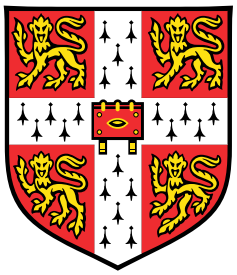
Structural analysis of phosphorylation hotspots and kinase target preferences.



Marta Julia Strumillo

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

HughesHallCollege 26.07.2019

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Marta Julia Strumillo
October 2019

Marta Julia Strumillo

Structural analysis of phosphorylation hotspots and kinase target preferences.

Abstract

Cells are constantly sensing and adapting to changes in conditions. Protein post-translational regulation is one of the fastest mechanism used by cells to relay signals from sensors to effectors during such adaptations. Mass spectrometry allows for the study of posttranslational modif cations on a very large scale and has been extensively applied to study protein phosphorylation. On the order of 75% of human proteins have been estimated to be phosphorylated and approximately 160,000 human phosphosites are listed in public repositories. This wealth of knowledge, remains mostly uncharacterized with around 5% of human phosphosites having an annotated regulatory role or known regulatory kinase. Devising ways to study the functional importance of phosphosites is therefore a crucial research question. The recognition of target sites by a kinase is thought to be determined by a short contiguous sequence motif around the target phosphosite. It has been reported that kinases can, in some cases, recognize a 3D epitope instead of a linear peptide sequence. However, the extent by which 3D epitopes are important for kinase recognition is unknown.

To study the usage of 3D kinase recognition motifs, I f rstly examined if known in vitro and in vivo human kinase targets can be explained by 3D epitopes. For this I devised a computational pipeline mapping known kinase target phosphosites to structural models. Using these I identif ed potential cases where the important specif city determinant residues are not observed in contiguous sequences in the targets but may exist as a 3D epitope, and performed docking simulations to examine the possible kinase interactions. The 3D epitope examples were found to be rather exceptions than a rule, and the analysis conf rms the general rule of linear motif recognition by kinases.

To better predict phosphorylation of high functional relevance I analysed phosphosites that are highly conserved across species within protein domains families. These regions of conserved phosphorylation, def ned as phosphorylation hotspots, were determined using phosphosite data for a total of 40 eukaryotic species. A total of 241 domain regions were identif ed as hotspots within 162 domain families that were then mapped to proteins structures. These regions were shown to predict known regulatory sites and overlap with important structural features (i.e. protein interfaces and residues near or at catalytic sites). To

further study the regulatory regions of protein domains I searched for regions of conserved ubiquitination and/or a high degree of recurrent mutations found in cancer. Of 68 domains that had enough data for analysis of all 3 types of hotspots I present the analysis of interesting cases and domains containing overlapping PTM and/or mutational hotspots.

I would like to dedicate this thesis to my loving parents.

Acknowledgements

There are several people who I would like to thank for help, support and advice throughout my long EBI - Cambridge adventure.

First and foremost, I would like to thank and acknowledge my supervisor Pedro Beltrao for welcoming me into his group, guiding me through the research, and being extremely patient about every little thing that describes me and my scientific work. The inspiration, creativity and scientific dazzle of Beltrao Lab was the main driving force of my research and a highlight of my everyday. David Ochoa, Brandon Invergo, and Marco Galardini: your insights, perception, personalities, and friendship are the input that will still influence me years after. I would like to offer my profound gratitude to all the members of Beltrao Lab, especially David Bradley and Haruna Imamura. Additional, special, thanks to Omar Wagih, who I missed everyday after he graduated, the way you miss your favourite cousin at a boring family party. My best friend, Nadezda Volkova, you are a star, I would not have made it without you. I also thank Greg Slodkowicz, for his friendship and all the scientific (and not) discussions, that we had since day one. Not seeing all of you everyday, and not working with you is the loss of which enormous size I am still estimating everyday.

I would like to thank my family for supporting me continuously throughout my studies and believing in me more than I did myself. I owe a debt of gratitude to my love Tony, for his patience, support, love and understanding.

With all the warmest memories I fully dedicate Chapter 6 to Capt. Benjamin 'Irish' Coffey and the 494th Fighter Squadron.

Table of contents

List of figures

Chapter 1

Introduction

1.1 Proteome

The complete sequencing of the human genome, defined as a the full content of all chromosome related information, was a hallmark of large-scale biology. Although it took immense effort and series of related and sometimes unintended innovations, the definition of entire genome is relatively well established. Human genome project revealed 20,300 protein coding genes, in contrast to the estimated 100,000 (Pruitt et al. [229]). This unexpected finding led to the recognition that the protein variation might be mostly responsible for the complexity of the biological organisms rather than a high number of distinct genes (Schlüter et al. [247]). A term 'proteoform' has been introduced to better refer to a product of a single gene (Smith et al. [266]) that is a protein but also encapsulates changes due to genetic variations, alternatively spliced RNA transcripts and post-translational modifications. The proteome, unlike the genome, has a non-linear character with a highly dynamic range and differences over time and space of proteins. In the case of proteomics, the definition of a complete, reference proteome presents a new, more complicated challenge. Characterization of all the possible isoforms and modification states of all expressed proteins may be impossible to experimentally discover, because of the astronomical number of possible combinations (Cox and Mann [51]). The limited view that comprahensive proteome can be defined as all the proteins identifiable by a state of art mass spectrometric methodology (Beck et al. [12]) has its practical aspect. Alternatively, the Chromosome-Centric Human Proteome Project identifies one or more protein representatives from all the protein coding genes in an organism (Paik et al. [215]). The most pragmatic and easiest, yet still very laborious definition to achieve, is the identification and quantification of at least one protein form from every genomic locus that is expressed in a given biological system. Such an

achievement would provide a very rich source even for the low abundant proteins and all the isoforms. So far, only partial information about the proteome and phosphoproteome are available in standardised databases. Efforts from OpenProt (Brunet et al. [31]) and UniProt (UniProt Consortium [290]) address the needs of the field and provide the most popular and extensive databases. The first description of a complete model proteome (de Godoy et al. [59]) along with the identification of proteins in human cell lines (Beck et al. [12], Nagaraj et al. [202]) took several months, and required immense effort to obtain. Improvements to mass spectrometers and sample preparation have enhanced the sensitivity and speed of such experiments, and nowadays, the same research requires only a fraction of the original time. Easier standard sample preparation and shorter time required for the MS machines, combined with straightforward bioinformatic analysis, has greatly improved the overall state of knowledge. However, only some specialised laboratories have well established robust pipelines.

1.2 PTMs and Phosphorylation

Posttranslational modifications and their structural and functional annotations present an analysis problem on their own, with arduous data gathering processes and individual function assignment. Over 300 different types of PTMs are known, ranging from single atom modifications (oxide) to small protein modifiers (ubiquitin) (Walsh et al. [296]). Protein phosphorylation is likely to be the most extensive and most well characterized PTM. On the order of 75% of the human proteome has been suggested to be phosphorylated and over 100,000 phosphosites have been discovered for these proteins (Sharma et al. [256]). Phosphorylation is catalysed by kinases and constitutes the transfer of the terminal phosphate group from ATP to the hydroxyl group of amino-acids. The transfer to serine (S), threonine (T), tyrosine (Y), histidine (H) and aspartic acid (D) residues is possible, however due to their prevalence in eukaryotic species and technical limitations, most studies have focused on S,T,Y phosphorylation (Mann et al. [179], Thomason and Kay [280]). The reverse mechanism (dephosphorylation) is catalyzed by phosphatases (Figure 1.1), the cooperation of both processes control actions such as molecular association, protein degradation, enzymatic activation, intracellular localization, etc (Hunter [116]).

