We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

3,900
Open access books available

116,000
International authors and editors

120M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
Abstract

Protein phosphorylation, catalyzed by protein kinases, is the main posttranslational modification in eukaryotes, regulating essential aspects of cellular function. Using mass spectrometry techniques, a profound knowledge has been achieved in the localization of phosphorylated residues at proteomic scale. Although it is still largely unknown, the protein kinases are responsible for such modifications. To fill this gap, many computational algorithms have been developed, which are capable to predict kinase–substrate relationships. The greatest difficulty for these approaches is to model the complex nature that determines kinase–substrate specificity. The vast majority of predictors is based on the linear primary sequence pattern that surrounds phosphorylation sites. However, in the intracellular environment the protein kinase specificity is influenced by contextual factors, such as protein–protein interactions, substrates co-expression patterns, and subcellular localization. Only recently, the development of phosphorylation predictors has begun to incorporate these variables, significantly improving specificity of these methods. An accurate modeling of kinase–substrate relationships could be the greatest contribution of bioinformatics to understand physiological cell signaling and its pathological impairment.

Keywords: protein kinases, phosphorylation, machine learning methods, docking sites

1. Introduction

Protein kinases is the second largest family of enzymes, composed by 518 members in the human genome [1]. These enzymes catalyze the transference of γ-phosphate moiety from adenosine triphosphate (ATP) to the hydroxyl group of serine, threonine, or tyrosine resi-
dues present in substrate proteins. The transient nature of this modification (reversed by dephosphorylation reactions, catalyzed by protein phosphatases) generates the main molecular switch, regulating each aspect of protein function, including interactions, conformations, subcellular localization, enzymatic activity, and turnover. Protein phosphorylation is also the most widespread post-translational modification, affecting at least three-quarters of the proteome [2].

The identification of phosphorylated sites (or phosphosites) has experienced an explosion with the utilization of mass spectrometry techniques. PhosphoSitePlus database [3] collects large part of the information obtained in these studies, including the localization of 144,899 serines, 61,654 threonines, and 41,273 phosphorylated tyrosines, but only 12,180 (5%) of them have annotated the protein kinase responsible for such modifications [3]. This is largely due to the expensive and time-consuming methodologies that need to be used in the identification of kinase–substrate relationships (KSRs).

This complex scenario has opened an important field for the development of computational strategies for phosphorylation site labeling with the specific protein kinase(s) responsible for its modifications in a whole proteome scale, in an effort to reconstruct the underlying regulatory networks. These approaches must overcome several challenges including the complexity of the regulatory networks itself, and the scarce information available about the molecular mechanisms that ensure recognition between protein kinases and substrates.

The list of currently developed tools for KSRs prediction is shown in Table 1. Of note, most of these tools are mostly based on classifiers designed to assign a phosphorylation site to a particular protein kinase considering only the sequence pattern surrounding the phosphorylation site, which provides an imperfect description of the kinase–substrate specificity. In this chapter, we will discuss the underlying biological rational of these tools and its potential for improvement.

