Chapter 5

Bioinformatics Approaches for Predicting Kinase–Substrate Relationships

Daniel A. Bórquez and Christian González-Billault

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/63761

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.

Method name	Approach	Contextual information	Training data	Number of kinase families	References	Website
HeteSim	Heterogeneous information networks	Yes	Phospho.ELM	210	[6]	No web implementation available
SlapRLS	Supervised Laplacian regularized least squares	No	Phospho.ELM	23	[7]	No web implementation available
PUEL	Positive-unlabeled ensemble	Yes	PhosphoSitePlus + Literature	2	[8]	https://github.com/PengyiYang/KSP-PUEL

Method name	Approach	Contextual information	Training data	Number of kinase families	References	Website
	learning (SVM and PSSM-based)					
Scansite 3	PSSM	No	Experimental data (*in vitro*)	62	[9]. Previous developments: [10] (Scansite) [11] (Scansite 2.0)	http://scansite3.mit.edu/
PhosphoPICK	Bayesian network	Yes	Phospho.ELM + HPRD	45	[12]	http://bioinf.scmb.uq.edu.au/phosphopick/
No name	SVM	No	dbPTM	16	[13]	No web implementation available
Kinome Xplorer	ANN and PSSM	Yes	Phospho.ELM	222	[14]. Previous developments: [15] (Net Phorest) [16] (NetworKIN)	http://kinomexplorer.info/
RegPhos 2.0	SVM	Yes	dbPTM	122	[17]. Previous developments: [18] (RegPhos)	http://csb.cse.yzu.edu.tw/RegPhos2/
PSEA	Set enrichment analysis (BLOSUM62 similarity)	No	PhosphoSitePlus	27	[19]	http://bioinfo.ncu.edu.cn/PKPred_Home.aspx
PhosNetConstruct	Similarity matrix	No	PhosphoSitePlus	61	[20]	http://202.54.249.142/~nikhil/network_reconstruction/index.html
No name	SVM	Yes	Phospho.ELM	21	[21]	No web implementation available
phos_pred	Random forest	Yes	Phospho.ELM	54	[22]	http://bioinformatics.ustc.edu.cn/phos_pred/
PKIS	SVM	No	Phospho.ELM	56	[23]	http://bioinformatics.ustc.edu.cn/pkis/
iGPS	Optimized	Yes	Phospho.ELM	69	[24]. Previous	http://igps.biocuckoo.org/

Method name	Approach	Contextual information	Training data	Number of kinase families	References	Website
	BLOSUM62 similarity		+ literature		developments: [25] (GPS 2.1) [26] (GPS 2.0) [27, 28] (GPS)	
ConDens	Conservation of local motif density	No	Unnecessary	All kinases with known motifs	[29]	http://www.moseslab.csb.utoronto.ca/andyl/
No name	PSSM	No	Unnecessary	492	[30, 31]	No web implementation available
PrediKin	PSSM	No	Unnecessary	All	[32]. Previous developments: [33–35]	http://predikin.biosci.uq.edu.au/
Musite	SVM	No	Phospho.ELM + Swiss-Prot + PhosphoPep	13	[36, 37]	http://musite.sourceforge.net/
Phos3D	SVM	No	Phospho. ELM	6	[38]	http://phos3d.mpimp-golm.mpg.de/
NetphosK	ANN	No		17	[39, 40]	http://www.cbs.dtu.dk/services/NetPhosK/
CRPhos	Conditional random fields	No	Phospho. ELM	18	[41]	http://www.ptools.ua.ac.be/CRPhos/
MetaPredPS	Meta predictor	No	Phospho.ELM + PhosphoSitePlus + Swiss-Prot	15	[42]	Web implementation no longer available
PhoScan	Log odds ratios	No	Phospho. ELM	48	[43]	http://bioinfo.au.tsinghua.edu.cn/phoscan/
Kinase Phos 2.0	SVM	No	Phospho. ELM + Swiss-Prot	71	[44]. Previous development: [45] (Kinase Phos 1.0)	http://kinasephos2.mbc.nctu.edu.tw/
PPSP	Bayesian	No	Phospho.	68	[46]	Web implementation

Method name	Approach	Contextual information	Training data	Number of kinase families	References	Website
	decision theory		ELM			no longer available
Pred Phospho	SVM	No	Phospho Base	4	[47]	Web implementation no longer available

Table 1. Computational methods for kinase–substrate relationships prediction.