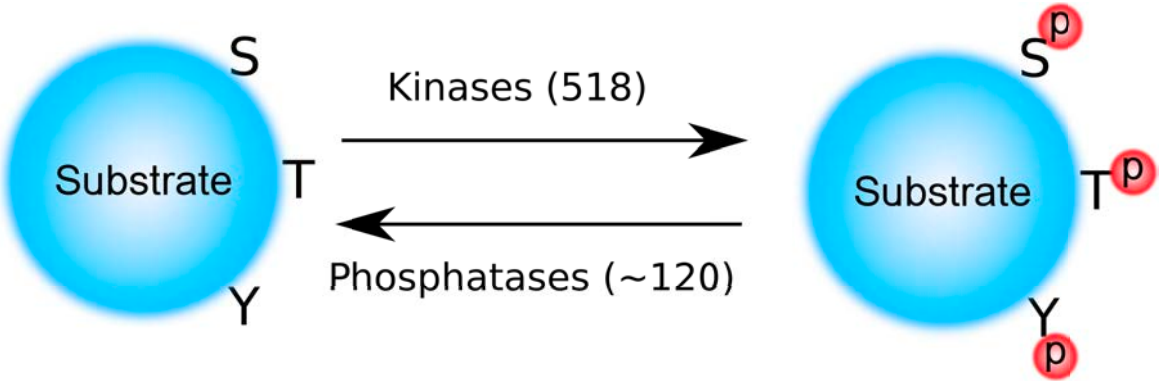


Fig. 1.1 There are 518 protein kinases in human genome, that are capable of phosphorylating substrate on Serine/Threonine/Tyrosine residues. Around 120 human phosphatases are responsible for phosphorylation removal.

1.3 Protein kinases and kinase specificity

1.3.1 Protein kinases classification

Eukaryotic protein kinases are structurally distinct from many other kinases found in prokaryotes. In prokaryotes, phosphorylation can be maintained by different molecular systems. In bacteria the two-component system is responsible for phosphate-based signal transduction, in which the sensor kinase first autophosphorylates on a Histidine residue and then transfers a phosphate to an Aspartate residue of a response regulator (Robinson et al. [239]). Prokaryotes also control the phosphorylation networks by proteins that are both kinases and phosphatases (e.g. isocitrate dehydrogenase kinase/phosphatase enzyme (Laporte et al. [160])) or function in distinct phosphotransferase systems (Kotrba et al. [152]). Eukaryotic-Like Kinases (ELKs) are evolutionary related to Eukaryotic Protein Kinases and phosphorylate small metabolites (Oruganty et al. [214]). ELKs also widely exist in eukaryotes, sharing the common kinase fold with eukaryotic protein kinases, and small sequence similarity (Figure 2). Eukaryotic protein kinases (ePKs) are the main type of kinases phosphorylating substrates in eukaryotic organisms. Integrative sequential and structural analysis of the ePKs and ELKs suggests that the ePKs diverged from the ELKs early during evolution (Oruganty et al. 2016). Eukaryotic kinases phosphorylating Histidines or other residues, do not share a common fold with those phosphorylating S/T/Y (ePKs).

The standard eukaryotic protein kinase classification scheme considers their evolutionary history, function, and structure (Manning et al. [181]). This scheme classifies kinases into 9 groups, 134 families and 196 subfamilies and covers human, yeast, worm and the fly kinome. The classification has been based on the previous research that included only 5 kinase groups

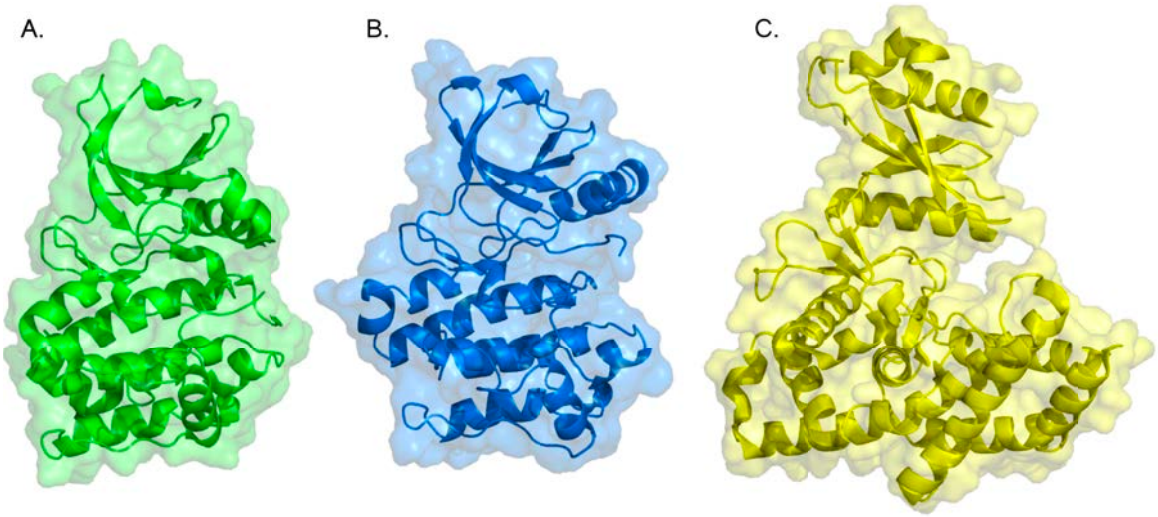


Fig. 1.2 Structural comparison of three different, human kinases. A. Serine/Threonine specific Aurora kinase (pdb: 1mq4), B. Tyrosine specific Kinase Hck (pdb: 2hck), C. Small molecule kinase - Choline Kinase alpha 2 (pdb: 2cko). Despite different specificities, all three kinases share the same kinase fold.

(Hanks and Hunter [96]). Most eukaryotic kinases from each group share a common catalytic domain, however sequence analysis shows major variations between the groups, with distinct and ancient functions. To map all groups of kinases across large evolutionary distances, each group can contain multiple families, which can contain multiple subfamilies. The process of classifying eukaryotic protein kinases includes clustering by: sequence similarity within the kinase domain, additional information from other domains within the kinase, evolutionary conservation, and known function. Besides the TK group which phosphorylates Tyrosine, all other groups phosphorylate S/T residues. This very practical, hybrid classification is still under development (especially the subfamilies) while new kinomes are being sequenced. In the list below all 9 groups of protein kinases are briefly described along with the Atypical and Other groups.

- **AGC** group is named after the Protein Kinase A, G and C families (PKA, PKG and PKC). Kinases within this group are mostly core intracellular signaling enzymes, modulated by cyclic nucleotides, phospholipids and calcium.
- **CMGC** group includes families CDK, MAPK, GSK3 and CLK, hence the name. The diversity of functions within the group includes cycle control, MAPK signalling, splicing and other.

- **CAMK** group includes Calmodulin/Calcium regulated kinases, split into CAMK1 and CAMK2 families. There are several families of non-calcium regulated kinases within this group.
- **CK1** is a small, ancient family with the name originating from Cell kinase 1. Members of this group are conserved from yeast to human.
- **STE** group contains 3 families that include homologs of the yeast STE7, STE11 and STE20. These kinases sequentially activate each other in a MAPK cascade in order to activate MAPK family.
- **TK** group stands for Tyrosine Kinase, which almost exclusively phosphorylates tyrosine residues. The evolutionary analysis indicates TK as the youngest group with the biggest number of distinct families.
- **TKL** are Tyrosine Kinase-Like kinases, that are mostly similar to TKs, but they are phosphorylating S/T substrates instead. More than half of plant kinomes include TKL kinases, such as receptor kinases and possible tyrosine-specific kinases in other lineages.
- **RGC** is a small group containing Receptor Guanylate Cyclases with an active guanylate cyclase domain and a catalytically inactive kinase domain.
- **PKL** contains several diverse families, sharing a Protein Kinase-Like fold and catalytic mechanism with other ePKs, like, lipid, sugar and other small-molecule kinases.
- **aPK** — the Atypical protein kinases in this group are shown experimentally to have protein kinase activity, but do not have structural similarity to ePKs.
- **Other** group includes several families, of which kinases clearly contain kinase domain, but do not fit into the remaining groups. The group includes e.g. Aurora Family, CAM kinase kinase (which activates CAMK1), PLK family and several more.