<table>
<thead>
<tr>
<th>Method name</th>
<th>Approach</th>
<th>Contextual information</th>
<th>Training data</th>
<th>Number of kinase families</th>
<th>References</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>HeteSim</td>
<td>Heterogeneous information networks</td>
<td>Yes</td>
<td>Phospho. ELM</td>
<td>210 [6]</td>
<td>No web implementation available</td>
<td></td>
</tr>
<tr>
<td>SlapRLS</td>
<td>Supervised Laplacian regularized least squares</td>
<td>No</td>
<td>Phospho. ELM</td>
<td>23 [7]</td>
<td>No web implementation available</td>
<td></td>
</tr>
<tr>
<td>Method name</td>
<td>Approach</td>
<td>Contextual information</td>
<td>Training data</td>
<td>Number of kinase families</td>
<td>References</td>
<td>Website</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------</td>
<td>------------------------</td>
<td>---------------</td>
<td>---------------------------</td>
<td>------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>No name</td>
<td>SVM</td>
<td>No</td>
<td>dbPTM</td>
<td>16</td>
<td>[13]</td>
<td>No web implementation available</td>
</tr>
<tr>
<td>Kinome Xplorer</td>
<td>ANN and PSSM</td>
<td>Yes</td>
<td>Phospho. ELM</td>
<td>222</td>
<td>[14], Previous developments: [15] (NetPhorest) [16] (NetworKIN)</td>
<td><a href="http://kinomexplore.info/">http://kinomexplore.info/</a></td>
</tr>
<tr>
<td>RegPhos 2.0</td>
<td>SVM</td>
<td>Yes</td>
<td>dbPTM</td>
<td>122</td>
<td>[17], Previous developments: [18] (RegPhos)</td>
<td><a href="http://csb.cse.yzu.edu.tw/RegPhos2/">http://csb.cse.yzu.edu.tw/RegPhos2/</a></td>
</tr>
<tr>
<td>No name</td>
<td>SVM</td>
<td>Yes</td>
<td>Phospho. ELM</td>
<td>21</td>
<td>[21]</td>
<td>No web implementation available</td>
</tr>
<tr>
<td>iGPS</td>
<td>Optimized</td>
<td>Yes</td>
<td>Phospho.ELM</td>
<td>69</td>
<td>[24], Previous</td>
<td><a href="http://igps.biocuckoo.org/">http://igps.biocuckoo.org/</a></td>
</tr>
<tr>
<td>Method name</td>
<td>Approach</td>
<td>Contextual information</td>
<td>Training data</td>
<td>Number of kinase families</td>
<td>References</td>
<td>Website</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>------------------------</td>
<td>---------------</td>
<td>---------------------------</td>
<td>------------</td>
<td>---------</td>
</tr>
<tr>
<td>ConDens</td>
<td>BLOSUM62 similarity</td>
<td>+ literature</td>
<td>All kinases with known motifs</td>
<td>492</td>
<td>[30, 31]</td>
<td><a href="http://www.moseslab.csb.utoronto.ca/andy1/">http://www.moseslab.csb.utoronto.ca/andy1/</a></td>
</tr>
<tr>
<td>No name</td>
<td>PSSM</td>
<td>No</td>
<td>Unnecessary</td>
<td>All kinases with known motifs</td>
<td>[25] (GPS 2.1) [26] (GPS 2.0) [27, 28] (GPS) [29]</td>
<td>No web implementation available</td>
</tr>
<tr>
<td>Phos3D</td>
<td>SVM</td>
<td>No</td>
<td>Phospho.ELM</td>
<td>6</td>
<td>[38]</td>
<td><a href="http://phos3d.mpimp-golm.mpg.de/">http://phos3d.mpimp-golm.mpg.de/</a></td>
</tr>
<tr>
<td>MetaPred PS</td>
<td>Meta predictor</td>
<td>No</td>
<td>Phospho.ELM + 15 PhosphoSite Plus + Swiss-Prot</td>
<td>42</td>
<td>[42]</td>
<td>Web implementation no longer available</td>
</tr>
<tr>
<td>Kinase Phos 2.0</td>
<td>SVM</td>
<td>No</td>
<td>Phospho.ELM + Swiss-Prot</td>
<td>71</td>
<td>[44] Previous development: [45] (Kinase Phos 1.0)</td>
<td><a href="http://kinasephos2.mbc.nctu.edu.tw/">http://kinasephos2.mbc.nctu.edu.tw/</a></td>
</tr>
<tr>
<td>PPSP</td>
<td>Bayesian</td>
<td>No</td>
<td>Phospho.ELM</td>
<td>68</td>
<td>[46]</td>
<td>Web implementation</td>
</tr>
</tbody>
</table>
Table 1. Computational methods for kinase–substrate relationships prediction.

<table>
<thead>
<tr>
<th>Method name</th>
<th>Approach</th>
<th>Contextual information</th>
<th>Training data</th>
<th>Number of kinase families</th>
<th>References</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred Phospho</td>
<td>SVM</td>
<td>No Phospho Base</td>
<td>4</td>
<td>[47]</td>
<td></td>
<td>no longer available</td>
</tr>
</tbody>
</table>

2. Comparing prediction tools: data, metrics, and methods

One of the most challenging problems in the field of prediction tools is to establish benchmarks between them, allowing a real assessment of the method itself. Each prediction tool requires for its testing (and often for training) a set of positive (actually phosphorylated sites) and negative data (sites actually not phosphorylated). The sources of phosphorylated sites for most of the predictors are limited to a few databases as Phospho.ELM [4] and PhosphositePlus [3]. These databases include information from different experimental approaches (in vivo and/or in vitro), which is processed homogeneously for training prediction algorithms. This can lead to a significant bias in the quality of predictions: protein kinases exhibit low specificity at in vitro experiments (which constitute the largest proportion of the information in databases), generating simpler motifs than those that may present in cells [5]. Moreover, using information derived only from in vivo experiments does not ensure that the observed phosphorylation site was directly phosphorylated by the protein kinase studied. Careful selection of the positive data set for training, including only the sites phosphorylated both in vivo (physiological) and in vitro (direct) by a protein kinase, can significantly improve the prediction [5].

