[Frontiers in Bioscience 16, 1261-1275, January 1, 2011]

# From sequence to structural analysis in protein phosphorylation motifs

**Allegra Via[1], Francesca Diella[2, 3], Toby James Gibson[2], Manuela Helmer-Citterich[4]**

*[1]Biocomputing group, Department of Biochemical Science "A. Rossi Fanelli", Sapienza University of Rome, P.le Aldo Moro 5, Rome, Italy, [2]European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg, Germany, [3]Biobyte solutions GmbH, Boxbergring 16, 69126 Heidelberg, Germany,[4]Center for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Italy*

## TABLE OF CONTENTS



## 1. ABSTRACT

Phosphorylation is the most widely studied post-translational modification occurring in cells. While mass spectrometry-based proteomics experiments are uncovering thousands of novel *in vivo* phosphorylation sites, the identification of kinase specificity rules still remains a relatively slow and often inefficacious task. In the last twenty years, many efforts have being devoted to the experimental and computational identification of sequence and structural motifs encoding kinase-substrate interaction key residues and the phosphorylated amino acid itself. In this review, we retrace the road to the discovery of phosphorylation sequence motifs, examine the progresses achieved in the detection of three-dimensional motifs and discuss their importance in the understanding of regulation and de-regulation of many cellular processes.

## 2. INTRODUCTION

Since the discovery of the phosphorylation of vitellin in 1906 (1) many thousands of articles about 'kinases and phosphorylation' have been published, showing that phosphorylation is by far the most widespread and studied post translational modification (PTM) in proteins. This is not surprising since phosphorylation is involved in almost every physiological process, including differentiation, control of metabolism, regulation of organogenesis during development, cytoskeleton remodeling, immunity, cell-cell interaction (2). However it was not before the 1970s that the general significance of protein phosphorylation came to be appreciated. As earlier mentioned by Philip Cohen it would be difficult to find anybody that would disagree with the statement that "the reversible phosphorylation of proteins regulates nearly

**Table 1.** Phosphorylation site databases. Note that, due to lack of resources, the mtcPTM database is no longer maintained

| Database | Organism | pmid | Link |
|---|---|---|---|
| UniProt | All | 12824418 | http://us.expasy.org/sprot/ |
| HPRD | Human | 18988627 | http://www.hprd.org/ |
| Phospho.ELM | Eukaryotes | 17962309 | http://phospho.elm.eu.org/ |
| PhosphoSite | Vertebrata | 15174125 | http://www.phosphosite.org/ |
| PHOSIDA | Human/Mouse/*Drosophila*/others | 17081983 | http://www.phosida.com/ |
| PlantP | Plants | 15308754 | http://plantsp.genomics.purdue.edu/html/ |
| PhosPhAt | Arabidopsis | 19880383 | http://phosphat.mpimp-golm.mpg.de |
| P3DB | Plants | 18931372 | http://www.p3db.org/ |
| Phospho3D | Eukaryotes | 17142231 | http://www.phospho3d.org/ |
| PhosphoPep | Human, *C.elegans, Drosophila, S.cerevisiae* | 19060867 | http://www.phosphopep.org/ |
| PhosphoGRID | *S.cerevisiae* | 20428315 | http://phosphogrid.org/ |
| mtcPTM | Eukaryotes | 17521420 | http://www.mitocheck.org |

every aspect of eukaryotic cell life" (3). As a consequence of this *ubiquitous* role, inappropriate activation of the protein kinases (e.g. by mutation or over expression) will lead to cellular deregulation, i.e. cancer (4).

It has been estimated that at least one-third of the cellular proteins are modified by phosphorylation (5). This finding has been confirmed by recent developments in mass spectrometry (MS) that has allowed the identification and quantification of thousands of *in vivo* phosphorylation sites in several organisms (6-10). These phosphorylation sites (or P-sites) have been collected and classified in several resources - some of which are listed in Table 1 - which represent a primary source of data for further biochemical and bioinformatics investigations.

The key players in the phosphorylation process are the protein kinase enzymes that represent ~2% of all the cellular proteins – in human they number about ~500 (11). Phosphorylation mainly occurs on Serine, Threonine and Tyrosine residues and the kinases have been grouped, according to their residue specificity, into three major families: Ser/Thr protein kinases, Tyr protein kinases and dual-specificity protein kinases. Kinases are pleiotropic enzymes phosphorylating a wide range of protein targets. At the same time they must be sufficiently specific and act only on a defined group of cellular substrates. How can kinases discriminate among a wide spectrum of substrates, which often display different structures and functions? It is nowadays well-known that the selectivity of protein kinases is ensured in two main different ways (12): 1) site specificity, i.e. the capability of a kinase to specifically recognize a number of features surrounding the phosphorylation site in its target substrates and (2) recruitment; we refer to recruitment as to any process that increases the likelihood that a kinase encounters its substrate (s). Recruitment can be achieved through different strategies such as kinase-substrate spatial proximity, docking interactions, modular domain mediated interactions or a cooperation of these mechanisms.

Spatial proximity can occur through compartimentalisation, i.e. a kinase and its substrate (s) are targeted, after synthesis, to the same cellular compartment, or by binding to the same scaffold protein. Docking interactions are interactions involving a binding region of the kinase that is distinct from that of the active site and that recognizes, in a specific fashion, a motif of the substrate that is separate from the phosphorylation site region. Modular recognition domains are dedicated interaction domains, such as SH2 and SH3 domains, that are separate from the catalytic domain and that specifically recognize substrate peptide motifs that do not belong to the phosphorylation region.

Nowadays, we appreciate that evolution assured phosphorylation specificity through a strategic cooperation of site specificity and recruitment, sometime favoring the former, sometime the latter, and sometime rewarding an intermediate strategy between the two (12).

In this review we will focus on our current understanding of the former of the aforementioned selectivity mechanism, i.e. we will delve into the idea that substrate recognition is based on distinct local amino acid patterns, called phospho-motifs, surrounding the local phosphorylation site and shared by globally different proteins targeted by a given kinase. Many efforts have been devoted to the detection and prediction of phospho-motifs and, as a result, *consensus* sequences and specificity rules have been established for some kinases and kinase groups. Nevertheless, the identification of such motifs is a difficult task that is still in progress, notably as far as three-dimensional motifs are concerned. In this regard, a clear definition of 3D phospho-motifs is still elusive, as we discuss in this review.

Here, we investigate phospho-motifs from a sequence and a structural perspective and retrace the experimental and computational milestones along the route to their identification, discussing step by step the reasons why it proves so difficult and in some cases even impossible to recognize/distinguish phosphorylatable residues and phosphorylation fingerprints encoding the protein kinase specificity.

The paper is organized as follows: first, we focus on phosphorylation linear motifs and discuss their identification in an historical perspective. Then, we try to formulate a suitable definition for 3D phospho-motifs - are they precise spatial arrangements of amino acids recognized by a specific kinase or are we only able to provide more blurred definitions - and review the results of the authors aimed at identifying local structural

characteristics that are specific for phosphorylation sites or classes of phosphorylation sites. In the last part of the review, we discuss how structural information gathered from phosphorylation sites and from their surrounding regions, is used to improve the performance of phosphorylation site predictors.

## 3. SUBSTRATE SPECIFICITY AND PHOSPHORYLATION LINEAR MOTIFS

Considering the broad distribution of the phosphorylation processes, it is not surprising that many efforts have been devoted to the identification of kinase substrates. However experimental methods are costly and time-consuming, therefore computational approaches have been used in order to minimize the amount of experimental work.

Identifying how the diverse kinases recognize their diverse substrates has become a major challenge in the field of phosphorylation. Early attempts to identify and classify the consensus phospho-motifs did not give satisfying results (13, 14) because at the time too few phosphorylation sites had been identified and, moreover, no three-dimensional structure of protein kinase/substrate complexes was available.

However, key points were identified regarding the specificity of recognition, which are still valid: a) the presence of a *consensus* sequence is necessary for its recognition as substrate by a kinase; b) the specificity is encoded in the residues surrounding the phospho-acceptor site; c) not all the amino acids in the surrounding peptide have equal weight in determining the recognition by the kinase (15).

A more mature work was published by Pinna and Ruzzene in 1996 (16) – based on the wealth of information derived from the first three-dimensional structure of a protein kinase (17) and the identification of the first protein-protein interaction modules, such as the Src homology 2 (SH2), Src homology 3 (SH3) and phosphotyrosine-binding (PTB) domains (18, 19). For the first time, a model for kinase regulation and for the specific recognition of the substrate could be proposed (16).