1.3.2 Phosphatases

While the main focus of the thesis is on kinase recognition and phosphosite function it is relevant to note the importance of phosphatases in phosphorylation signaling. Protein Phosphatases (PPs) are capable of dephosphorylating amino acids that has been previously phosphorylated by a kinase. Phosphatase uses water molecules to cleave a phosphoric acid

monoester into a phosphate ion and an alcohol. Hence they are classified as a subcategory of hydrolases because of their capability of hydrolysing their substrate. Protein phosphatases can be grouped into three main classes based on sequence, structure and catalytic function. Subdivision of phosphatases based upon their substrate specificity can be distinguished as:

- Tyrosine-specific phosphatases
- Serine-/threonine- specific phosphatases
- Dual specificity phosphatases (serine/threonine or tyrosine)
- Histidine phosphatase

The human genome encodes around 200 PPs, with ~40 Ser/Thr specific, ~100 Tyrosine specific and ~50 of dual specificity (Moorhead et al. [198]). Specificity of phosphatases appears to manifest mostly through the association of phosphatase catalytic domains with particular regulatory subunits (Ubersax and Ferrell [289]). The specificity of the phosphatases and the so called "phosphatase code" are still important research tasks that are not yet fully understood. Despite the impression of recognizing multiple substrates *in vitro*, in *in vivo* experiments protein phosphatases remained extremely specific (Sacco et al. [243]).

1.3.3 Protein kinase domain

The catalytic unit of a S/T kinase is a structurally conserved protein domain, classified in Pfam as PF00069. Similarly Tyrosine specific kinases are depicted in Pfam as PF007714, however the structural differences between these catalytic domains are mostly related to the substrate binding pocket. The domain is constructed from ca. 300 amino acids, with a catalytic pocket of 10 residues (Hanks and Hunter [96], Manning et al. [181], Kannan and Neuwald [134]). Other structural domains that a kinase can contain are usually identified as regulatory or targeting modules (Scott and Pawson [250]) and are often the origin of the kinase name (e.g. Polo kinase contains a Polo domain along with the ePK domain). There are 9546 architectures of kinase domain containing genes reported in Pfam. Such variety of architectures means that the kinase domain may have a lot of other domains (SH2, SH3, WD40, PH, Death etc) in close proximity, connected by linkers, that all together create a functioning protein. In Figure 1.3 a structural representation of a kinase domain is shown with a highlighted N-lobe (consisting mostly of beta sheets) and a larger, helical C-lobe. ATP binds in the cleft between the lobes, where the adenine group of ATP intercalates with hydrophobic residues of the pocket (Figure 1.3) (Hu et al. [108], Nolen et al. [208]).

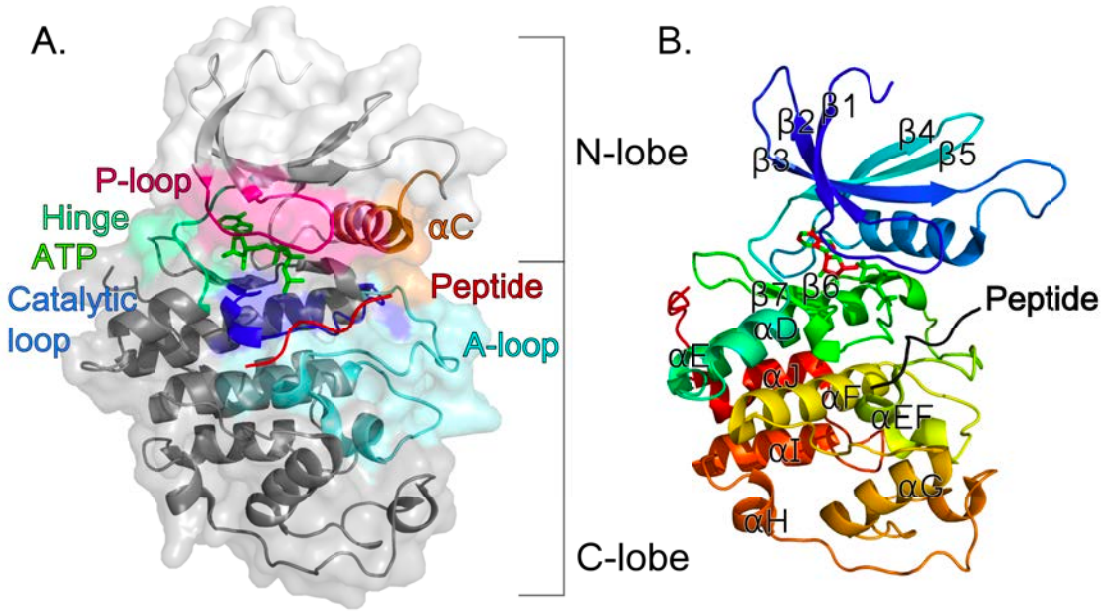


Fig. 1.3 Structural representation of the kinase. A. Catalytic pocket is located in the cavity between catalytic loop, P-loop and A-loop. Peptide binding inside the pocket can easily access the ATP for phosphoryl transfer. The structure used for presentation is that of CyclinB (pdb: 1qmz). B. The rainbow coloring allows to see the numbering of the helices in the C-lobe.

Shown in Figure 1.3 the N-lobe is constructed of five β sheets (β1-β5) with a Gly-rich loop (called P-loop) between β1 and β2 sheets. This flexible P-loop contains an important hydrophobic residue that contributes to the coordination and positioning of the phosphates of ATP, folding over the nucleotide (Cowan-Jacob [48]). A single helix within the N-lobe is called the C-helix (αC). The C-helix plays an important role facilitating the catalysis — rotating the N-terminus into C-helix-out position (suboptimal for catalysis) results in inactivation of the kinase, and provides contacts for other motifs (Cowan-Jacob [48], Kannan et al. [135], Fabbro et al. [74]). The helical C-lobe includes helices D, E, EF, F, G, H, I and J, and from two to four β sheets (β6-β9). These β sheets transform into loops quite easily — some of the crystalised structures (like the CyclinB presented in Figure 1.2) contain only some of the β sheets, eg. β6 and β7. Strand β9 is usually contained within the activation loop, and the catalytic loop includes β6 and β7. The DFG-motif (Aspartate-Glycine-Phenylalanine) is located between β8 and β9, and its Aspartate recognizes one of the Mg^{2+} ions. Also in the DFG-motif the Phe makes hydrophobic contacts with the C-helix facilitating the Lys-Glu salt bridge.

The DFG-motif extends into the A-loop, which ends at the beginning of the F-helix. The A-loop is a very flexible region, which regulates the on and off state of the kinase by changing its conformation and access to the pocket. The activation loop (A-loop) occurs in an open (activated) or various closed conformations (inactivated) (Nolen et al. [208]). The platform for the substrate binding is made by the activated A-loop together with the helices of the C-lobe (Ubersax and Ferrell [289]). The GHI domain includes three helices (αG, αH and αI) on the bottom of the C-lobe and is unique to ePKs. Many substrate and regulatory proteins bind to the GHI domain. The activation of a protein kinase results in the re-orientation of the A-loop and the C-helix. The Glu from the C-helix comes into proximity of the active site Lys (from the AXK-motif) and the A-loop. In many kinases, the Phe from the DFG motif also changes position upon activation - from the DFG-out (inactive) to DFG-in (active conformation). The A-loop can be stabilized in the active conformation by phosphorylation or other interactions with the additional regulatory proteins. The catalytic loop (β6/β7) contains the Y/HRD-motif. Residues in the Y/HRD motif are conserved throughout all ePKs and ELKs. Tyr/His from this motif serves as a central scaffold for binding of the Asp and making a contact with the Phe from the DFG motif. Asp of the Y/HRD is responsible for the correct orientation of the phosphosite hydroxyl acceptor group in the peptide substrate. The catalytic loop does not change the conformation during the activation of the kinase.