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 vivo*), 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)):

$$S_n = \frac{TP}{TP + FN}$$

$$S_p = \frac{TN}{TN + FP}$$

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:

$$Ac = \frac{TP + TN}{TP + FP + TN + FN}$$

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

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

http://dx.doi.org/10.5772/63761

the positionwise independence. Accordingly, the predictors that incorporate second-order information were less accurate than those who consider only first-order information, over fitting 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

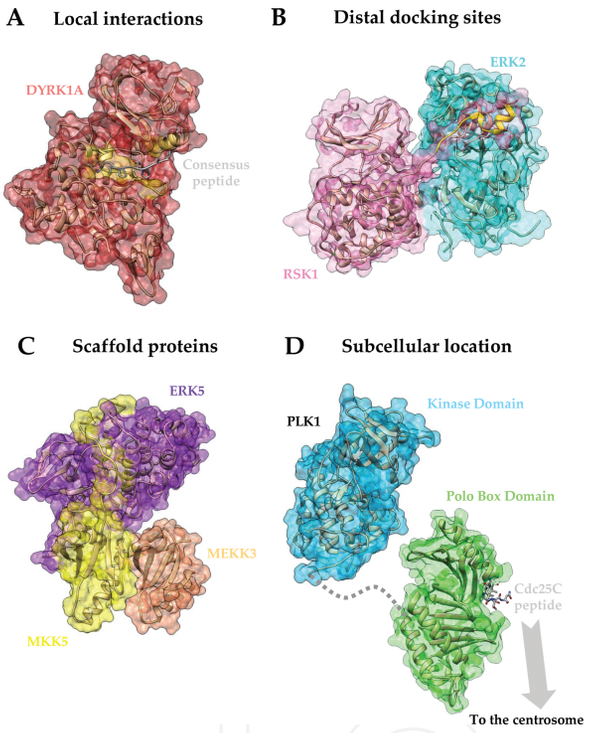

Figure 1. A structural view of protein kinases specificity determinants. (A) Local interactions. Extensive contacts are established between the protein kinase active site and the surrounding region of phophosite, which partly defines the specificity of protein kinases. For example, a complex between Dual-specificity tyrosine-phosphorylation-regulated kinase (DYRK)-1A and a consensus substrate peptide (ARPGT*PAL) 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 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: 2OU7)-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} (PPAR^{3}). 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].

http://dx.doi.org/10.5772/63761

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. Bórquez[1] and Christian González-Billault[2,3*]

*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

[1] Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002;298:1912–34. DOI:10.1126/science.1075762.

[2] Sharma K, D'Souza RC, Tyanova S, Schaab C, Wisniewski JR, Cox J et al. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. Cell Reports. 2014;8:1583–94. DOI:10.1016/j.celrep.2014.07.036.

[3] Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Research. 2015;43:D512–20. DOI:10.1093/nar/gku1267.

[4] Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ et al. Phospho.ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Research. 2011;39:D261–7. DOI:10.1093/nar/gkq1104.

[5] Borquez DA, Olmos C, Alvarez S, Di Genova A, Maass A, Gonzalez-Billault C. Bioinformatic survey for new physiological substrates of Cyclin-dependent kinase 5. Genomics. 2013;101:221–8. DOI:10.1016/j.ygeno.2013.01.003.

[6] Li H, Wang M, Xu X. Prediction of kinase–substrate relations based on heterogeneous networks. Journal of Bioinformatics and Computational Biology. 2015;13:1542003. DOI: 10.1142/S0219720015420032.

[7] Li A, Xu X, Zhang H, Wang M. Kinase identification with supervised laplacian regularized least squares. PLoS One. 2015;10:e0139676. DOI:10.1371/journal.pone. 0139676.

[8] Yang P, Humphrey SJ, James DE, Yang YH, Jothi R. Positive-unlabeled ensemble learning for kinase substrate prediction from dynamic phosphoproteomics data. Bioinformatics. 2016;32:252–9. DOI:10.1093/bioinformatics/btv550.

[9] Ehrenberger T, Cantley LC, Yaffe MB. Computational prediction of protein–protein interactions. Methods in Molecular Biology. 2015;1278:57–75. DOI: 10.1007/978-1-4939-2425-7_4.