In particular, the development of new experimental strategies based on degenerated peptide libraries fostered the notion that the short regions (8-10 amino acids) surrounding the phospho-acceptor residue (20, 21) are crucial in substrate recognition. Nevertheless, the peptide library approach has some limitations since it does not take into account important elements, such as the role of the structural context and the influence of distant residues located on the same polypeptide chain.

Another issue is that some kinases show a wider substrate specificity than others: comparing the peptide specificity deduced from the known substrates of different kinases, Kreegipuu *et al.* (21) suggested that Ser/Thr kinases are generally more specific than Tyr kinases. They also suggested that Ser/Thr kinases can be divided into three groups, according to their substrate specificity: i) the proline-directed kinases (e.g. MAPK) that require a proline in position +1 (where the position 0 is the one occupied by the phospho-residue); ii) the acidophilic kinases (e.g. CKII) that prefer acidic amino acids surrounding the phospho-residue; and iii) the basophilic kinases (e.g. PKA) that favor basic amino acids.

The results obtained from library screening experiments enable the construction of *consensus* motifs and position specific scoring matrices (PSSM). These can be used either to predict novel substrates or to infer the kinase responsible for the phosphorylation of a known substrate (22, 23).

Phosphoproteomic experiments, whereby phosphorylated amino acids are identified in purified samples with mass spectrometry techniques, provide a growing and potentially valuable source of data for understanding the mechanisms of protein regulation. This enormous amount of data has spurred the development of algorithms for the *de novo* discovery of motifs in unaligned sequences. Schwartz and Gygi made the first relevant attempt to extract motifs from a mixture of phosphorylated peptides: the algorithm relies on the intrinsic alignment of phospho-residues and the extraction of motifs through iterative comparison with a dynamic statistical background (24). The Motif-X algorithm has been applied in several proteomic studies and represents a promising approach for the discovery of sequence specificities of uncharacterized kinases or phosphatases in signaling pathways (9, 25, 26).

Since closely related protein kinases usually share very similar phospho-site motifs and, in addition, kinases belonging to more divergent families can share a minimal *consensus* (i.e. MAPKs, HIPKs and CDKs phosphorylate substrates on Ser/Thr-Pro), peptide recognition cannot be the only force driving the kinase-substrate specificity. Protein kinases have adopted additional mechanisms for the selective recruitment of their substrates: It is likely that *in vivo* the intrinsic specificity of the catalytic domain cooperates with other factors which raise the effective local concentration of protein substrates. Thus substrate specificity could also be modulated by additional interactions, distal from the phospho-site, involving domains, scaffolds or docking sites.

Tyrosine kinases tend to use non-catalytic modules, e.g. SH2 and SH3 domains, to bind target peptides in protein adaptors or substrates in order to position themselves near the potential phospho-acceptor sites. In the case of Src kinase the presence of an SH2 domain-binding site in a substrate has been shown to increase the rate of phosphorylation by 10-fold (27).

Many Ser/Thr kinases have evolved an analogous strategy to increase the enzyme-substrate specificity, i.e. the utilization of direct kinase domain docking interactions, which are capable of establishing specific connections via small peptide motifs. The role that the docking sites have in enhancing the substrate specificity has been well studied in the MAPKs system (28). These sites are ubiquitous on the MAPK substrates and are

usually found 50-100 residues away from the phospho-acceptor site. In contrast to Tyr kinases, additional modular domains tend to be less common in Ser/Thr kinases (though there are striking exceptions such as the kinase domain in the giant protein titin).

Another clear example of the importance of the spatial and biological regulation is provided by the scaffolding A-kinase anchoring proteins (AKAPs) that bind almost simultaneously the regulatory subunit of the inactive PKA and a subcellular anchor that juxtapose PKA to its substrates (29).

Another important factor is the localization in distinct subcellular compartments, which can improve the specificity by regulating accessibility of the kinases to their substrates. Already in 1996 Faux and Scott had reported that discrete localization would aid in conferring specificity to the mechanisms of protein phosphorylation (30). However it was not until recently that this kind of information has been integrated into phospho-predictor resources. Taking advantage of the recently increased information on protein-protein interaction networks, Linding *et al.* have developed NetworKIN, an algorithm that combines predicted sites of known phosphorylation motifs (from ~112 human kinases) with protein interaction data obtained from various literature sources and pathway databases (31). With this approach not only is the confidence of the kinase-substrate predictions increased, but it is also possible to assign sites to specific kinases rather than to a whole group with similar specificities. It can be expected that, as the knowledge of phosphorylation motifs and interaction data increases, such approaches will become even more powerful.

## 4. STRUCTURAL FEATURES OF PHOSPHORYLATION MOTIFS

### 4.1. Can phosphorylation 3D motifs be defined?

In order to discuss the features of three-dimensional (3D) phosphorylation motifs, we first need to understand if a definition for such motifs can be found. Therefore, the question we initially want to try to answer here is the following: is it possible to detect a spatial arrangement of amino acids on the three-dimensional (3D) structure of a phosphorylatable protein, representing a unique and conserved feature that a kinase or a family of kinases specifically recognizes in the phosphorylation event?

Since protein kinases phosphorylate a wide spectrum of protein substrates with varying structures and often exerting different functions, it is reasonable to hypothesize that the majority of kinases recognize a number of local structural features surrounding the phospho-residue in the substrate rather than the substrate protein in its integrity. As discussed above, in several cases actual site specificity seems to be entirely located in the primary structure of the phosphoacceptor site and it is based on a local *consensus* or *motif* shared by all the proteins targeted by a given kinase. This conjecture is also supported by the experimental observation that a wide number of protein kinases was able to phosphorylate small peptides mimicking the phospho sites, with kinetics comparable to those of the intact protein substrate (16). However, as more and more protein structures have been determined and targeted experiments carried out in the last years, it has been observed that sequence *consensus* is not sufficient for kinase recognition, and that an appropriate structural context is necessary for the interaction to occur. In these cases the partner kinases are able to phosphorylate peptides reproducing the primary structure around the phospho residue with much less efficiency than the intact proteins(16).

Both early and recent studies report that many phosphorylation sequences show a preference for an observed or predicted beta-turn or for a loop conformation and, more in general, that any exposed and flexible structure represents a feature that facilitates phosphorylation (32-38). These results are consistent with the idea that a phosphorylation event would be better assisted if the P-site is presented to the catalytic domain of a kinase in an exposed and flexible structural form. In 1991, Sowadski and collaborators (39) published the crystal structure of the PKA-PKI peptide complex, which exhibited the *consensus* sequence hosting the P-site accommodated in an extended conformation inside a cleft between the two lobes of the kinase. This evidence showed for the first time that the flexibility of the phosphorylation motif represents, as for many types of interaction, an unambiguously advantageous feature for the interaction.

The P-site tendency to reside in accessible and flexible regions of substrate proteins was specifically investigated by Dunker and collaborators (40, 41), who analyzed and exploited the importance of structural disorder for protein phosphorylation.

These authors used a robust procedure to build datasets of reliable positive (P) and negative (NP) phosphorylation sites by i) extracting sequences of (-12,+12) residues centered at (annotated and unannotated, respectively) S, T and Y sites from eukaryotic proteins and subsequently ii) applying a number of filters in order to discard site redundancy and other potential biases. Then, they compared the P and NP datasets for a set of properties of the amino acids surrounding each site and determined which residues were enriched or depleted at specific positions. The properties analysed range from site accessibility, to site hydrophobicity, charge and flexibility. The most prominent result is that the distribution of order- and disorder-promoting amino acids around P-sites is similar to the one that is characteristic for intrinsically disordered protein regions. Moreover, the authors found that the distribution of sequence complexity for P-sites, as defined by Wootton (42), is very similar to that for disordered segments, whereas the complexity distribution for NP-sites is similar to that for ordered globular segments. The authors furthermore predicted that only 1% of serines and 1% of threonines in the ordered regions could be phosphorylated, whereas tyrosines appear to be phosphorylatable both in intrinsically disordered and surface exposed ordered states. These predictions are

supported by several examples indicating that phosphorylation commonly occurs within intrinsically disordered protein regions; for example, the crystal structure of some protein kinase-substrate complexes shows bound substrates and inhibitor peptides having essentially no intra-chain backbone hydrogen bonding while having extensive hydrogen bonding between their backbones and the backbones or side chains of their kinase partner (43-46): The formation of these hydrogen bonds would not be possible if the P-sites were located within ordered regions. In other words, a disordered conformation can guarantee that substrates have available backbone hydrogen bonding potential whilst this availability is incompatible with an ordered structure. It must be noticed that the authors describe also some counter examples, i.e. crystal complexes showing phosphorylated residues occurring in ordered protein regions and suggest three possible reasons explaining them: a) the disorder is not always required; b) the P-site undergoes a transition to disorder just prior to association with the kinase and c) the observed crystal structures are artifacts.