The residue controlling the access into the deep pocket is called a gatekeeper and is located within the Hinge (Figure 1.3). Mutation of the gatekeeper amino acid can cause

resistance to the inhibitors (Cowan-Jacob [48], Taylor and Kornev [279], Moebitz and Fabbro [196]). The resistance due to the gatekeeper mutation is usually causing a steric clash with the inhibitor, or significant increase of the affinity towards ATP. The optimal structure of the active kinase includes also the formation of the Regulatory spine (R-spine) and the Catalytic spine (C-spine). The R-spine is built by four residues- one from the $\beta 4$ sheet, one from the C-helix, the Phe from the DFG motif in the N-lobe and the Y/H from the catalytic loop. Hydrophobic interaction of these four residues supports the scaffold between the N- and C-lobes that supports the optimal kinase activity. The assembly of the R-spine is a hallmark of an active kinase. The C-spine contains the Val from the $\beta 2$ sheet and the Ala from the $\beta 3$ (from the AXK-motif) which are connected with adenine ring of ATP.

Although the sequences of many kinases vary, three sequence motifs within the kinase are necessary for the catalysis. Firstly, placed in the $\beta 3$ sheet, is the AXK-motif. The AXK-motif contains the active site Lys that forms a salt bridge with the Glu from the C-helix. This motif interacts with the phosphates of ATP to anchor and orient the ATP inside the pocket. Secondly, the Y/HRD-motif inside the catalytic loop ($\beta 6/\beta 7$), in which the Asp functions as the acceptor for the proton transfer. Lastly, the DFG-motif within the A-loop contains the Asp that binds the Mg^{2+} ions. The Mg^{2+} ions coordinate the phosphates of ATP in the cleft to position the ATP for the phosphate transfer.

1.3.4 Kinase-substrate recognition

Eukaryotic protein kinases are generally subdivided into S/T kinases, Y kinases, and dual-specificity kinases, based on their favoured substrates. Kinase preferences are mostly determined by conserved features of the kinase catalytic pocket, which is unique to each class of kinases. However, kinases target specific substrates through several types of physical interactions, not only within the catalytic pocket (Manning et al. [181], Ubersax and Ferrell [289]).

Efficient phosphorylation of a substrate requires a binding of a discrete peptide sequence inside the catalytic pocket of the kinase. In crystal structures the substrate generally binds inside the cleft in an extended conformation, however it might be due to the crystallization techniques. Crystallized peptide makes β-sheet-like interactions with a part of the kinase activation loop. As a consequence, residues flanking the phosphorylation site are recognised within the cleft, however this interaction alone is insufficient to mediate selection of protein substrates (Bose et al. [29], Goldsmith et al. [90]). Substrate specificity is usually enhanced by docking interactions (examples presented in Figure 1.4). Distal to the phosphorylation, parts

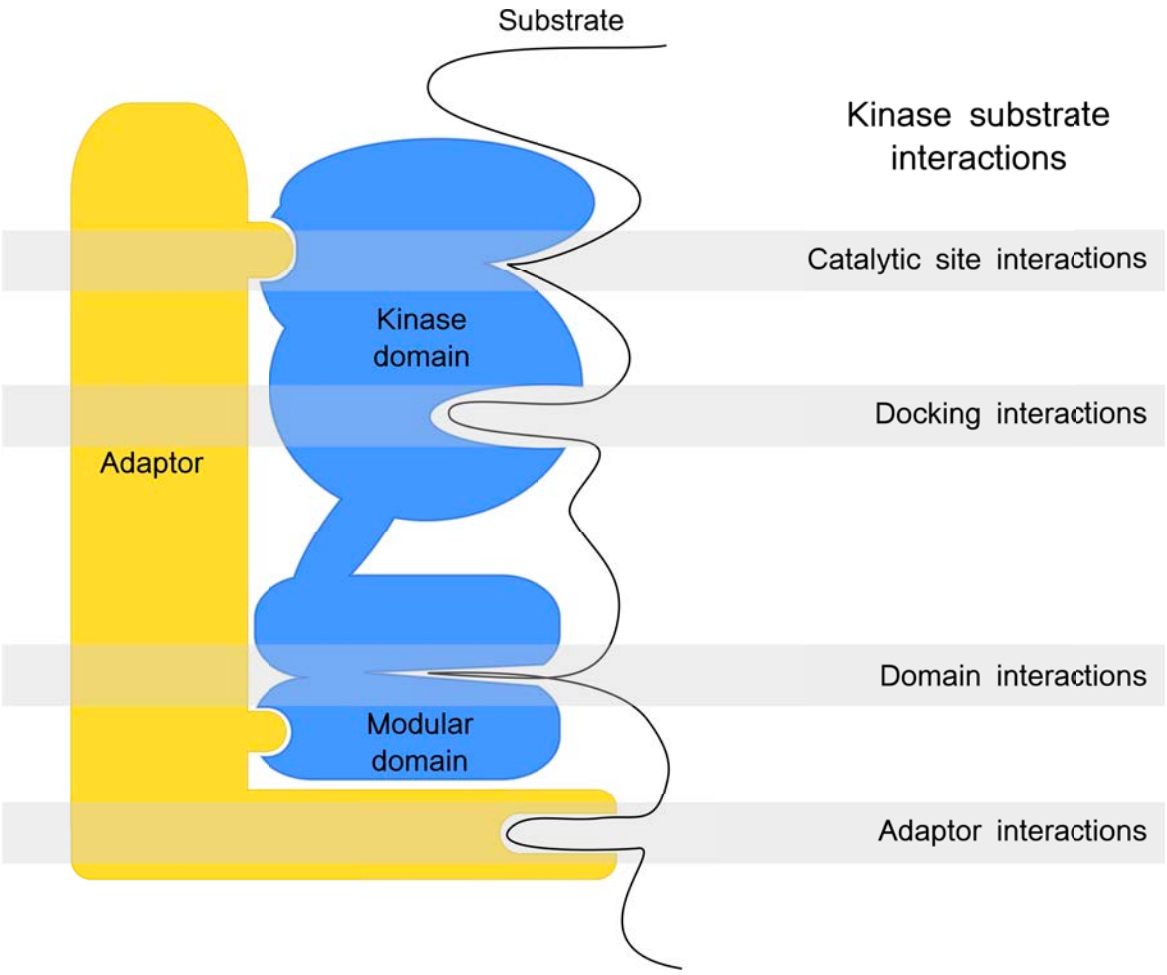


Fig. 1.4 Substrates interact with kinases through a combination of catalytic domain interactions, and proximal or distal interactions to the active sites. Interactions involve short linear sequence motifs recognition by protein modules, and indirect interactions mediated by adaptor/scaffold proteins.

of the substrate bind to other pockets in the kinase surface and/or to its surrounding proteins in order to improve substrate affinity and specificity. Similarly to phosphosite interactions, docking interactions can involve short linear sequence motifs that can be recognised by scaffold proteins interacting with the kinase. Adaptor and scaffold proteins can promote phosphorylation through induced proximity, controlling kinase subcellular localisation, changing the conformation of substrates and serving as hubs for substrate recruitment.