[10] Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, Cantley LC. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nature Biotechnology. 2001;19:348–53. DOI:10.1038/86737.

[11] Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Research. 2003;31:3635–41. DOI:10.1093/nar/gkg584

[12] Patrick R, Le Cao KA, Kobe B, Boden M. PhosphoPICK: modelling cellular context to map kinase–substrate phosphorylation events. Bioinformatics. 2015;31:382–9.DOI: 10.1093/bioinformatics/btu663.

[13] Su M-G, Huang K-Y, Tung C-H, Lee T-Y. A new scheme to predict kinase-specific phosphorylation sites on protein three-dimensional structures. International Journal of Bioscience, Biochemistry and Bioinformatics. 2013;3: 473. DOI:10.7763/IJBBB. 2013.V3.258

[14] Horn H, Schoof EM, Kim J, Robin X, Miller ML, Diella F et al. KinomeXplorer: an integrated platform for kinome biology studies. Nature Methods. 2014;11:603–4. DOI: 10.1038/nmeth.2968.

[15] Miller ML, Jensen LJ, Diella F, Jorgensen C, Tinti M, Li L et al. Linear motif atlas for phosphorylation-dependent signaling. Science Signaling. 2008;1:ra2. DOI:10.1126/ scisignal.1159433.

[16] Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, Miron IM et al. Systematic discovery of in vivo phosphorylation networks. Cell. 2007;129:1415–26. DOI:10.1016/j.cell.2007.05.052.

[17] Huang KY, Wu HY, Chen YJ, Lu CT, Su MG, Hsieh YC et al. RegPhos 2.0: an updated resource to explore protein kinase–substrate phosphorylation networks in mammals. Database: the Journal of Biological Databases and Curation. 2014;2014:bau034. DOI: 10.1093/database/bau034.

[18] Lee TY, Bo-Kai Hsu J, Chang WC, Huang HD. RegPhos: a system to explore the protein kinase–substrate phosphorylation network in humans. Nucleic Acids Research. 2011;39:D777–87. DOI:10.1093/nar/gkq970.

[19] Suo SB, Qiu JD, Shi SP, Chen X, Liang RP. PSEA: Kinase-specific prediction and analysis of human phosphorylation substrates. Scientific Reports. 2014;4:4524. DOI:10.1038/ srep04524.

[20] Damle NP, Mohanty D. Deciphering kinase–substrate relationships by analysis of domain-specific phosphorylation network. Bioinformatics. 2014;30:1730–8. DOI: 10.1093/bioinformatics/btu112.

[21] Xu X, Li A, Zou L, Shen Y, Fan W, Wang M. Improving the performance of protein kinase identification via high dimensional protein–protein interactions and substrate structure data. Molecular Biosystems. 2014;10:694–702. DOI:10.1039/c3mb70462a.

[22] Fan W, Xu X, Shen Y, Feng H, Li A, Wang M. Prediction of protein kinase-specific phosphorylation sites in hierarchical structure using functional information and random forest. Amino Acids. 2014;46:1069–78. DOI:10.1007/s00726-014-1669-3.

[23] Zou L, Wang M, Shen Y, Liao J, Li A, Wang M. PKIS: computational identification of protein kinases for experimentally discovered protein phosphorylation sites. BMC Bioinformatics. 2013;14:247. DOI:10.1186/1471-2105-14-247.

[24] Song C, Ye M, Liu Z, Cheng H, Jiang X, Han G et al. Systematic analysis of protein phosphorylation networks from phosphoproteomic data. Molecular & Cellular Proteomics: MCP. 2012;11:1070–83. DOI:10.1074/mcp.M111.012625.

[25] Xue Y, Liu Z, Cao J, Ma Q, Gao X, Wang Q et al. GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. Protein Engineering, Design & Selection: PEDS. 2011;24:255–60. DOI:10.1093/protein/gzq094.

[26] Xue Y, Ren J, Gao X, Jin C, Wen L, Yao X. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Molecular & Cellular Proteomics: MCP. 2008;7:1598–608. DOI:10.1074/mcp.M700574-MCP200.

[27] Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X. GPS: a comprehensive www server for phosphorylation sites prediction. Nucleic Acids Research. 2005;33:W184–7. DOI: 10.1093/nar/gki393.