These results indicate that phosphorylation sites are preferentially located in unstructured regions of proteins - suggesting a limited relevance of any structurally well-defined binding motif for the specific recognition of kinases. Moreover, as mentioned above, several experiments demonstrate that the specificity of kinases is often found to be encoded in substrate sequence *consensi* rather than in structural determinants (16, 20, 47-53). Nevertheless, it is plausible that what indeed a kinase recognizes is the three-dimensional conformation of the polypeptide at the P-site, and not only the primary structure (16, 54, 55), and that kinase specificity is dependent on the P-site tertiary structure. This observation is further supported by recent systematic structural analysis of P-sites, based on structural information predicted from thousands of *in vivo* P-sites identified by mass spectrometry-based proteomics (37) and from P-sites retrieved from literature and databases (33) and by novel studies investigating the possibility that many P-sites in a phosphoproteome are non-functional (56, 57). The non-functionality of several P-sites might result from the off-target activity of protein kinases. In particular, Landry *et al.* (57) suggested, in order to identify functional P-sites, exploiting the information about their conservation across related species and their presence within *consensus* kinase recognition motifs, and showed that disordered regions are particularly rich in supposedly non-functional P-sites. The abundance of "noisy phosphorylation events" in disordered regions suggests that kinases recognize more specifically phosphorylatable sites in structured regions and within kinase recognition motifs than in unstructured regions, outside of such motifs. However, this is quite a controversial argument since site conservation is only a way to estimate the importance of a phosphorylation site. Linding and collaborators claimed that non-conserved P-sites could be mediating species-specific cellular functions and reported few examples where non-positionally conserved sites can still be functional, e.g. the position of Ser46 in the human p53 (58). In addition, multiple phosphorylaton sites on a protein might have similar function, e.g. modulating protein interactions, or might be crucial for the molecular flexibility. In this regard, Figure 1A reports the interesting case of the Phospho.ELM BLAST (59) result for SRRM2. SRRM2 (UniprotKB/SwissProt: Q9UQ35) is part of pre- and post-splicing multiprotein mRNP complexes and has been found to contain huge amounts of phosphorylation sites. All of these P-sites have been detected by means of phosphoproteomics experiments (e.g. (60, 61)) and SRRM2 is predicted to be natively disordered, as shown in Figure 1B. It is possible that a minority of SRRM2 P-sites are functional only in specific context and, in some cases, even the result of experimental artifacts. In these cases, we expect that they do not carry recognition specificity features and should be carefully considered in procedures aimed at building phosphorylation sequence and/or structure *consensi*. Additionally, as for 3D motifs, disordered regions are intrinsically unstructured by definition and cannot be used to study 3D determinants of kinase specificity.

Even though well-defined 3D recognition motifs – defined as recurring spatial arrangements of specific sets of amino acids - are yet unrecognized, several structural features have been identified that characterize phosphorylation sites and likely promote kinase-substrate interaction. More specifically, on the basis of the structural analyses carried out until now, we cannot in general claim that "phosphorylation 3D *consensi*" exist, as we can do for P-site sequence *consensi*. Consistently with this observation, Shaw and collaborators (62) showed that the specificity of interaction of AGC and CAMK kinases is mediated in part by the disfavor, rather than the preference, for a specific amino acid positioned nearby the P-site. More in detail, they observed that the occurrence of a proline at the P-site+1 position prevented AGC and CAMK kinases from phosphorylating substrates that can be phosphorylated by proline-directed kinases. It should be noted that this mechanism might indeed be reasonably common when a substrate protein is phosphorylated at two or more P-sites by distinct kinases and, therefore, precision in phosphorylation is particularly critical.

However, it has been observed that, in a number of cases, P-sites assume precise structural conformations and are surrounded by specific 3D arrangements of residues (63, 64). Such structural environments define 3D signature motifs, i.e. local spatial amino acid distributions that may contribute to, or even determine, the specificity for a kinase or kinase family.

With regard to structural phosphorylation motifs, in principle we can envisage three scenarios: i) the 'motif' consists of coarse (charge, secondary structure, protein surface, etc.) preferences of the amino acids surrounding in space the P-site, which are not necessarily significantly distinguishable from the corresponding preferences of a background distribution of residues surrounding any non-phosphorylated serine, threonine or tyrosine; ii) the motif consists of a set of local properties (charge, secondary structure, solvent accessibility, etc) displaying favored values in the vicinity of P-sites, yet this cannot be related to

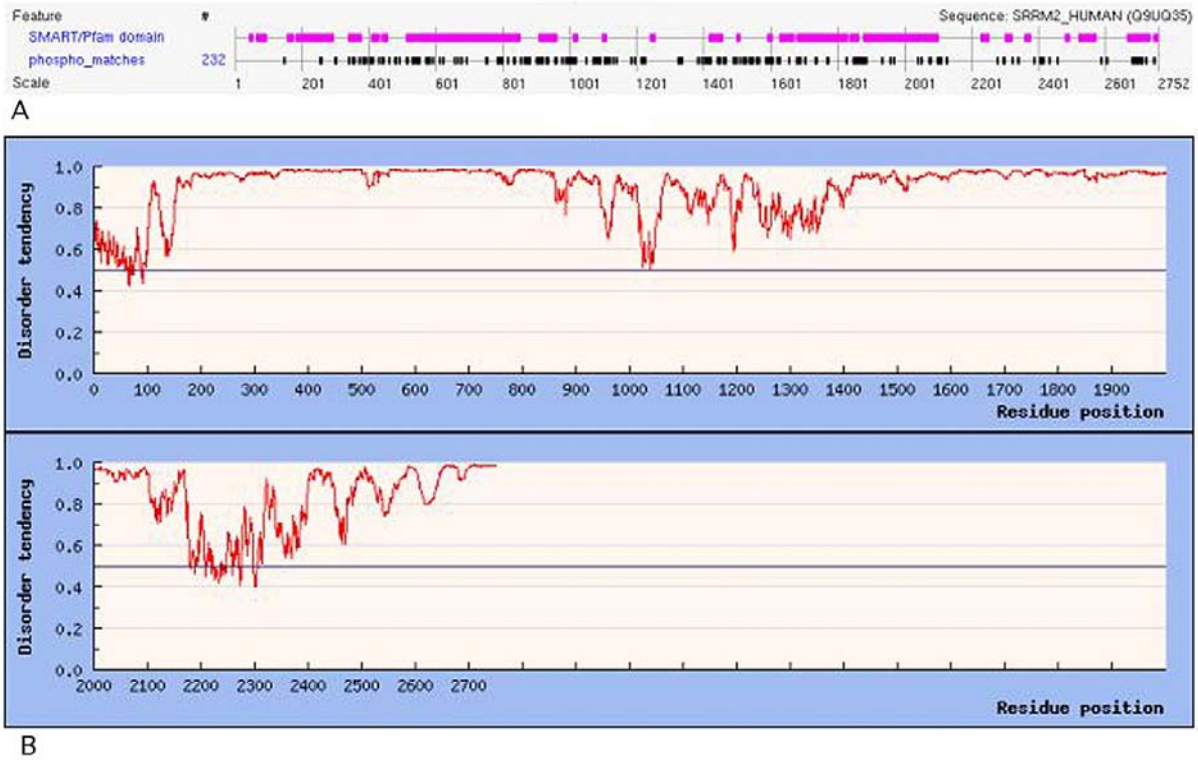


**Figure 1**. a) Phospho.elm BLAST result for SRRM2 protein (UniprotKB/SwissProt Q9UQ35) showing multiple phosphorylation events occurring in a protein that is apparently totally natively disordered, as predicted by the IUPRED (15955779) disorder predictor; b) Plot results of the IUPRED predictor for SRRM2. Residues with a score above 0.5 can be regarded as disordered.

a detectable and conserved arrangement of amino acids in space; iii) the motif consists of a specific and conserved arrangement of residues in space exhibiting a definite pattern of biochemical features. Also, a '3D motif' might be found to be peculiar of phosphorylation sites as a whole (low specificity), of specific classes of acceptor sites (Ser, Thr, Tyr) (mild specificity), or even of sets of P-sites phosphorylated by specific kinases or kinase families (pronounced specificity).