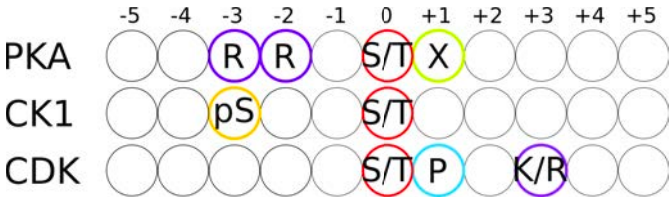


Fig. 1.5 Three example sequential motifs for kinases. PKA, Protein kinase A requires two arginies on positions -3 and -2, and a hydrophobic residue (X) on the position +1. CK1, Casein kinase-1 requires already a phosphorylated Serine at position -3. CDK, Cyclin-dependent kinase preferes peptides with Proline on +1 position and Arginine or Lysine on the +3 position.

1.3.5 Kinase recognition Linear Motifs

The primary sequence of the substrate plays an essential role in kinase recognition. A sequence motif usually consisting of up to five residues surrounding the phosphosite are critical for efficient phosphorylation. Some examples of such kinase target motifs are shown in Figure 1.5. Related kinases may have identical phosphorylation site motifs, emphasizing the role of other ways of determining specificity such as cellular localization. The relevance of individual determinants can greatly vary, both on the kinase and on the motif side. Some of the residues within the motif can be more important than the others — an important determinant at the canonical position can be replaced by the same residue in a nearby position. Sequential motifs are usually described as “preferable”, while negative determinants can also greatly influence correct phosphorylation. Negative determinants prohibit a kinase from phosphorylating an otherwise preferential peptide. Often in case of a deleterious mutation, a residue from the motif mutates into a negative determinant and hence compromises the ability of the catalytic binding.

Preferences of S/T kinases can be roughly divided into three categories. The first one contains basophilic kinases, using as determinants basic residues (Arg and Lys) and often hydrophobic residues (Ala, Gly and more). The second category contains proline-directed kinases, preferring Proline and basic residues. The last category contains acidophilic/phosphate-directed kinases, with carboxylic and previously phosphorylated residues (Pinna and Ruzzene [224]). Although it has been widely recognized that the specific motifs can be a useful predictor of kinase preference (Kemp and Pearson [140], Miller et al. [194]), they do not have enough resolving power alone to assign the substrates to a single kinase with high confidence. Motifs often share many common phosphosites even between kinases from different families — Needham and colleagues has shown that fewer than 18% of sites match a single motif (Needham et al. [205]). In addition, kinases with well-defined motifs can also phos-

phorylate atypical sites in certain circumstances. New structural mechanisms of achieving kinase-substrate fidelity are still being discovered. In the study of Duarte and colleagues a noncontiguous motif for the PKC kinase has been presented - PKC prefers a linear motif of a basic amino acid at either P-3 or P-2 (on position -3/-2 from the P-site) (Duarte et al. [69]). The study showed that a basic amino acid from a distal part of the protein is able to form a structural noncontiguous motif recognized by the kinase. Although this has been shown for one example, it is not yet known how often kinases can recognize targets via non-contiguous motifs.

1.3.6 Methods of identifying kinase substrates

Measuring kinase activity and identifying its substrates is essential in discovering phosphorylation pathways. Regardless of the approach, defining substrates can be described as showing that a particular target kinase phosphorylates the substrate more efficiently than any other generic protein. Important proof of *in-vivo* relevance that should be supplied along with the interaction information, is that of the presence of the kinase and substrate in the same cellular compartment. Finally, it should also be demonstrated that the phosphorylated residue is the same *in-vivo* as in *in-vitro*, and also what is the biological process regulated by this phosphorylation. Amongst many available methods of establishing kinase-substrate relationships, I am describing the most popular ones with their advantages and downsides. These methods do not include a range of Mass Spectrometry based approaches that are described in the Paragraph 1.4.

Genetic screening

Historically genetic screening has been often applied to identify protein substrates of kinases. This approach has been often applied in model organisms such as yeast, worms or flies, however, until recently, it has been difficult to use in mammalian species due to the bottleneck of high throughput mutagenesis. The outcomes however can still be useful for human studies by using homology and predictions of kinase substrates. Genetic screening methods include the establishment of the phenotype for a mutated target kinase, to then screen the genes that can either suppress the phenotype in mutants or mimic the kinase mutant phenotype in wild-type. The high throughput genetic manipulation (recently including also siRNA and CRISPR technology) can be performed on a genome-wide scale. Products of selected genes are later tested as substrates of the kinase by various biochemical approaches (Leberer et al. [163], Paradis and Ruvkun [217], Clark et al. [43], Sha et al. [253]). This method identifies

relevant proteins directly, however the actual relationship between identified genes and the kinase might be very diverse.

In vitro kinase assay

The most popular biochemical method to determine kinase activity towards the substrates is the *in vitro* kinase assay. During the experiment the selected, purified kinase is incubated with possible substrates and ATP. After the incubation period the phosphorylation of the substrate can be assessed with a variety of methods, such as calorimetric, radioactive, chemiluminescence or fluorometric detection (Johnson and Hunter [131]). The main limitation of this method is that the phosphorylation *in vitro* may be different to the phosphorylation that takes place in the living cell. Often kinases may require additional scaffolding and binding proteins, and not all of the substrates may be present in the same subcellular compartment as the kinase. Due to these facts, elimination of false positives has to be performed in *in vivo* studies (Delom and Chevet [61]). This combination of methods provides trustworthy results, but is very laborious and low throughput. There are number of approaches that attempt to screen potential kinase substrates in a high throughput manner like microarrays and phage display.

Protein and peptide microarrays

A peptide microarray (also called peptide chip or peptide epitope microarray) is a surface-based collection of peptides, usually displayed on a glass or plastic chip. They are used to study binding properties and kinetics of protein-protein interactions. Microarrays can be used to profile an enzyme (a selected kinase) to find key residues for protein binding. When a chip containing the human proteome has been created the high throughput analysis of kinase target has become widely available (Jeong et al. [128]). The array technique usually requires small amounts of the purified kinase and other reagents, while providing a sensitive and rapid assay. Since a typical peptide microarray consists of hundreds of peptides derived from a specific organism, kinase primary sequence preferences can be easily established. The phosphorylated peptides, obtained after incubating the chip with the kinase can be analysed with autoradiography, fluorescence, or immunoblotting (Buss et al. [32], Lesaicherre et al. [167]). Although the microarrays do not highlight actual kinase-substrate relationship, they provide valuable information of kinase sequence preferences. General concerns of microarray efficiency has been mentioned in cases where a third adaptor protein is necessary or a peptide happens to be in the incorrect conformation (Huang et al. [113]).

Phage display

Phage display is a popular approach for studying protein-peptide and protein-ligand interaction (Fukunaga and Hunter [87]). In the case of identifying kinase targets, a custom cDNA library is cloned into phage expression vectors. Proteins coded by individual cDNA clones are massively expressed after adding phage plaque on lawns of Escherichia coli. The expressed proteins are subsequently immobilized on the solid phase and then phosphorylated by the kinase of interest in the presence of ATP. The phosphorylated substrates can be enriched by panning over phospho-selective antibodies and identified by sequencing the phage plaques (Dente et al. [62]). To improve the sensitivity, multiple cycles of selection can enrich the positive phages with phosphorylation signals. After the selection, distinct substrate sequences can be defined by phage cloning. This approach identifies the candidate substrate by isolating the clone from the phosphorylation-positive plaque, however potential problems may arise from the incorrect folding of cDNA-encoded proteins in bacteria expression system (Pillay [223]).