[28] Zhou FF, Xue Y, Chen GL, Yao X. GPS: a novel group-based phosphorylation predicting and scoring method. Biochemical and Biophysical Research Communications. 2004;325:1443–8. DOI:10.1016/j.bbrc.2004.11.001.

[29] Lai AC, Nguyen Ba AN, Moses AM. Predicting kinase substrates using conservation of local motif density. Bioinformatics. 2012;28:962–9. DOI:10.1093/bioinformatics/bts060.

[30] Safaei J, Manuch J, Gupta A, Stacho L, Pelech S. Prediction of 492 human protein kinase substrate specificities. Proteome Science. 2011;9 Suppl 1:S6. DOI:10.1186/1477-5956-9-S1-S6.

[31] Safaei J, Manuch J, Gupta A, Stacho L, Pelech S, editors. Prediction of human protein kinase substrate specificities. Prediction of human protein kinase substrate specificities; 2010: IEEE.

[32] Ellis JJ, Kobe B. Predicting protein kinase specificity: Predikin update and performance in the DREAM4 challenge. PloS One. 2011;6:e21169. DOI:10.1371/journal.pone.0021169.

[33] Brinkworth RI, Breinl RA, Kobe B. Structural basis and prediction of substrate specificity in protein serine/threonine kinases. Proceedings of the National Academy of Sciences of the United States of America. 2003;100:74–9. DOI:10.1073/pnas.0134224100.

[34] Saunders NF, Brinkworth RI, Huber T, Kemp BE, Kobe B. Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. BMC Bioinformatics. 2008;9:245. DOI: 10.1186/1471-2105-9-245.

[35] Saunders NF, Kobe B. The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. Nucleic Acids Research. 2008;36:W286–90. DOI:10.1093/nar/gkn279.

[36] Gao J, Xu D. The Musite open-source framework for phosphorylation-site prediction. BMC Bioinformatics. 2010;11(Suppl 12):S9. DOI:10.1186/1471-2105-11-S12-S9.

[37] Gao J, Thelen JJ, Dunker AK, Xu D. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. Molecular & Cellular Proteomics: MCP. 2010;9:2586–600. DOI:10.1074/mcp.M110.001388.

[38] Durek P, Schudoma C, Weckwerth W, Selbig J, Walther D. Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. BMC Bioinformatics. 2009;10:117. DOI:10.1186/1471-2105-10-117.

[39] Miller ML, Blom N. Kinase-specific prediction of protein phosphorylation sites. Methods in Molecular Biology. 2009;527:299–310. DOI:10.1007/978-1-60327-834-8_22.

[40] Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics. 2004;4:1633–49. DOI:10.1002/pmic.200300771.

[41] Dang TH, Van Leemput K, Verschoren A, Laukens K. Prediction of kinase-specific phosphorylation sites using conditional random fields. Bioinformatics. 2008;24:2857–64. DOI:10.1093/bioinformatics/btn546.

[42] Wan J, Kang S, Tang C, Yan J, Ren Y, Liu J et al. Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. Nucleic Acids Research. 2008;36:e22. DOI:10.1093/nar/gkm848.

[43] Li T, Li F, Zhang X. Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. Proteins. 2008;70:404–14. DOI:10.1002/prot.21563.

[44] Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH et al. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic Acids Research. 2007;35:W588–94. DOI:10.1093/nar/gkm322.

[45] Huang HD, Lee TY, Tzeng SW, Horng JT. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Research. 2005;33:W226–9. DOI:10.1093/nar/gki471.

[46] Xue Y, Li A, Wang L, Feng H, Yao X. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. BMC Bioinformatics. 2006;7:163. DOI:10.1186/1471-2105-7-163.

[47] Kim JH, Lee J, Oh B, Kimm K, Koh I. Prediction of phosphorylation sites using SVMs. Bioinformatics. 2004;20:3179–84. DOI:10.1093/bioinformatics/bth382.

[48] Neuberger G, Schneider G, Eisenhaber F. pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase–substrate binding model. Biology Direct. 2007;2:1. DOI:10.1186/1745-6150-2-1.

[49] Joughin BA, Liu C, Lauffenburger DA, Hogue CW, Yaffe MB. Protein kinases display minimal interpositional dependence on substrate sequence: potential implications for the evolution of signalling networks. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2012;367:2574–83. DOI:10.1098/rstb.2012.0010.