### 4.2. Characterizing the P-site 3D environment (or signature motifs)

In order to detect P-site 3D signature motifs, one fundamental step consists of studying and characterizing the local 3D environment surrounding P-sites and identifying structural features that are unique to such environment.

Fan and Zhang (63) carried out a study in 2005 aimed at exploring the spatial microenvironment of protein phosphorylated sites and were able to identify some significant properties specific for such sites. These authors fetched the P-sites from the Phospho.ELM database (59); a negative set was constructed by considering all the Ser, Thr, and Tyr residues extracted from the Phospho.ELM sequences, which were not annotated as experimentally verified P-sites. The subsets of positive and negative sites concurrently reported in the PDB (65) were taken as the final datasets. A microenvironment was defined as the shell around a site at a distance range of 2-16 angstrom accumulating the spatial distribution of 80 biophysical and biochemical properties. The property distributions for P-sites and negative sites were subsequently statistically compared using a standard nonparametric test of significance. Only properties whose distributions significantly differed in the two datasets were retained and discussed. The structural descriptors ranged from atom-based properties, such as atom type and charge, to chemical group, residue and secondary structure-based properties. Also B-factor, mobility and solvent accessibility were taken into account.

The major result of Fan and Zhang is that they detect an impoverishment of some property around P-sites: More specifically, the serine acceptor site microenvironment is depleted in Ile and Phe residues and, more in general, nonpolar and hydrophobic amino acids. Similarly, the Threonine acceptor site microenvironment is poor in Val residue and hydrophobic and non-polar residues. These results are consistent with the report that several protein kinases are able to phosphorylate their substrates in both Ser and Thr acceptor sites. The Tyr acceptor site microenvironment is characterized by only one enriched property, namely the charge, and is depleted in Cys and Pro, which are neutral residues. These findings were already known from studies based on the protein sequences flanking the P-sites. Other findings, such as the absence of some specific atom types and a smaller van der

Waals volume around the Thr acceptor sites, are novel. Overall, the authors concluded that their analysis highlights the depletion of some properties rather than the enrichment pointed out by sequence-based methods; they propose that deficiency rather than enrichment of some properties in P-site microenvironments has much more effect on protein phosphorylation. These conclusions both support and are supported by the Shaw and collaborators (62) findings and may account for the fine-tuning mechanism of kinase specificity when multiple P-sites are present in the same substrate protein.

Another related study, focusing on charge environments, has recently been published by Kitchen and co-workers (64). The authors point out that only about one third of the phosphorylated protein structures in their dataset display particularly favorable charge interactions at the phosphorylation site. A significant fraction of this third of P-sites, many of which are hosted within kinases, appear to be electrostatically stabilized largely through interactions with basic side chains, even though only minor or no differences were found for the majority of P-sites compared to background Ser, Thr and Tyr residues. This approach does not distinguish between Ser, Thr and Tyr P-sites and describes only general charge propensities of a small subset of P-site environments; therefore it cannot be used in order to identify a specific charge behaviour and conserved electrostatic interactions characterizing the 3D surroundings of P-sites.

Durek and collaborators, in a recent paper (66), attempt to characterize 3D-signature P-site motifs and use them to improve the prediction of P-sites. The main hypothesis of this work is that the analysis of P-sites across all kinase families might impair the identification of motifs that are specific for particular kinases or kinase classes.

In this regard, we want to point out that, in the case of phosphorylation *sequence* motifs (see above), kinase-family specific motifs are much more informative than motifs detected irrespective of kinase-family. This observation is also supported by the greater success rate of kinase-family specific prediction methods compared with kinase-agnostic methods. The work of Durek *et al.* is the first where these principles, well-established in the case of phosphorylation sequence motifs, are used as a criterion for detecting P-site structural signatures.

These authors define Ser/Thr/Tyr phosphorylation sequence motifs as the 13-mer peptides hosting, in their middle, phosphorylation sites obtained from the Phospho.ELM database. By exact matching phosphorylation sequence motifs onto PDB structures and discarding redundancy, they obtained a final dataset of 750 structural phosphorylation motifs, for 307 of which the kinase is known. They also created a set of *bona fide* non–phosphorylation sites. The analysis of solvent accessibility, secondary structure and B-factor (which is a measure of local structural rigidity) showed that P-sites have a tendency to be more exposed and more frequently in random coil regions and less in alpha-helical or beta-strand regions than their unphosphorylated counter-parts. Moreover, the analysis of B-factor distributions highlighted that P-sites are found more often than non-P-sites in flexible regions. These results are in agreement with those of other structural studies of phosphorylation sites.

In order to identify P-site spatial signatures, the authors studied enrichment and depletion of each amino acid type within distances from 2 to 10 angstrom from central P-sites. Interestingly, the amino acid types were counted within radial distances and displayed in radial-radial cumulative propensity plots.

When looking at the P-site 3D surroundings irrespectively of their association with a kinase family, no significant differences with the set of unphosphorylated sites were found; in other words, Durek and collaborators could not identify Ser/Thr/Tyr phosphorylation 3D signatures. Interestingly, they found that, in many cases, phosphorylation 3D motifs (i.e. enrichment and/or depletion of specific amino acid types in the 3D region surrounding the P-site) are detectable for Ser P-sites when studying kinase families individually; for Thr and Tyr kinases the analysis was not possible due to the lack of kinase-target pairs structural data. The inclusion of this 3D information in a Support Vector Machine predictor showed a performance gain compared to the use of sequence information alone.

There are two resources that attempt to define 3D phosphorylation motifs as a specific and conserved arrangement of residues in space, encoding the determinants of kinase specificity. One is the Phospho3D database (67), housing 701 P-site-containing protein structures at the time of publication, and the other is the program PREDIKIN (68), that performs automated predictions of protein Ser/Thr kinase optimal substrates.

Phospho3D stores structural information mapped from the experimentally verified phosphorylation instances collected in the phospho.ELM database. The database is enriched with structural information and annotations at the residue level and also collects the results of a large-scale structural comparison procedure providing clues for the identification of new putative phosphorylation sites. Zanzoni *et al.* define, for each P-site, a 3D *zone* as the set of residues whose distance from the P-site does not exceed 12 angstrom. The *zones* are annotated at the residue level with solvent accessibility value, secondary structure assignment and residue conservation as from the Consurf-HSSP database (69). In addition, for each *zone*, the database contains the results of a large-scale local structural comparison, based on structural and biochemical similarity criteria, versus a representative dataset of PDB protein chains from eukaryotic organisms. The idea underlying the 3D *zones* is the one that better matches the idea of P-site 3D motifs intended as specific 3D arrangements of residues exhibiting definite biochemical features. The authors of Phospho3D do not report whether the large-scale local structural comparison made it possible to identify conserved 3D motifs or not. A new enhanced release of Phospho3D is in preparation.

The P-site predictor PREDIKIN implements a set of rules extracted on the basis of the analysis of the crystal structures of protein Ser/Thr kinase-peptide complexes. As part of the analysis, the 3D structures of cAMP-dependent protein kinase A (PKA), phosphorylase kinase (PHK), and cyclin-dependent kinase (CDK) 2 with bound substrate peptides were studied to find significant contacts between the catalytic domain and the side-chains of the peptide. First of all, enzyme residues contacting the side-chains of the residues surrounding the P-site were identified. Subsequently the complementarity with the six (-3 to +3) positions surrounding the substrate P-site were analyzed in terms of size, polarity, charge, and hydrogen-bonding potential. The inspection of protein-peptide crystal complexes allowed the authors to make some observations, such as that all protein Ser/Thr kinases adopt a similar fold in the binding cleft and bind substrates in a similar extended conformation, and that residues at analogous positions in the kinase pocket bind the same substrate residue (s). Based on these assumptions and on the identification of kinase-peptide complementary positions, interaction rules were extracted. An example rule is: "For kinases in the AGC and CaMK groups, Glu or Asp in position 6 of the enzyme binding pocket results in a preference for Arg or Lys at position -3 of the substrate". Hence, Brinkworth *et al.* were able to identify P-site key residue features and positions in space, that are specifically related to key residues and positions in the binding pocket of the associated kinases. It should be noticed that they report several important exceptions to their rules, such as: a) some protein kinases are not very specific; b) the specificity may sometimes depend on substrate residues outside the range (-3,+3); c) the specificity might be affected by kinase residues that can in turn be phosphorylated; d) P-sites can be surrounded by other phosphorylation sites; e) some residues in the substrate (-3,+3) peptide sequence can make few or even no contacts with the kinase, etc.

