. Von Heijne, Improved membrane protein topology prediction by domain assignments, Protein Sci. 14 (2005) 1723–1728.
[33] A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. von Heijne, A. Elofsson, Prediction of membrane-protein topology from first principles, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 7177–7181.
[34] M. van Geest, J.S. Lolkema, Membrane topology and insertion of membrane proteins: search for topogenic signals, Microbiol. Mol. Biol. Rev. 64 (2000) 13–33.
[35] G.A. Khoury, R.C. Baliban, C.A. Floudas, Proteome-wide post-translational modification statistics: frequency analysis and curation of the Swiss-Prot database, Sci. Rep. 1 (2011).
[36] R. Apweiler, H. Hermjakob, N. Sharon, On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database, Biochim. Biophys. Acta 1473 (1999) 4–8.
[37] J.K. Welply, P. Shenbagamurthi, W.J. Lennarz, F. Naider, Substrate recognition by oligosaccharyltransferase. Studies on glycosylation of modified Asn-X-Thr/Ser tripeptides, J. Biol. Chem. 258 (1983) 11856–11863.
[38] E. Bause, Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes, Biochem. J. 209 (1983) 331–336.
[39] Y. Gavel, G. von Heijne, Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering, Protein Eng. Des. Sel. 3 (1990) 433–442.
[40] I.M. Nilsson, G. von Heijne, Determination of the distance between the oligosaccharyltransferase active site and the endoplasmic reticulum membrane, J. Biol. Chem. 268 (1993) 5798–5801.
[41] M. Popov, J. Li, R.A. Reithmeier, Transmembrane folding of the human erythrocyte anion exchanger (AE1, Band 3) determined by scanning and insertional N-glycosylation mutagenesis, Biochem. J. 339 (Pt 2) (1999) 269–279.
[42] M. Popov, L.Y. Tam, J. Li, R.A. Reithmeier, Mapping the ends of transmembrane segments in a polytopic membrane protein. Scanning N-glycosylation mutagenesis of extracytosolic loops in the anion exchanger, band 3, J. Biol. Chem. 272 (1997) 18325–18332.
[43] C. Landolt-Marticorena, R.A. Reithmeier, Asparagine-linked oligosaccharides are localized to single extracytosolic segments in multi-span membrane glycoproteins, Biochem. J. 302 (Pt 1) (1994) 253–260.
[44] J.C. Cheung, R.A. Reithmeier, Scanning N-glycosylation mutagenesis of membrane proteins, Methods 41 (2007) 451–459.
[45] T. Pawson, J.D. Scott, Protein phosphorylation in signaling — 50 years and counting, Trends Biochem. Sci. 30 (2005) 286–290.
[46] T. Hunter, Tyrosine phosphorylation: thirty years and counting, Curr. Opin. Cell Biol. 21 (2009) 140–146.
[47] C.D. Wood, T.M. Thornton, G. Sabio, R.A. Davis, M. Rincon, Nuclear localization of p38 MAPK in response to DNA damage, Int. J. Biol. Sci. 5 (2009) 428–437.
[48] J. Zhang, G.V. Johnson, Tau protein is hyperphosphorylated in a site-specific manner in apoptotic neuronal PC12 cells, J. Neurochem. 75 (2000) 2346–2357.
[49] D.E. Kalume, H. Molina, A. Pandey, Tackling the phosphoproteome: tools and strategies, Curr. Opin. Chem. Biol. 7 (2003) 64–69.
[50] N. Blom, T. Sicheritz-Ponten, R. Gupta, S. Gammeltoft, S. Brunak, Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence, Proteomics 4 (2004) 1633–1649.
[51] R. Gupta, S. Brunak, Prediction of glycosylation across the human proteome and the correlation to protein function, Pac. Symp. Biocomput. (2002) 310–322.
[52] K. Julenius, A. Molgaard, R. Gupta, S. Brunak, Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites, Glycobiology 15 (2005) 153–164.