Protein interaction based screening

Discovering protein-protein interaction is a common path to screen potential kinase substrates. The idea of coidentifying kinase-substrate pairs interacting is very tempting (Staudinger et al. [271], Tien et al. [281], Amano et al. [6]), however phosphorylation is commonly considered a transient protein-protein interaction. A single kinase is able to phosphorylate multiple substrates in a very short time, thus such interactions are difficult to be trapped or identified. Kinase-substrate complexes are hardly captured in affinity purification experiments and the interaction does not trigger the reporter gene transcription in the yeast-two-hybrid system. Because of these difficulties, few studies have identified kinase substrates by two–hybrid (Yang et al. [311], Vadlamudi et al. [292]) or affinity purification (Daub et al. [57], Belozerov et al. [14]). The biggest problem of such approaches is the number of false positives, because a potential large number of other proteins, not only the substrates, interacts with the target kinases physically (like adaptor or scaffold proteins). Confirmation of candidate substrates by a secondary method is therefore necessary.

Bioinformatics predictions

The central hypothesis of kinase substrate prediction is that the substrate consensus motif plays a determining role in kinase recognition. Through the use of peptide library screening and other advanced technologies, many kinases have been examined for their sequence

preferences (Songyang et al. [269]). Once the quantitative kinase preference is established, it can be used to examine any known phosphorylated protein sequence to predict which kinase may account for the phosphorylation event. Although the prediction of the specificity or the substrate itself is not a complete proof of phosphorylation, such predictions can supply quick screening to narrow down the options for further biological testing. One of the first developed bioinformatic tools that offers predictions of kinase motifs is Scansite (Obenauer et al. [210]), where each sequence motif is represented as a position-specific scoring matrix (PSSM). Scansite and other bioinformatic tools take great advantage of the massive biological information that is still being developed to generate better predictions. Other matrix based methods include PHOSITE and PhoScan, that assign weights to all 20 amino acids, rather than representing only the most popular ones like Scansite (Koenig and Grabe [146], Li et al. [168]). NetPhorest is a popular tool using artificial neural networks (Miller et al. [194]), to model inter-positional dependencies for some of the kinases. KinomeXplorer (Horn et al. [103]) is a platform integrating NetPhorest and NetworKIN (Linding et al. [169]) which are computational approaches that combine consensus sequence motifs and protein–protein interaction networks to supply better predictions. Like NetPhorest, the iGPS (Song et al. [267]) approach also predicts kinase specific interactions using neural networks.

Predictors capable of including the structural information of the kinase use machine learning methods (neural networks and support vector machines) (Trost and Kusalik [288]). Incorporating many different features of the kinase and substrate into one method has been achieved in NetPhos (Blom et al. [26]), Phos3D (Durek et al. [70]), and PhosK3D (Su and Lee [277]). In Phos3D and PhosK3D kinase specificity is analysed as a radial pattern of amino acids biases in the vicinity of the phosphosite rather than a sequence motif. Although the structural information does not significantly improve the overall score of the predictions, it is thought that larger numbers of kinase–substrate structural models will greatly improve the existing methods (Durek et al. [70], Plewczynski et al. [226]).

Predictors using the sequence of the kinase of interest to predict substrate specificity rely upon the homology of the query kinase to a set of kinases for which the specificity has been already experimentally determined. The most popular methods in this category are Predikin (Saunders et al. [244]) and KINspect (Creixell et al. [53]). These sequence based approaches are very valuable in the analysis of the kinases for which no structure is known and those with unknown substrates.

1.4 Mass Spectrometry analysis of the proteome

Advances in mass spectrometry have significantly improved PTMs analysis within the last years. High speed, high resolution and direct analysis have contributed to the MS dominance in PTM identification and analysis technology. Compared to other aforementioned techniques, like microarrays or phage display system, MS not only identifies phosphoproteins, but also directly highlights the phosphosites. Nowadays, thanks to mass-spectrometric advances, it is possible to perform large-scale analysis of whole proteomes of multiple organisms at once. Mass-spectrometry has fundamentally improved the methods of analysing single proteins as well as systematic measurements. Available techniques are able to quantify and identify almost any expressed protein, localise and identify post translationally modified amino acids, and provide insights into topology and composition of subunits in complexes. Extreme sensitivity of the method provides inherent specificity of identification, however in practice it is challenging to realise and truly use the full potential of existing techniques. A comprehensive and reliable mass spectrometry based proteome map is a prerequisite for mechanistic, hypothesis-driven investigations and for large-scale studies. There are two main approaches to study proteins with mass spectrometry: top-down (Tran et al. [285]) and bottom-up (Meissner et al. [189]) proteomics. In the top-down experiments proteins can be studied as intact entities. The advantage of this method is that all modifications that occur on the same molecule can be measured together, which enables direct identification of the proteoform. Despite many advantages of the top-down approach, it is the bottom-up proteomics that has proven to be more reliable and is the most widespread proteomic workflow. In the bottom-up approach proteins are extracted from the source material and enzymatically digested into peptides. Three main methods used within the bottom-up approach are: shotgun (discovery) proteomics, aimed at achieving unbiased coverage of the proteome, using Data Dependent Acquisition (DDA); targeted proteomics using selected reaction monitoring, aimed at acquisition of known peptides of interest; multiplexed fragmentation of all peptides, aimed at generating comprehensive peptide libraries, using data-independent acquisition (DIA) strategies that rely on information coming from high quality spectral libraries.

Because of such a rich choice of methodologies, there is no consistent protocol in obtaining the data. Multiple attempts are showing different mass spectrometry-based strategies for the kinase substrates exploration (Huang et al. [112], Amanchy et al. [5], Coba et al. [44]). Different approaches are complementary and a full set of data can be compiled from multiple sources. This comprehensive characterisation of the proteome could become the most efficient way of obtaining whole proteome and is predicted to be soon a routine experiment

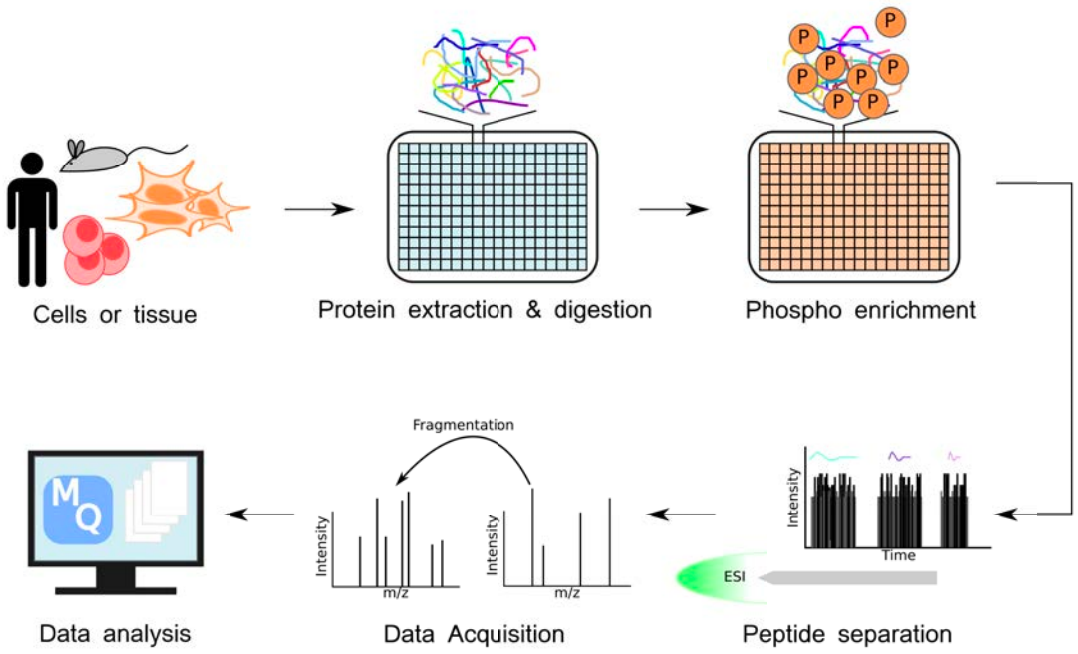


Fig. 1.6 The usual protocol for phosphoproteome analysis includes several obligatory steps. Firstly, peptides are extracted from the provided sample, and prepared for the experiment by phospho enrichment and/or labeled. Peptides are separated using a variety of techniques usually involving chromatography and then forwarded into Tandem Mass Spectrometers. Data analysis with multiple of bioinformatics tools allows for quantitation of spectra.