Finally, it is worth mentioning a tool, developed in the group of Rychlewski (70), for the molecular modeling of the local structure around phosphorylation sites. The paper describes essentially a predictor of P-sites targeted by PKA and PKC kinases, and the reason for quoting it here is that it introduces a new 3D representation for phosphorylation motifs. The authors collected annotated phosphorylation sites from the Swiss-Prot database (71) and built a database of experimentally determined structures of backbone segments around P-sites using the PSI-blast server running on the PDB database (http://bioinfo.pl/). These local structure segments (LLSs) are then described using the symbolic Baker representation (72), consisting of 11 symbols, each constrained to certain regions of backbone dihedral angles. The analysis of P-site sequence and structure surroundings, showed a clear difference between the sequence composition of PKA and PKB P-site motifs, but much less significant differences were found between the local structures of the two types. Most of the P-sites were found in unstructured parts of proteins, whose coordinates are missing in the crystal structure.

### 4.3. Structural features identified in systematic structural analyses of phosphorylation sites

The recent significantly increased number of P-sites experimentally determined by technologies such as mass spectrometry-based proteomics, with concurrently available 3D structures of the related proteins, is giving rise to systematic structural analyses of P-sites and their spatial surrounding regions.

In 2007, Jimenez and co-workers (33) published the mtcPTM database (see Table 1), an online repository of 13,116 human and 8,889 mouse P-sites (http://www.mitocheck.org/cgi-bin/mtcPTM/search) retrieved from literature, protein annotation, and other databases. The database contains also atomic models for a high number of P-sites. These models have been automatically built by homology to experimentally determined structures using a conservative procedure in order to minimize modeling errors and maximize the reliability of structural data. The authors also report a systematic structural analysis of 264 serine/threonine and 219 tyrosine P-sites, based on 324 non-redundant three-dimensional (3D) models of human and mouse proteins stored in the database. Interestingly, only 10% of these P-sites were found in structurally defined regions, suggesting that P-sites tend to be hosted in flexible, unstructured segments and in linkers between domains. Notice that these results are in agreement with those of Iakoucheva and collaborators (41) mentioned above. In particular, the authors found that 30% of phosphorylatable serine/threonine and 10% of tyrosines of their dataset belong to unstructured terminal regions preceding or following globular domains and were able to depict three different structural contexts hosting P-sites in these tails: 1) the P-site-hosting region is an important part of the interface of interaction with other molecules; 2) the P-site is in the short linker joining adjacent domains and 3) the P-site is in long and unstructured *termini* relatively far away from the globular domain. Moreover, they report that the side chains of phosphorylated amino acids within domains tend to be – as expected – more exposed than the average for side chains of serine/threonine or tyrosine and that this tendency is more marked for phosphoserine/threonine than for phosphotyrosines. Interestingly, around 15% of all P-sites exhibit less than 10% accessibility of their side chains in the phosphorylatable state of the protein; in this regard, the manuscript covers a detailed analysis and several examples of these buried P-sites, and illustrates hypotheses on how the phosphorylation of buried residues may affect functional sites, trigger changes in the relative positioning between domains, or cause structural instability and therefore structural rearrangements, including local or even total unfolding. As for secondary structure propensity, Jimenez *et al.* report that serine and threonine P-sites (within domains) show a marginal preference for loops whereas phosphotyrosines do not belong to any particular secondary structure context.

In 2006, Olsen and collaborators (7) presented PHOSIDA, a phosphorylation site database collecting thousands of high-confidence *in vivo* P-sites identified by mass spectrometry-based proteomics in various eukaryotic

and prokaryotic organisms. Interestingly, the database integrates time course data in response to growth factor stimulation, thus providing quantitative data on the relative level of phosphorylation. Additionally, the authors report a structural study of the phosphoproteome stored in PHOSIDA. In this regard, the authors predicted solvent accessibility and secondary structure (using the SABLE 2.0 program (73)) for sets of phosphorylated and non-phosphorylated serines, threonines and tyrosines taken from phosphoproteins; the analysis of these predicted values showed that the accessibilities of phosphoserine, phosphothreonine and phosphotyrosines are significantly higher than non-phosphorylated serine, threonines and tyrosines and that P-sites are largely localized in hinges and loops (93.0% of phosphoserines, 88.5% of phosphothreonines and 67.3% of phosphotyrosines, whereas non-P-sites have a significantly lower tendency to be located in these regions. Furthermore, the *in vivo* phosphorylation sites were mapped onto three-dimensional coordinates taken from the PDB, resulting in 26 phosphogroups in 16 different structures. The analyses of accessibility and secondary structure of these phosphogroups, which were assigned using DSSP (74), highlighted that they are always hosted in highly accessible parts of the proteins and that, in all but one case, they are found in flexible regions.

## 5. STRUCTURAL INFORMATION IN PHOSPHORYLATION SITE PREDICTORS

The majority of computational methods for P-site prediction are based on machine learning approaches. The descriptions of prediction methodologies and performances are outside the scope of this review. However we want to briefly examine what type of structural information some P-site predictors use and how, and why, in most cases, it entails an increase in performance.

As described in the previous section, results reported by many authors are consistent with the idea that P-site 3D motifs, intended as specifically recognizable and conserved arrangements of residues in the space, do not exist, but that it is possible to identify structural preferences of P-sites and P-site-hosting local regions, in terms of solvent accessibility, secondary structure and structural flexibility (e.g. B-factor). These local structural preferences have been used by some authors for P-site prediction or to increase the performance of sequence-based P-site predictors. Notice that when structural information is also included in training the method, predictions are nevertheless carried out on protein sequences.

The Brunak group was the first, in 1999, that included structural information in a P-site prediction method (75). The local structures of P-sites were integrated in the form of *predicted* local contact maps of peptide fragments centered on the acceptor residues. Given that the P-site prediction using 3D information was itself based on a prediction, the sensitivity turned out to be impressively high, even though the specificity was rather low. In any case, the authors observed that their sequence-based network was overall performing better than the structure-based one on the same dataset.

As already described in the previous section, Brinkworth *et al.* built PREDIKIN, a predictor of Ser/Thr kinase substrates, which takes only the amino acid sequence of a protein kinase as input but applies a set of specificity rules extracted from the crystal structures of kinases and substrates. A new version of PREDIKIN was published in 2008 together with PredikinDB, a database of P-sites that links substrates to kinase sequences (76).

Plewczynski *et al.* in 2005 (70, 77) published two methods based on SVMs (Support Vector Machines) for P-site prediction making use of 3D information. The authors claim that the incorporation of structural information into the description of P-site neighborhood implies a significant improvement in the accuracy of predictions, suggesting that structural information should be added whenever possible to sequence information for more effective predictions of P-sites. Gnad *et al.* (32) are of like mind. These authors built PHOSIDA, a P-site predictor based on a SVM, and applied it to various sets of different types of P-site 1D and 3D descriptors, ranging from the primary sequence surrounding the P-site only, to an assortment of properties including P-site secondary structure, its predicted accessibility, and evolutionary conservation. They found that the performance of the prediction based on the sequence only was already very high but increased when including the structural and conservation information.

Durek *et al.* (66) also applied SVMs to P-site prediction using both sequence information only and subsequently adding P-site 3D-context information, such as secondary and tertiary structure preferences, solvent accessibility, structural disorder indices and others. The incorporation of 3D-context information implied a small but consistent performance improvement compared to the SVM using sequence information only.

From these results, it appears that P-site structural context conveys significant information content. This observation also stems from the fact that what a kinase actually recognizes is the three-dimensional shape of the substrate rather than its sequence. It should be noticed that nearly all authors mentioned above complain about the lack of experimental structural data, which causes poor statistics and low predictor performances. One reason for the paucity of determined structures at P-sites, is due to the fact that they often occur in the unstructured region of proteins (41), thus making it impossible to clearly determine P-site structural descriptors. The tendency of P-sites of being in unstructured regions has nevertheless been used by Dunker and his collaborators (40) to build DISPHOS, a phosphorylation site predictor incorporating disorder information to improve the ability to discriminate between P-sites and non P-sites.A list of major P-site predictors is reported in Table 2.