Another problem is the construction of a negative data set used in machine-learning methods. Although experiments can verify that a protein kinase phosphorylates a given residue, it is very difficult to demonstrate that a particular residue in a protein is not phosphorylated at any situation. A good approximation is made by Neuberger et al. [48], which consider any residue present in a protein which is phosphorylated by a particular protein kinase and which has not been reported as phosphorylated in databases, as part of the negative data set.

The sensitivity ($S_n$) and specificity ($S_p$) are commonly used to assess the performance of prediction algorithms. For a set of data predicted as positive, real positive (previously experimentally determined as phosphorylated) are called true positives (TP), while the remaining are called false positives (FP). Concomitantly, for data predicted as no phosphorylated sites, the real ones are called true negatives (TN) whereas phosphorylated sites are considered false negatives (FN). The ratio of positive sites can be correctly classified is named sensitivity ($S_n$). On the other hand, proportion of negatives sites correctly identified, is called specificity ($S_p$). Both parameters are calculated as follows (Eq. (1)):
Other common parameter used to evaluate the predictor performance is the accuracy that denotes the percentage of correct prediction in both negative and positive data sets. Also, Matthews correlation coefficient (MCC) is widely used as a general estimator for the performance of a predictor. It considers the four numbers described in Eq. (1), giving a balanced assessment of the performance of the predictor even if these parameters are very different. Both parameters are calculated as follows:

\[
S_\text{p} = \frac{TP}{TP + FN} \\
S_\text{n} = \frac{TN}{TN + FP}
\]

\[
\text{Ac} = \frac{TP + TN}{TP + FP + TN + FN}
\]

\[
\text{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}
\]

The ROC (receiver operating characteristic) curve is commonly used for evaluation and comparison of classifiers. The true positive rate (sensitivity) is plotted against the false positive rate (1-specificity). For a perfect classifier the area under curve (AUC) is 1, while a poor classifier achieved values near 0.5 (which defines a random guess).

It may be considered that the parameters previously described can only be compared when the data sets of TP and TN are similar, which is relatively common for the former, but very uncommon for the later. It should be necessary to define standard data sets that are occupied by the research community or define benchmarks that are independent of the training data sets. One approach is to compare the predictors based on their ability to assign to known phosphorylation sites the lowest ranking in a proteomic search for phosphosites of a given protein kinase [5].

Apart from the difficulties to make quantitative comparisons between predictors, there is a perception that machine-learning algorithms, as artificial neural networks (ANN) or support vector machines (SVM), provide a better predictability of protein kinase substrates that simpler methods such as the position-specific scoring matrices (PSSM). Such an idea is substantiated in the assumption that machine-learning algorithms are capable of classifying highly complex sequences, in which correlations amongst positions are important. Such an assumption was recently questioned by studying the sequence interpositional dependence for ataxia telangiectasia mutated (ATM/ATR) kinase, casein kinase 2 (CK2), and cyclin-dependent kinase 2 (CDK2) substrates. Through statistical analysis, Joughin and colleagues [49] found few pairs of positions in the sequences of the phosphorylated sites that are significantly deviated from...
the positionwise independence. Accordingly, the predictors that incorporate second-order information were less accurate than those who consider only first-order information, overfitting the training data [49]. This strongly suggests that a good strategy to develop most accurate predictive tools is to integrate simple sequence models with contextual information, such as protein–protein interactions, subcellular localization, and distal recognition sites.