[50] Won AP, Garbarino JE, Lim WA. Recruitment interactions can override catalytic interactions in determining the functional identity of a protein kinase. Proceedings of the National Academy of Sciences of the United States of America. 2011;108:9809–14. DOI:10.1073/pnas.1016337108.

[51] Ngo JC, Chakrabarti S, Ding JH, Velazquez-Dones A, Nolen B, Aubol BE et al. Interplay between SRPK and Clk/Sty kinases in phosphorylation of the splicing factor ASF/SF2

is regulated by a docking motif in ASF/SF2. Molecular Cell. 2005;20:77–89. DOI:10.1016/j.molcel.2005.08.025.

[52] Gogl G, Schneider KD, Yeh BJ, Alam N, Nguyen Ba AN, Moses AM et al. The structure of an NDR/LATS Kinase-Mob complex reveals a novel kinase-coactivator system and substrate docking mechanism. PLoS Biology. 2015;13:e1002146. DOI:10.1371/journal.pbio.1002146.

[53] van de Weerdt BC, Littler DR, Klompmaker R, Huseinovic A, Fish A, Perrakis A et al. Polo-box domains confer target specificity to the Polo-like kinase family. Biochimica et Biophysica Acta. 2008;1783:1015–22. DOI:10.1016/j.bbamcr.2008.02.019.

[54] Koivomagi M, Valk E, Venta R, Iofik A, Lepiku M, Morgan DO et al. Dynamics of Cdk1 substrate specificity during the cell cycle. Molecular Cell. 2011;42:610–23. DOI:10.1016/j.molcel.2011.05.016.

[55] Cheng KY, Noble ME, Skamnaki V, Brown NR, Lowe ED, Kontogiannis L et al. The role of the phospho-CDK2/cyclin A recruitment site in substrate recognition. The Journal of Biological Chemistry. 2006;281:23167–79. DOI:10.1074/jbc.M600480200.

[56] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC et al. UCSF Chimera – a visualization system for exploratory research and analysis. Journal of Computational Chemistry. 2004;25:1605–12. DOI:10.1002/jcc.20084.

[57] Mottin M, Souza PC, Skaf MS. Molecular Recognition of PPARgamma by Kinase Cdk5/p25: Insights from a combination of protein–protein docking and adaptive biasing force simulations. The Journal of Physical Chemistry B. 2015;119:8330–9. DOI:10.1021/acs.jpcb.5b04269.

[58] Piserchio A, Ramakrishan V, Wang H, Kaoud TS, Arshava B, Dutta K et al. Structural and dynamic features of F-recruitment site driven substrate phosphorylation by ERK2. Scientific Reports. 2015;5:11127. DOI:10.1038/srep11127.

[59] Whisenant TC, Ho DT, Benz RW, Rogers JS, Kaake RM, Gordon EA et al. Computational prediction and experimental verification of new MAP kinase docking sites and substrates including Gli transcription factors. PLoS Computational Biology. 2010;6. DOI:10.1371/journal.pcbi.1000908.

[60] Gordon EA, Whisenant TC, Zeller M, Kaake RM, Gordon WM, Krotee P et al. Combining docking site and phosphosite predictions to find new substrates: identification of smoothelin-like-2 (SMTNL2) as a c-Jun N-terminal kinase (JNK) substrate. Cellular Signalling. 2013;25:2518–29. DOI:10.1016/j.cellsig.2013.08.004.

[61] Bardwell AJ, Bardwell L. Two hydrophobic residues can determine the specificity of mitogen-activated protein kinase docking interactions. The Journal of Biological Chemistry. 2015;290:26661–74. DOI:10.1074/jbc.M115.691436.

[62] Garai A, Zeke A, Gogl G, Toro I, Fordos F, Blankenburg H et al. Specificity of linear motifs that bind to a common mitogen-activated protein kinase docking groove. Science Signaling. 2012;5:ra74. DOI:10.1126/scisignal.2003004.

[63] Zeke A, Bastys T, Alexa A, Garai A, Meszaros B, Kirsch K et al. Systematic discovery of linear binding motifs targeting an ancient protein interaction surface on MAP kinases. Molecular Systems Biology. 2015;11:837. DOI:10.15252/msb.20156269.