**Table 2.** Phosphorylation site predictors

| Resource | Methods | Training dataset | Accuracy | Kinase predicted | Link | PMID | Since |
|---|---|---|---|---|---|---|---|
| Scansite | Position-specific scoring matrix (PSSM) | Oriented peptide library and phage display | | yes | http://scansite.mit.edu | 12824383 | 2001 |
| NetPhos | Artificial neural network (ANN) | Experimentally verified phosphorylation sites | | no | http://www.cbs.dtu.dk/services/NetPhos | 10600390 | 1999 |
| NetPhosK | Artificial neural network (ANN) | Experimentally verified phosphorylation sites | Specificity (0.90) and sensitivity (0.84) | yes | http://www.cbs.dtu.dk/services/NetPhosK | 15174133 | 2004 |
| NetPhosYeast | Artificial neural network (ANN) | ~1140 experimentally verified phosphorylation sites | | no | http://www.cbs.dtu.dk/services/NetPhosYeast | 17282998 | 2007 |
| GPS | Markov Cluster Algorithm (MLC) | ~ (13,000) experimentally verified phosphorylation sites | | yes | http://bioinformatics.lcd-ustc.org/gps2 | 15980451, 18463090 | 2005 |
| PPSP | Bayesian decision theory (BDT) | Experimentally verified phosphorylation sites | | yes | http://bioinformatics.lcd-ustc.org/PPSP | 16549034 | 2005 |
| NetworKIN | Integration of consensus substrate motifs with contextual information on substrates and kinases | Experimentally verified phosphorylation sites | | yes | http://networkin.info | 17981841 | 2007 |
| KinasePhos | Profile hidden Markov model (HMM) | ~ (1880) experimentally verified phosphorylation sites | 0.86% for serine, 0.91% for threonine and 0.84% for tyrosine | yes | http://KinasePhos.mbc.nctu.edu.tw/ | 15980458 | 2005 |
| PREDIKIN | Combine substrate-determining residues (SDR) with structural information | Oriented peptide library | | Yes, only for S/T | http://predikin.biosci.uq.edu.au | 18501020 | 2003 |
| DISPHOS | Integration of position-specific amino acid frequencies and disorder information | Experimentally verified phosphorylation sites | 76% for serine, 81% for threonine and 83% for tyrosine | no | hftp://www.ist.temple.edu/ | 14960716 | 2004 |
| GANNPhos | Genetic algorithm integrated neural network (GANN) | ~ 2500 (7200) experimentally validated phosphorylation sites | 81.1% for serine, 76.7% for threonine and 73.3% for tyrosine | no | n.a. | 17652129 | 2007 |
| PredPhospho | Support vector machines (SVM) | Experimentally verified phosphorylation sites | 83 to 95% at the kinase family level | yes | http://pred.ngri.re.kr/PredPhospho.htm | 15231530 | 2004 |
| NetPhorest | A pipeline selects sequence models of linear motifs using a tree-structured hier-archy | Experimentally verified phosphorylation sites | | yes | http://netphorest.info | 18765831 | 2008 |
| Plewczynski *et al.* | Support vector machines (SVM) integrating local structure information | Swiss-Prot experimentally verified phosphoraylation sites | | yes | n.a. | 15809681 | 2005 |
| Plewczynski *et al.* | Support vector machines (SVM) integrating local structure information | Swiss-Prot experimentally verified phosphoraylation sites | | yes | n.a. | 16094535 | 2005 |
| PHOSIDA | Support vector machines (SVM) integrating context information (from the primary sequence to evolutionary conservation and structural information) | 4731 pS, 664 pT, 107 pY Experimentally verified phosphorylation sites | 89.85 to 91.75% for pS, 74.24 to 81.06% for pT, 66.67 to 76.19% for pY | no | http://www.phosida.com | 18039369 | 2007 |
| Phos3D | Support vector machines (SVM) including P-site 3D-context information | ~750 experimentally verified phosphorylation sites with associated structure | depending on the kinase or kinase group | yes | http://phos3d.mpimp-golm.mpg.de | 19383128 | 2009 |

## 6. CONCLUSIONS

Peptide screen library assays and mass spectrometry experiments made it possible to identify *consensus* sequences – or phospho-motifs - surrounding phospho-acceptor sites and encoding kinase specificity. Such motifs, however, if used to scan a proteome in order to predict novel P-sites, are susceptible to tremendous over-prediction, so that the few true motifs are lost amongst a plethora of false positives. This observation accounts for

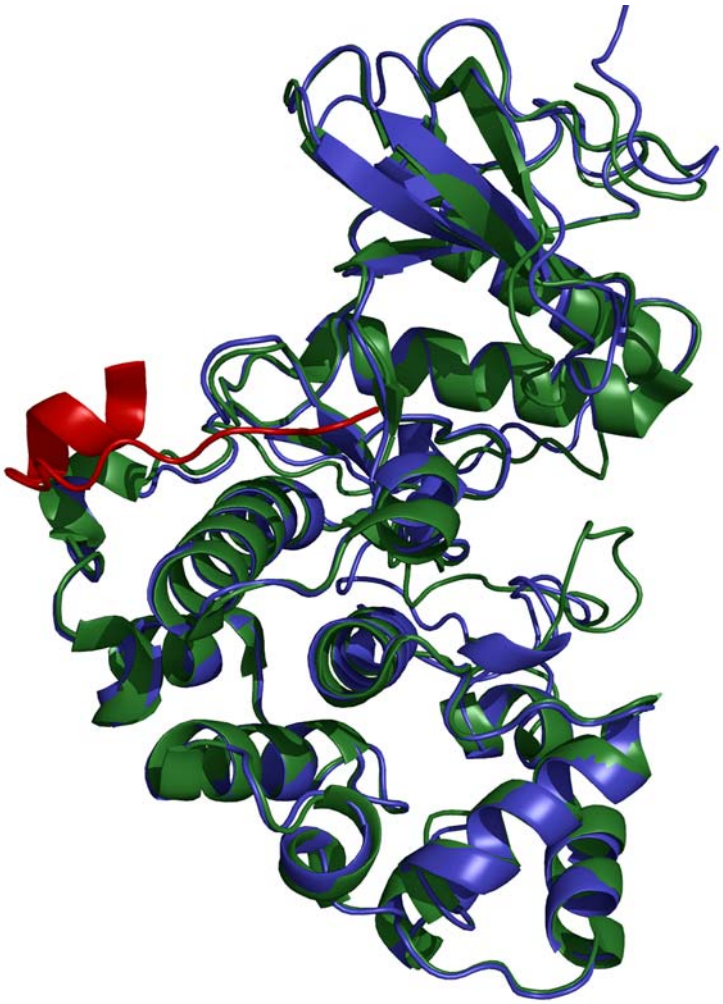

**Figure 2.** Superimposition of the inactive ERK2 (blue) (pdb: 1ERK) and the EKR2 (green) which bind the pepHePTP peptide (red) (pdb: 2ERK). The binding of the docking peptide to ERK2 induces a conformational change in the activation loop which becomes available for the kinase-substrate interaction

more subtle rules governing the specificity of phosphorylation in the cell.

As regard of structure motifs, the honest answer to the question: "Can phosphorylation 3D motifs be defined?" we asked at the beginning of section "Structural features of phosphorylation motifs" is: not yet. However, at the same time, we cannot say that phosphorylation 3D motifs do not exist at all: If we consider the looser 3D phospho motif definitions given in i) and ii) in the same section, we can state that many kinases recognize their substrates through structural patterns made up basically of specific local charge distributions and/or the presence or absence of some specific atom and/or amino acid types. In particular, it has been observed that, in the case of AGC and CAMK kinases, the specificity is mediated not just by favour for positively charged residues but also by the disfavor for a proline residue located close to the P-site. Since mechanisms based on disfavor might be more difficult to ascertain than those based on preference, it is possible that they are more frequent than hitherto observed. Further structural features characterizing these 'motifs' are that most of the time they have a beta-turn, loop or unstructured conformation and that they lie on the substrate surface. These latter properties facilitate the phosphorylation event by making the P-site and its microenvironment clearly visible to the kinase and by making available to the interaction its backbone hydrogen bonding potential. As clearly reported by several authors, phosphorylation is a complex process, which may involve events such as conformational changes (Figure 2), allosteric modifications and recognition of distal docking sites. Moreover, it is well-established that substrate recruitment (e.g. the co-localization of the kinase and the substrate to the same region of the cell or the binding of the kinase and the substrate to the same scaffolding protein, etc.) plays a crucial role in the specificity of many phosphorylation events, and – at least in some cases – it might act as the prevailing – if not the unique – mechanism of specificity. In these cases, we have to abandon the idea of a delimited sequence and/or structural region surrounding the P-site and displaying well-defined biophysical and biochemical

properties that a kinase specifically recognizes and to which the kinase binds in a complementary fashion.