3. Beyond the sequence: improving substrate prediction with contextual information

There are two factors, which are important to determine the specific phosphorylation of substrates by protein kinases: recruitment and phosphorylation site recognition. Recruitment

![Figure 1](image)

**Figure 1.** A structural view of protein kinases specificity determinants. (A) Local interactions. Extensive contacts are established between the protein kinase active site and the surrounding region of phosphosite, which partly defines the specificity of protein kinases. For example, a complex between Dual-specificity tyrosine-phosphorylation-regulated kinase (DYRK)-1A and a consensus substrate peptide (ARPPPAL) is shown. Active site residues (F170, F196, Y246, D287, K289, E291, Y321, S324, Y327, E353; colored in yellow) establish contacts with the peptide substrate. (B) Distal docking sites. Often the interaction of the substrate with the active site of the protein kinase is not enough to ensure specificity. For example, Extracellular signal regulated kinase (ERK)-2 and its substrate Ribosomal S6 kinase (RSK)-1, establish additional contacts distal to the protein kinase active site, through a linear binding motif (colored in yellow). (C) Scaffolds proteins. Many kinases utilize scaffold proteins to be placed close to their substrates. For example, mitogen-activated protein kinase (MAPK) kinase (MKK)-5 organizes MAPK kinase (MEKK)-3, MKK5, and ERK5 in a signaling complex. (D) Subcellular location. Through protein–protein interactions, protein kinases are located in specific subcellular structures, wherein phosphorylate specific substrates. For example, Polo-like kinase (PLK)-1 interacts through its Polo Box Domain with Cdc25C, allowing its centrosomal localization. Molecular graphics were performed with the UCSF Chimera package [56] based on the following structures: DYRK1A-substrate complex (PDB: 2WO6), ERK2-RSK1 complex (PDB: 2Y4I), ERK5-MKK5-MEKK3 ternary complex (made by superimposition of PDB: 4IC7 and PDB: 2O2V) and PLK1 kinase domain (PDB: 2OJX)-Polo box domain-Cdc25c complex (PDB: 2OJX).
is related with a number of determining factors that promote productive interaction between protein kinases and substrates. Phosphorylation site recognitions are related with the preference of individual residues surrounding the modified residue (Figure 1).

The relative importance of these factors has been rarely studied experimentally in the functional specificity of protein kinases. For example, in yeast, the high specificity of the mitogen-activated protein kinase kinase (MAPKK) is ensured mainly by the use of docking motifs and scaffolding interactions [50].

3.1. Distal docking sites

A transient physical interaction between protein kinases and their substrates could place them in close proximity and in the correct orientation, creating the opportunity for post-translational modification. These interactions are based on short linear motifs, termed docking sites, that reside in disordered regions of the proteins, and that only adopt a defined structure upon binding. The utilization of docking sites seems to be a widespread strategy to improve the specificity of protein kinases to phosphorylate defined phosphosites, as evidenced by studies of SR protein-specific kinase-1 (SRPK1) [51], Cbk1 [52], Polo-like kinases [53], and Cdks [54, 55]. Due to the lack of structural models of interaction between complete substrates and protein kinases, it is not yet possible to measure the importance of distal interactions to the site of phosphorylation in specificity determination. However, by combining protein–protein docking and adaptive biasing force molecular dynamics simulations, Mottin et al. obtained a structural model of the interaction between an active protein kinase (the complex Cdk5/p25) and a complete substrate: Peroxisome proliferator-activated receptor ³ (PPAR³). This model suggests that the protein kinase establishes two distal docking sites with the substrate, pinpointing the importance of those contact sites for proper positioning of the phosphosite in the kinase active site [57].

Mitogen-activated protein kinases (MAPKs) are the prototypical example of using docking sites to enhance the specificity of their promiscuous active sites, which phosphorylate their substrates at most in a weak consensus site (Ser/Thr-Pro). Two types of docking sites have been characterized for MAPKs, called D-sites and F-sites. D-sites interacts with a D-recruitment site (DRS), consisting of a negatively charged region and a shallow hydrophobic pocket, located on the opposite side from the kinase active site. On the other hand, F-sites bind a hydrophobic docking groove (F-recruitment site or FRS). Although all MAPKs have a DRS (hence, it is referred, too, as Common Docking or CD), the FRS seems to be a characteristic only of ERK1/2 and p38α.

Structural studies of the interaction between extracellular-regulated kinase 2 (ERK2) and a peptide containing the F-site from Elk-1 suggests that the main effect is to increase the local phosphosite concentration, enhancing productive encounters that enhance phosphorylation [58].