[53] S.A. Chen, T.Y. Lee, Y.Y. Ou, Incorporating significant amino acid pairs to identify O-linked glycosylation sites on transmembrane proteins and non-transmembrane proteins, BMC Bioinformatics 11 (2010) 536.
[54] The_Uniprot_Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt), Nucleic Acids Res. 40 (2012) D71–D75.
[55] G.E. Tusnady, Z. Dosztanyi, I. Simon, PDB_TM: selection and membrane localization of transmembrane proteins in the Protein Data Bank, Nucleic Acids Res. 33 (2005) D275–D278.
[56] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[57] P.V. Hornbeck, J.M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, M. Sullivan, PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse, Nucleic Acids Res. 40 (2012) D261–D270.
[58] H. Dinkel, C. Chica, A. Via, C.M. Gould, L.J. Jensen, T.J. Gibson, F. Diella, Phospho.ELM: a database of phosphorylation sites — update 2011, Nucleic Acids Res. 39 (2011) D261–D267.
[59] R. Gupta, H. Birch, K. Rapacki, S. Brunak, J.E. Hansen, O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins, Nucleic Acids Res. 27 (1999) 370–372.
[60] G.N. Tsaousis, K.D. Tsirigos, X.D. Andrianou, T.D. Liakopoulos, P.G. Bagos, S.J. Hamodrakas, ExTopoDB: a database of experimentally derived topological models of transmembrane proteins, Bioinformatics 26 (2010) 2490–2492.
[61] P. Fariselli, P.L. Martelli, R. Casadio, A new decoding algorithm for Hidden Markov Models improves the prediction of the topology of all-beta membrane proteins, BMC Bioinformatics 6 (Suppl. 4) (2005) S12.
[62] A. Hennerdal, A. Elofsson, Rapid membrane protein topology prediction, Bioinformatics 27 (2011) 1322–1323.
[63] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.
[64] A. Zemla, C. Venclovas, K. Fidelis, B. Rost, A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment, Proteins 34 (1999) 220–223.
[65] K.D. Tsirigos, A. Hennerdal, L. Kall, A. Elofsson, A guideline to proteome-wide alpha-helical membrane protein topology predictions, Proteomics 12 (2012) 2282–2294.
[66] H. Kim, K. Melen, M. Osterberg, G. von Heijne, A global topology map of the *Saccharomyces cerevisiae* membrane proteome, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 11142–11147.
[67] B. Yan, W. Zhang, J. Ding, P. Gao, Sequence pattern for the occurrence of N-glycosylation in proteins, J. Protein Chem. 18 (1999) 511–521.
[68] T.S. Nuhse, A. Stensballe, O.N. Jensen, S.C. Peck, Phosphoproteomics of the *Arabidopsis* plasma membrane and a new phosphorylation site database, Plant Cell 16 (2004) 2394–2405.
[69] S. Ben-Dor, N. Esterman, E. Rubin, N. Sharon, Biases and complex patterns in the residues flanking protein N-glycosylation sites, Glycobiology 14 (2004) 95–101.
[70] A.N. Nguyen Ba, A.M. Moses, Evolution of characterized phosphorylation sites in budding yeast, Mol. Biol. Evol. 27 (2010) 2027–2037.
[71] C.R. Landry, E.D. Levy, S.W. Michnick, Weak functional constraints on phosphoproteomes, Trends Genet. 25 (2009) 193–197.
[72] R. Malik, E.A. Nigg, R. Korner, Comparative conservation analysis of the human mitotic phosphoproteome, Bioinformatics 24 (2008) 1426–1432.
[73] F. Gnad, S. Ren, J. Cox, J.V. Olsen, B. Macek, M. Oroshi, M. Mann, PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites, Genome Biol. 8 (2007) R250.
[74] B. Zhao, T. Pisitkun, J.D. Hoffert, M.A. Knepper, F. Saeed, CPhos: a program to calculate and visualize evolutionarily conserved functional phosphorylation sites, Proteomics 12 (2012) 3299–3303.