In this elaborate panorama, mass spectrometry experiments are identifying many events of hyper-phosphorylation, especially in protein-disordered regions. It has been argued that such P-sites can be in some cases non-functional, being the harmless result of kinase off-target activities. Apart from this consideration, we want to conclude by underscoring that, in reality, phosphoproteomics data are increasingly corroborating the state-of-the-art view of cell regulation (78), which reasonably considers cell-signaling systems as being networked rather than a series of linear pathways, supports the scenario where there is not a unique or a favorite mechanism accounting for the preference of a kinase for its substrate (s), but where the phosphorylation event is rather the consequence of different processes and (temporal and spatial) contexts that may either cooperate or act independently.

## 7. ACKNOWLEDGEMENTS

We thank Pier Federico Gherardini for critically reading and commenting on the manuscript. This work was supported by the 7th Framework Programme of the European Commission through a grant to the LEISHDRUG project. The authors are grateful to the German Academic Exchange Service BMBF/DAAD and the Italian MIUR for supporting travelling expenses within the bilateral Vigoni-programme.

## 8. REFERENCES

1. Levene P.A. and C.L. Alsberg: The cleavage products of vitellin**.** *J Biol Chem* 2, 127–133 (1906)

2. Cohen P.: The regulation of protein function by multisite phosphorylation--a 25 year update. *Trends Biochem Sci* 25(12), 596-601 (2000)

3. Cohen P.: The origins of protein phosphorylation. *Nat Cell Biol* 4(5), E127-30 (2002)

4. Tsatsanis C. and D.A. Spandidos: The role of oncogenic kinases in human cancer (Review). *Int J Mol Med* 5(6), 583-90 (2000)

5. Blume-Jensen P. and T. Hunter: Oncogenic kinase signalling. *Nature* 411(6835), 355-65 (2001)

6. Ballif B.A., J. Villen, S.A. Beausoleil, D. Schwartz and S.P. Gygi: Phosphoproteomic analysis of the developing mouse brain. *Mol Cell Proteomics* 3(11), 1093-101 (2004)

7. Olsen J.V., B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen and M. Mann: Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 127(3), 635-48 (2006)

8. Zhai B., J. Villen, S.A. Beausoleil, J. Mintseris and S.P. Gygi: Phosphoproteome analysis of Drosophila melanogaster embryos. *J Proteome Res* 7(4), 1675-82 (2008)

9. Wilson-Grady J.T., J. Villen and S.P. Gygi: Phosphoproteome analysis of fission yeast. *J Proteome Res* 7(3), 1088-97 (2008)

10. Kersten B., G.K. Agrawal, H. Iwahashi and R. Rakwal: Plant phosphoproteomics: a long road ahead. *Proteomics* 6(20), 5517-28 (2006)

11. Manning G., D.B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam: The protein kinase complement of the human genome. *Science* 298(5600), 1912-34 (2002)

12. Zhu G., Y. Liu and S. Shaw: Protein kinase specificity. A strategic collaboration between kinase peptide specificity and substrate recruitment. *Cell Cycle* 4(1), 52-6 (2005)

13. Tessmer G.W., J.R. Skuster, L.B. Tabatabai and D.J. Graves: Studies on the specificity of phosphorylase kinase using peptide substrates. *J Biol Chem* 252(16), 5666-71 (1977)

14. Kemp B.E. and R.B. Pearson: Protein kinase recognition sequence motifs. *Trends Biochem Sci* 15(9), 342-6 (1990)

15. Kennelly P.J. and E.G. Krebs: Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. *J Biol Chem* 266(24), 15555-8 (1991)

16. Pinna L.A. and M. Ruzzene: How do protein kinases recognize their substrates? *Biochim Biophys Acta* 1314(3), 191-225 (1996)

17. Zheng J.H., D.R. Knighton, J. Parello, S.S. Taylor and J.M. Sowadski: Crystallization of catalytic subunit of adenosine cyclic monophosphate-dependent protein kinase. *Methods Enzymol* 200, 508-21 (1991)

18. Mayer B.J.: SH3 domains: complexity in moderation. *J Cell Sci* 114(Pt 7), 1253-63 (2001)

19. Schlessinger J. and M.A. Lemmon: SH2 and PTB domains in tyrosine kinase signaling. *Sci STKE* 2003(191), RE12 (2003)

20. Songyang Z., S. Blechner, N. Hoagland, M.F. Hoekstra, H. Piwnica-Worms and L.C. Cantley: Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr Biol* 4(11), 973-82 (1994)

21. Kreegipuu A., N. Blom, S. Brunak and J. Jarv: Statistical analysis of protein kinase specificity determinants. *FEBS Lett* 430(1-2), 45-50 (1998)

22. Yaffe M.B., G.G. Leparc, J. Lai, T. Obata, S. Volinia and L.C. Cantley: A motif-based profile scanning approach for genome-wide prediction of

signaling pathways. *Nat Biotechnol* 19(4), 348-53 (2001)

23. Fujii K., G. Zhu, Y. Liu, J. Hallam, L. Chen, J. Herrero and S. Shaw: Kinase peptide specificity: improved determination and relevance to protein phosphorylation. *Proc Natl Acad Sci U S A* 101(38), 13744-9 (2004)

24. Schwartz D. and S.P. Gygi: An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol* 23(11), 1391-8 (2005)

25. Pan C., F. Gnad, J.V. Olsen and M. Mann: Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors. *Proteomics* 8(21), 4534-46 (2008)

26. Dephoure N., C. Zhou, J. Villen, S.A. Beausoleil, C.E. Bakalarski, S.J. Elledge and S.P. Gygi: A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci U S A* 105(31), 10762-7 (2008)

27. Pellicena P., K.R. Stowell and W.T. Miller: Enhanced phosphorylation of Src family kinase substrates containing SH2 domain binding sites. *J Biol Chem* 273(25), 15325-8 (1998)

28. Tanoue T., M. Adachi, T. Moriguchi and E. Nishida: A conserved docking motif in MAP kinases common to substrates, activators and regulators. *Nat Cell Biol* 2(2), 110-6 (2000)

29. Carlisle Michel J.J., K.L. Dodge, W. Wong, N.C. Mayer, L.K. Langeberg and J.D. Scott: PKA-phosphorylation of PDE4D3 facilitates recruitment of the mAKAP signalling complex. *Biochem J* 381(Pt 3), 587-92 (2004)

30. Faux M.C. and J.D. Scott: More on target with protein phosphorylation: conferring specificity by location. *Trends Biochem Sci* 21(8), 312-5 (1996)

31. Linding R., L.J. Jensen, G.J. Ostheimer, M.A. van Vugt, C. Jorgensen, I.M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J.G. Park, L.D. Samson, J.R. Woodgett, R.B. Russell, P. Bork, M.B. Yaffe and T. Pawson: Systematic discovery of in vivo phosphorylation networks. *Cell* 129(7), 1415-26 (2007)

32. Gnad F., S. Ren, J. Cox, J.V. Olsen, B. Macek, M. Oroshi and M. Mann: PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* 8(11), R250 (2007)

33. Jimenez J.L., B. Hegemann, J.R. Hutchins, J.M. Peters and R. Durbin: A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. *Genome Biol* 8(5), R90 (2007)

34. Marin O., A. Donella-Deana, A.M. Brunati, S. Fischer and L.A. Pinna: Peptides reproducing the phosphoacceptor sites of pp60c-src as substrates for TPK-IIB, a splenic tyrosine kinase devoid of autophosphorylation activity. *J Biol Chem* 266(27), 17798-803 (1991)

35. Meggio F., J.W. Perich, H.E. Meyer, E. Hoffmann-Posorske, D.P. Lennon, R.B. Johns and L.A. Pinna: Synthetic fragments of beta-casein as model substrates for liver and mammary gland casein kinases. *Eur J Biochem* 186(3), 459-64 (1989)

36. Pinna L.A., A. Donella-Deana and F. Meggio: Structural features determining the site specificity of a rat liver cAMP-independent protein kinase. *Biochem Biophys Res Commun* 87(1), 114-20 (1979)

37. Small D., P.Y. Chou and G.D. Fasman: Occurrence of phosphorylated residues in predicted beta-turns: implications for beta-turn participation in control mechanisms. *Biochem Biophys Res Commun* 79(1), 341-6 (1977)