Although systematic in silico exploration of docking sites could be helpful to find new putative substrates for MAPKs, it has been hampered by low stringency of the sequence, generating a high rate of false positives that can stochastically occur.
Whisenant et al. [59] developed D-finder, a computational tool that uses a hybrid pattern matching algorithm/hidden Markov model, to find sequences of docking sites at protein kinase substrates. Trained with 20 experimentally identified D-sites for a c-Jun N-terminal kinase (JNK), they identified 394 proteins with putative D-sites, and experimentally validated some of them [59]. In order to prioritize the predictions made by D-finder, this was combined with a predictor of JNK phosphorylation sites based on position weight matrices, achieving the identification and experimental validation of one additional substrate of JNK [60].

Significant efforts have been done to study the determinants of specificity of the DRS [61, 62], results that undoubtedly will improve the design of predictors. More recently, Zeke and colleagues [63] addressed the problem by separating the D-sites in 4 class of motifs based on experimental studies and structural and evolutionary analysis, generating PSSMs that allowed them to generate specific interactomes for each class of motif [63].

Utilization of docking motifs to enhance the accuracy of the predictions is still incipient for other kinases. Bibi and collaborators [64] conducted a search for new substrates of Plk1, using both the consensus sequence of phosphorylation site as the recognition motif of the Polo-box domain, responsible for the recruitment of the substrates and activation of the protein kinase [64].

### 3.2. Scaffold proteins

Scaffold proteins are important to ensure the encounter of the protein kinases with its substrates, as suggested by a large number of these signaling proteins associated with phosphorylation [65], especially in the MAPK pathway, where a recent interactome analysis identified 10 scaffold proteins associated [66]. A canonical example of utilization of scaffold proteins is found in the signaling of cAMP-activating protein kinase (PKA), where more than 50 A-kinase anchoring proteins (AKAPs) are responsible for associating the protein kinase with its substrates [67].

The first attempt to integrate protein interactions with PhosphoSite sequence patterns to improve KSR prediction was the development of NetworKIN [16], a two-stage algorithm. In the first term, artificial neural networks (from NetPhosK) and PSSMs (from Scansite) are used to label a given phosphosite sequence with a kinase or kinase family. In a second stage, the contextual information is included, calculating the proximity of the substrate to all kinases (the most likely route connecting them) in a network of functional relationships extracted from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [68].

For a group of well-studies kinase families (CDK, PKC, PIKK and INSR) NetwortKIN doubled prediction accuracy (to 64%) over only sequence-based methods. However, NetworKIN includes a circular logic system, which can overestimate the values of accuracy: the performance assessment was performed with known phosphosites derived from the literature and the STRING database also includes information from text-mining (as co-occurrence in abstracts), therefore associates many kinases with its substrates. NetworKIN has recently been upgraded into KinomeXplorer platform [14]. This introduces improvements in the accuracy of prediction through a new scoring scheme based on naive Bayes method, to overcome the bias in the
network structure caused by highly studied proteins. Another algorithm based on sequence and contextual information is PhosphoPICK. It integrates information from previously known KSRs, protein–protein interactions and protein abundance profiles through the cell cycle in a Bayesian network model. The average AUC for the 59 protein kinases evaluated was 0.86. The main determinant of this good performance was the inclusion of data from protein–protein interactions, while the protein expression throughout the cell cycle had a modest contribution [12].

To our knowledge, no predictor used only physical protein–protein interaction information, which better translates the concept that kinases phosphorylate proteins that are in close proximity. Recently, studies carried out by affinity purification coupled with mass spectrometry revealed the structure of many protein complexes in human cells [69, 70], including some specific for protein kinases [71, 72] or signaling pathways associated with them [73, 74], which can be exploited to improve sequence-only based predictors.

Contextual information can also be given by the location of the protein kinase in a specific subcellular structure, which would have privileged access to certain substrates. Alexander and colleagues elegantly demonstrated that substrate specificity between mitotic kinases, Aurora A/B, Cdk1, Plk1 and Nek2 are based on mutually exclusive motifs in the case of protein kinases presenting overlapping location and for similar motifs in protein kinases showing an exclusive distribution [75].