[64] Bibi N, Parveen Z, Rashid S. Identification of potential Plk1 targets in a cell-cycle specific proteome through structural dynamics of kinase and Polo box-mediated interactions. PloS One. 2013;8:e70843. DOI:10.1371/journal.pone.0070843.

[65] Hu J, Neiswinger J, Zhang J, Zhu H, Qian J. Systematic prediction of scaffold proteins reveals new design principles in scaffold-mediated signal transduction. PLoS Computational Biology. 2015;11:e1004508. DOI:10.1371/journal.pcbi.1004508.

[66] Bandyopadhyay S, Chiang CY, Srivastava J, Gersten M, White S, Bell R et al. A human MAP kinase interactome. Nature Methods. 2010;7:801–5. DOI:10.1038/nmeth.1506

[67] Calejo AI, Tasken K. Targeting protein–protein interactions in complexes organized by A kinase anchoring proteins. Frontiers in Pharmacology. 2015;6:192. DOI:10.3389/fphar.2015.00192.

[68] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research. 2015;43:D447–52. DOI:10.1093/nar/gku1003.

[69] Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell. 2015;163:712–23. DOI:10.1016/j.cell.2015.09.053.

[70] Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J et al. The BioPlex Network: a systematic exploration of the human interactome. Cell. 2015;162:425–40. DOI:10.1016/j.cell.2015.06.043.

[71] Varjosalo M, Keskitalo S, Van Drogen A, Nurkkala H, Vichalkovski A, Aebersold R et al. The protein interaction landscape of the human CMGC kinase group. Cell Reports. 2013;3:1306–20. DOI:10.1016/j.celrep.2013.03.027.

[72] Varjosalo M, Sacco R, Stukalov A, van Drogen A, Planyavsky M, Hauri S et al. Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. Nature Methods. 2013;10:307–14. DOI:10.1038/nmeth.2400.

[73] Couzens AL, Knight JD, Kean MJ, Teo G, Weiss A, Dunham WH et al. Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. Science Signaling. 2013;6:rs15. DOI:10.1126/scisignal.2004712.

[74] Hauri S, Wepf A, van Drogen A, Varjosalo M, Tapon N, Aebersold R et al. Interaction proteome of human Hippo signaling: modular control of the co-activator YAP1. Molecular Systems Biology. 2013;9:713. DOI:10.1002/msb.201304750.

[75] Alexander J, Lim D, Joughin BA, Hegemann B, Hutchins JR, Ehrenberger T et al. Spatial exclusivity combined with positive and negative selection of phosphorylation motifs is the basis for context-dependent mitotic signaling. Science Signaling. 2011;4:ra42. DOI:10.1126/scisignal.2001796.

[76] Duarte ML, Pena DA, Nunes Ferraz FA, Berti DA, Paschoal Sobreira TJ, Costa-Junior HM et al. Protein folding creates structure-based, noncontiguous consensus phosphorylation motifs recognized by kinases. Science Signaling. 2014;7:ra105. DOI:10.1126/scisignal.2005412.

[77] Nadzirin N, Gardiner EJ, Willett P, Artymiuk PJ, Firdaus-Raih M. SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. Nucleic Acids Research. 2012;40:W380–6. DOI:10.1093/nar/gks401.

[78] Kumar N, Mohanty D. MODPROPEP: a program for knowledge-based modeling of protein–peptide complexes. Nucleic Acids Research. 2007;35:W549–55. DOI: 10.1093/nar/gkm266.

[79] Kumar N, Mohanty D. Identification of substrates for Ser/Thr kinases using residue-based statistical pair potentials. Bioinformatics. 2010;26:189–97. DOI:10.1093/bioinformatics/btp633.

[80] Chen C, Ha BH, Thevenin AF, Lou HJ, Zhang R, Yip KY et al. Identification of a major determinant for serine-threonine kinase phosphoacceptor specificity. Molecular Cell. 2014;53:140–7. DOI:10.1016/j.molcel.2013.11.013.

[81] Creixell P, Palmeri A, Miller CJ, Lou HJ, Santini CC, Nielsen M et al. Unmasking determinants of specificity in the human kinome. Cell. 2015;163:187–201. DOI:10.1016/j.cell.2015.08.05

IntechOpen