38. Tinker D.A., E.A. Krebs, I.C. Feltham, S.K. Attah-Poku and V.S. Ananthanarayanan: Synthetic beta-turn peptides as substrates for a tyrosine protein kinase. *J Biol Chem* 263(11), 5024-6 (1988)

39. Knighton D.R., J.H. Zheng, L.F. Ten Eyck, V.A. Ashford, N.H. Xuong, S.S. Taylor and J.M. Sowadski: Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* 253(5018), 407-14 (1991)

40. Dunker A.K., C.J. Brown, J.D. Lawson, L.M. Iakoucheva and Z. Obradovic: Intrinsic disorder and protein function. *Biochemistry* 41(21), 6573-82 (2002)

41. Iakoucheva L.M., P. Radivojac, C.J. Brown, T.R. O'Connor, J.G. Sikes, Z. Obradovic and A.K. Dunker: The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 32(3), 1037-49 (2004)

42. Wootton J.C. and S. Federhen: Statistics of local complexity in amino acid sequences and sequence databases *Computers and chemistry* 17(2), 149-163 (1993)

43. Bossemeyer D., R.A. Engh, V. Kinzel, H. Ponstingl and R. Huber: Phosphotransferase and substrate binding mechanism of the cAMP-dependent protein kinase catalytic subunit from porcine heart as deduced from the 2.0 A structure of the complex with Mn2+ adenylyl imidodiphosphate and inhibitor peptide PKI(5-24). *EMBO J* 12(3), 849-59 (1993)

44. Hubbard S.R.: Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J* 16(18), 5572-81 (1997)

45. Lowe E.D., M.E. Noble, V.T. Skamnaki, N.G. Oikonomakos, D.J. Owen and L.N. Johnson: The crystal structure of a phosphorylase kinase peptide substrate

complex: kinase substrate recognition. *EMBO J* 16(22), 6646-58 (1997)

46. Narayana N., S. Cox, S. Shaltiel, S.S. Taylor and N. Xuong: Crystal structure of a polyhistidine-tagged recombinant catalytic subunit of cAMP-dependent protein kinase complexed with the peptide inhibitor PKI(5-24) and adenosine. *Biochemistry* 36(15), 4438-48 (1997)

47. Friedmann M., M.S. Nissen, D.S. Hoover, R. Reeves and N.S. Magnuson: Characterization of the proto-oncogene pim-1: kinase activity and substrate recognition sequence. *Arch Biochem Biophys* 298(2), 594-601 (1992)

48. Meggio F. and L.A. Pinna: One-thousand-and-one substrates of protein kinase CK2? *FASEB J* 17(3), 349-68 (2003)

49. Obata T., M.B. Yaffe, G.G. Leparc, E.T. Piro, H. Maegawa, A. Kashiwagi, R. Kikkawa and L.C. Cantley: Peptide and protein library screening defines optimal substrate motifs for AKT/PKB. *J Biol Chem* 275(46), 36108-15 (2000)

50. Pearson R.B. and B.E. Kemp: Protein kinase phosphorylation site sequences and consensus specificity motifs: tabulations. *Methods Enzymol* 200, 62-81 (1991)

51. Songyang Z., K.L. Carraway, 3rd, M.J. Eck, S.C. Harrison, R.A. Feldman, M. Mohammadi, J. Schlessinger, S.R. Hubbard, D.P. Smith, C. Eng and et al.: Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. *Nature* 373(6514), 536-9 (1995)

52. Songyang Z., K.P. Lu, Y.T. Kwon, L.H. Tsai, O. Filhol, C. Cochet, D.A. Brickey, T.R. Soderling, C. Bartleson, D.J. Graves, A.J. DeMaggio, M.F. Hoekstra, J. Blenis, T. Hunter and L.C. Cantley: A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1. *Mol Cell Biol* 16(11), 6486-93 (1996)

53. Till J.H., P.M. Chan and W.T. Miller: Engineering the substrate specificity of the Abl tyrosine kinase. *J Biol Chem* 274(8), 4995-5003 (1999)

54. Johnson L.N., E.D. Lowe, M.E. Noble and D.J. Owen: The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases. *FEBS Lett* 430(1-2), 1-11 (1998)

55. Johnson L.N., M.E. Noble and D.J. Owen: Active and inactive protein kinases: structural basis for regulation. *Cell* 85(2), 149-58 (1996)

56. Landry C.R., E.D. Levy and S.W. Michnick: Weak functional constraints on phosphoproteomes. *Trends Genet* 25(5), 193-7 (2009)

57. Lienhard G.E.: Non-functional phosphorylations? *Trends Biochem Sci* 33(8), 351-2 (2008)

58. Tan C.S., C. Jorgensen and R. Linding: Roles of "junk phosphorylation" in modulating biomolecular association of phosphorylated proteins? *Cell Cycle* 9(7)

59. Diella F., C.M. Gould, C. Chica, A. Via and T.J. Gibson: Phospho.ELM: a database of phosphorylation sites--update 2008. *Nucleic Acids Res* 36(Database issue), D240-4 (2008)

60. Beausoleil S.A., M. Jedrychowski, D. Schwartz, J.E. Elias, J. Villen, J. Li, M.A. Cohn, L.C. Cantley and S.P. Gygi: Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* 101(33), 12130-5 (2004)

61. Molina H., D.M. Horn, N. Tang, S. Mathivanan and A. Pandey: Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc Natl Acad Sci U S A* 104(7), 2199-204 (2007)

62. Zhu G., K. Fujii, N. Belkina, Y. Liu, M. James, J. Herrero and S. Shaw: Exceptional disfavor for proline at the P + 1 position among AGC and CAMK kinases establishes reciprocal specificity between them and the proline-directed kinases. *J Biol Chem* 280(11), 10743-8 (2005)

63. Fan S.C. and X.G. Zhang: Characterizing the microenvironment surrounding phosphorylated protein sites. *Genomics Proteomics Bioinformatics* 3(4), 213-7 (2005)

64. Kitchen J., R.E. Saunders and J. Warwicker: Charge environments around phosphorylation sites in proteins. *BMC Struct Biol* 8, 19 (2008)

65. Berman H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne: The Protein Data Bank. *Nucleic Acids Res* 28(1), 235-42 (2000)

66. Durek P., C. Schudoma, W. Weckwerth, J. Selbig and D. Walther: Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics* 10, 117 (2009)

67. Zanzoni A., G. Ausiello, A. Via, P.F. Gherardini and M. Helmer-Citterich: Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res* 35(Database issue), D229-31 (2007)

68. Brinkworth R.I., R.A. Breinl and B. Kobe: Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc Natl Acad Sci U S A* 100(1), 74-9 (2003)

69. Glaser F., Y. Rosenberg, A. Kessel, T. Pupko and N. Ben-Tal: The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins* 58(3), 610-7 (2005)

70. Plewczynski D., L. Jaroszewski, A. Godzik, A. Kloczkowski and L. Rychlewski: Molecular modeling of phosphorylation sites in proteins using a database of local structure segments. *J Mol Model* 11(6), 431-8 (2005)

71. Gasteiger E., A. Gattiker, C. Hoogland, I. Ivanyi, R.D. Appel and A. Bairoch: ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* 31(13), 3784-8 (2003)

72. Bystroff C. and D. Baker: Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* 281(3), 565-77 (1998)

73. Adamczak R., A. Porollo and J. Meller: Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* 59(3), 467-75 (2005)

74. Kabsch W. and C. Sander: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12), 2577-637 (1983)

75. Blom N., S. Gammeltoft and S. Brunak: Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* 294(5), 1351-62 (1999)

76. Saunders N.F., R.I. Brinkworth, T. Huber, B.E. Kemp and B. Kobe: Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC Bioinformatics* 9, 245 (2008)

77. Plewczynski D., A. Tkacz, A. Godzik and L. Rychlewski: A support vector machine approach to the identification of phosphorylation sites. *Cell Mol Biol Lett* 10(1), 73-89 (2005)

78. Gibson T.J.: Cell regulation: determined to signal discrete cooperation. *Trends Biochem Sci* 34(10), 471-82 (2009)

**Key Words:** Phosphorylation Site, Protein Structure, Predictor, Bioinformatic, Computational, Review

**Send correspondence to:** Manuela Helmer-Citterich, Center for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Italy, Tel: 39 06 72594324 Fax: 39 06 2023500, E-mail: citterich@uniroma2.it

http://www.bioscience.org/current/vol16.htm