4. Structural aspects of phosphosites

It has been traditionally assumed that the phosphorylation site can be described as a linear sequence and this assumption underlies almost all predictors developed to date. However, it has been recently identified that the Thr253 of the α-tubulin is located in a nonlinear motif, comprising residues distant in primary sequence but which are folded in a consensus site phosphorylated by PKC [76]. Durek and collaborators [38] had previously addressed this problem by characterizing the structural motifs (3D) present in the phosphorylation sites. Using only the radial distance from the phosphorylated residue and regardless of the angular information, they achieved a modest increase in performance over only linear sequence-predictions [38]. This can be explained if only a limited number of residues are recognized by protein kinases based on a structural epitope or by the low complexity of the model used. Currently, they have developed more sophisticated tools to find patterns in protein 3D structures that could be useful to identify KSRs based on nonlinear motifs. For example, Amino acid pattern Search for Substructures And Motifs (ASSAM) [77]) uses 3D coordinates of a motif comprising up to 12 amino acids, for matching in a structure database as the Protein Data Bank (PDB). ASSAM represents protein structures as a graph in which nodes consist of a vector between two pseudoatoms, representing the side chains of the individual amino acids, while the edges are the distances between the corresponding vectors, providing a more exhaustive representation that used by Durek and collaborators [38].
To date only one predictor, denominated MODPROPEP [78], is based exclusively on structural features of the interaction between the protein kinase and the phosphorylation site. MODPROPEP modulates putative substrate peptides into the active site of the protein kinases, using 49 kinase-peptide complexes with structure available in the Protein Data Bank (PDB) as templates. The scoring function is based on the binding energy of the peptides with the kinase calculated using a statistical-based residue pair potential [79]. Owing to the low accuracy in some families of kinases, a new scoring scheme based on molecular mechanics Poisson-Boltzmann Surface Area (MM-PBSA) binding energy values was generated. The performance with this improved scoring system was similar to predictors based only on sequence as GPS, PPSP, Scansite and NetPhosK, but with the advantage that MODPROPEP not require a training set of known sites, which is a remarkable advantage for prediction of previously uncharacterized protein kinases.

5. The kinase side: determinants of specificity

The combinations of residues in the kinase domain that allow substrate specificity are known determinants of specificity (DoS). For example, discrimination between Ser and Thr by eukaryotic protein kinases appears to depend on the nature of a single residue immediately adjacent to the DFG motif (DFG+1). Whereas most of the Thr-specific kinases have a β-branched aliphatic residue at this position (as Ile, Val or Thr), Ser-specific kinases have large hydrophobic residues (predominantly Leu, Phe, and Met). All non-selective kinases have Leu or Ser as the DFG+1 residue. The mutation of this residue is enough to modify the amino acid preference of multiple protein kinases [80].

The idea of building predictors based on the characteristics of the kinase domain can overcome the problem of building a positive data set of phosphorylated sequences, an aspect especially complicated for poorly characterized protein kinases, and ideally would serve to perform explorations at whole-kinome level. The first DoS-based predictor was Predikin, which allows automatic prediction of peptide substrates using only the amino acid sequence of the protein kinase [33]. This algorithm has been continuously improved [34], achieving the most accurate prediction of experimentally obtained position weight matrices in the category of Domain Recognition Peptide/Kinase protein of Dialogue for Reverse Engineering Assessments and Methods (DREAM4) challenge [32].

On the other hand, Safaei and colleagues [30] using the primary sequence of the kinase catalytic domains, generate PSSMs describing substrate specificity for 492 protein kinases. For this purpose, residues act as DoS are identified using multiple alignments of the kinase domain and the correlation between these and those residues present in the consensus sequences derived from known KSRs [30]. Although the performance of this strategy was similar to NetPhorest, the advantage it does not require information about substrates.

A computational methodology called KINspect (based on learning classifier systems) was recently developed, in order to predict DoS through an iterative process based on randomly generated specificity masks, which are progressively improved in its predictive ability through
mutation and crossover [81]. This sophisticated approach, transferred to a predictor of KSRs, possibly will overcome previous advances.

6. Conclusions

In silico prediction of KSRs may become the most important contribution of bioinformatics toward the understanding of cell signaling. To date, there are no high-throughput experimental techniques allowing pairing thousands of phosphorylation sites (identified by mass spectrometry) with protein kinases that catalyzing such reactions. Recent studies provide a deepest characterization of the determinants of specificity of protein kinases for their substrates, enabling a more realistically modeling of the KSRs. These models reduce the rate of false positives and support the construction of feasible regulatory networks.

Author details

Daniel A. Bórquez¹ and Christian González-Billault²,³

*Address all correspondence to: chrgonza@uchile.cl

1 Biomedical Research Center, School of Medicine, Diego Portales University, Santiago, Chile
2 Department of Biology, Faculty of Sciences, University of Chile, Santiago, Chile
3 FONDAP Geroscience Center, Brain Health and Metabolism, Santiago, Chile

References