(Mann et al. [178]). Figure 1.6 is demonstrating the usual necessary steps included in most of the protocols. Given the importance of MS derived phosphorylation information for this thesis, I provide here a brief introduction to mass spectrometry and the steps required to perform a phosphoproteomic experiment.

Mass Spectrometry

A mass spectrometer sorts ions into a spectrum, based on their mass-to-charge ratio. Besides proteomics, MS techniques can be used in many different fields and can be applied to very complex mixtures of ions as well as pure samples. The outcome of the experiment is a mass spectrum, which is a plot of the ion signal as a function of the mass-to-charge ratio. In the phosphoproteomic experiments the spectra are analysed to identify the peptides and localise phosphorylations. General steps from MS protocol usually involve digestion of proteins into peptides, which are separated by liquid chromatography and then ionised. These ions are then separated typically by the use of electro-magnetic field (Figure 1.7). Ions of different

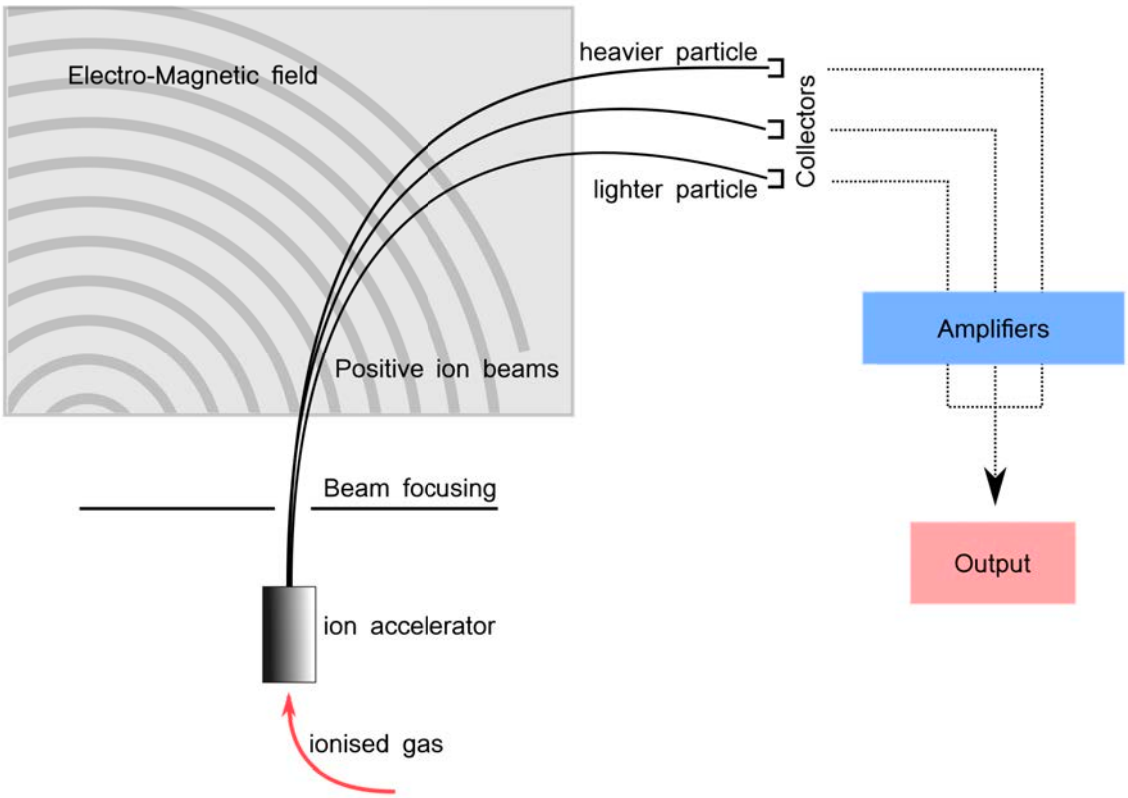


Fig. 1.7 Basic components of a mass spectrometer. The ionizer converts the sample into ions. The ions are then sorted according to their mass-to-charge ratio inside the electromagnetic field and detected. The detector measures the value of an indicator quantity and provides data for calculating the abundances of each ion present.

mass-to-charge ratios will undergo different amounts of deflection in the spectrometer. After the deflection, charged particles can be detected with different tools, commonly with an electron multiplier. Results are displayed as a function of the mass-to-charge ratio, and can be analysed with widely accessible software. The peptides or molecules from the sample can usually be identified by correlating known substrates with a characteristic fragmentation pattern.

A mass spectrometer consists of three main components: an ion source, mass analyzer, and a detector. The two mainly used methods for ionization of protein samples are electrospray (ESI) and matrix-assisted laser desorption/ionization (MALDI). The oldest and still used technique is electron bombardment ionization (EI). Electrospray is the most delicate technique of ionization, which allows even fragile molecules to remain intact. In MALDI samples are embedded with a solid matrix, and the ions are created by pulses of laser light. The chosen method of ionization may influence the experiment, for example,

electrospray produces more multiply-charged ions than MALDI, however they are more likely affected by contaminants, buffers and additives. Recently, hybrid electron-transfer and high-energy collision dissociation (EthcD) fragmentation (Frese et al. [85]) and ultraviolet photodissociation (UVPD) (Fort et al. [82], Robinson et al. [238]) became very popular in large-scale phosphopeptide identification. Mass analyzers commonly used in proteomics include time-of-flight (TOF) or Fourier transform ion cyclotron resonance (FT-ICR). These two methods are usually preferred because of their wide mass range. Detectors record the charge induced by the deflected ion or the current it has produced. Usually some type of an electron multiplier is used as a detector, although Faraday cups or ion-to-photon detectors are also popular. The choice of the detector might be correlated with the choice of the mass analyzer. In proteomics the multipliers are often necessary because of the small masses of peptides.

Tandem Spectrometry

Tandem Mass Spectrometry known as MS/MS, also symbolised as MS2, is a spectrometry method providing high speed and accuracy. In tandem mass spectrometer ions of a particular mass-to-charge ratio are separated in the first stage of mass spectrometry (MS1). The precursor ions that are selected from MS1 are fragmented for the second time and then detected in the second stage of mass spectrometry MS2 (Figure 1.8)

Tandem mass spectrometry experiment can be performed in space or in time. In the first approach, during the selection of ions for MS2, the separation elements are physically separated and distinct. This means between MS1 and MS2 ions are going through sectors, transmission quadrupole, or are divided by the time-of-flight method. In the second in time approach, separation is accomplished by a quadrupole ion trap or Fourier transform ion cyclotron resonance (FTICR). The separation is accomplished with ions trapped in the same place, with multiple steps taking place over time.

Sample preparation

Depending on the sample origins, the preparation for MS experiment can include unique steps. Lysis protocols for cells and tissues usually include inhibition of endogenous phosphatases (Lundby et al. [171]). Because of the lability of phosphate groups, pH and temperature must be highly controlled during the preparation. Due to the low stoichiometric amount of phosphoproteins in the proteome and low ionization efficiency of negatively charged phosphate groups, many approaches have been developed to enrich the samples prior to

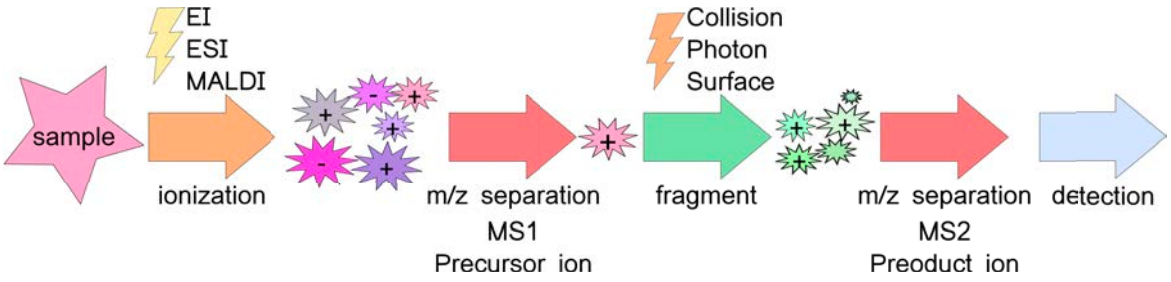


Fig. 1.8 Schematic representation of Tandem Mass spectrometry. The ionized sample is separated at the first stage of mass spectrometry (MS1). Ions of a particular mass-to-charge ratio (i.e. specific peptides) are further fragmented to be analysed in a second stage of mass spectrometry (MS2).

MS analysis (Lemeer and Heck [166]). Historically, the enrichment specificity has been a serious, limiting problem in phosphoproteomic profiling. Nowadays, a variety of enrichment strategies are available to improve the ease and accuracy of measuring changes in protein phosphorylation. Methods of enrichment for detecting phosphosites include immobilized metal ion affinity chromatography (IMAC), metal oxide affinity chromatography (MOAC), polymer-based metal ion affinity capture (PolyMAC), and tyrosine peptide enrichment. IMAC (Andersson and Porath [8], Posewitz and Tempst [227], Zhou et al. [321]) captures phosphorylated peptides using the electrostatic interactions between them and the metal ions (Fe^{3+}), Ga^{3+}, Ti^{4+}, Zr^{4+}, etc.) that are immobilized on the surface of solid supporters. MOAC (Larsen et al. [162], Wolschin et al. [301], Kweon and Håkansson [155]) uses the affinity of metal oxide particles (TiO_2, ZrO_2, Fe_3O_4, etc.) to retain the phosphoryl groups on the solid matrices. Non-phosphopeptides, mostly acidic, have been found to bind both to IMAC and MOAC matrices. To prevent such nonspecific bindings, many protocols include additional steps prior to enrichment that eliminate the unwanted peptides. PolyMAC strategy immobilizes the metal ions not on the solid surface, but on the water-soluble dendrimer (Iliuk et al. [122]). Identification and enrichment of phosphotyrosine sites is still challenging because of their significantly lower abundance (comparing to that of pS/pT). A common method of pY enrichment includes immunoprecipitation (Palma et al. [216], Mijn et al. [193]), however better tyrosine phosphatase inhibitors are still required.

Approaches that allow for the quantification of changes in the phosphoproteome across states (e.g. stimulated vs. unperturbed) can be divided into two categories: containing isotopic labels or label-free (Figure 1.9). In the labeling category, metabolic, chemical or isobaric labels can be distinguished. Isotope labeling methods have an advantage of diminishing the sample preparation bias, because the samples can be combined early within the experiment. Metabolic labels (Figure 1.9.B) usually involve an organism's own cellular

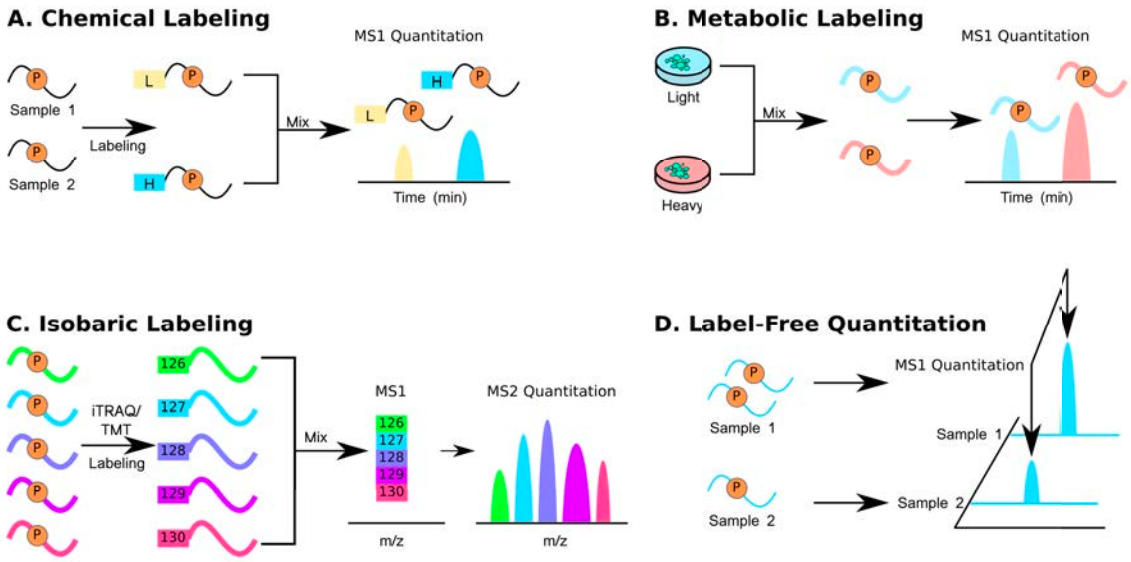


Fig. 1.9 Variety of popular quantification methods that can be used to compare phosphorylations changes across different states in LC-MS/MS.

processes to incorporate isotope labeled amino acids into its proteins through enriched media/diet by 15N isotope (Conrads et al. [47], McClatchy et al. [187]) or in a stable isotope labeling by amino acids in cell culture (SILAC) (Ong et al. [213]). SILAC has been proven very successful in multiple research projects (Rigbolt et al. [235], Iesmantavicius et al. [121], Matic et al. [184]). Chemical labeling (Figure 1.9.A) produces isotopic variants through reductive dimethylation (Boersema et al. [28]) or proteolytic digestion in $_{18}$O enriched water (Yao et al. [313]). This type of labeling may be available for more sample types than metabolic and can be already distinguished at the MS1 level. Isobaric labeled peptides (Figure 1.9.C) cannot be distinguished in the MS1 level as they have the same chromatographic behaviour. The peptides release the reporter ions after the MS1 phase that provide quantitative information (Rauniyar et al. [232]). Commercially available isobaric tags include iTRAQ (Ross et al. [241]) and TMT (Dayon [58]). To analyze the changes in phosphorylations of the sample, a half of it is usually labeled by one of isotopic methods and mixed together with the other half into one sample. Labeled peptides are slowly reaching the MS1 (due to slower chromatography process), which allows the quantification and identification in the spectra. Label-free approach (Figure 1.9.D) analyses two samples separately in order to find differences between spectras. Data analysis tools for label-free samples include MaxQuant MaxLFQ (Cox et al. [49]), Skyline (Schilling et al. [246]) and OpenMS (Röst et al. [242]), among others, that are capable of aligning and normalizing MS data from two different samples.

Peptide Chromatographic separations

Despite the high speed of MS experiments, the depth of phosphoproteome available in a single shot analysis is still limited. The coverage of the phosphoproteome can be extended by separation of peptides into batches of similar masses, and hence performing multiple scans of derived groups. Multidimensional chromatographies have been coupled with MS/MS to analyze low abundand phosphopeptides. Simplification of the complex samples can be achieved with multiple methods, that can be generally divided into two types: HPLC column-based format, and StageTip-based format. In High Performance Liquid Chromatography (HPLC) format, the beads are packed into a column that uses linear mobile gradients. The column format requires HPLC equipment, which might be a limiting step in some laboratories. StageTip (Rappsilber et al. [231]) is constructed with an extraction membrane packed into a loading tip and does not require any additional equipment. It has become popular recently, due to the low cost, high throughput and no need for HPLC equipment.

Spectra Analysis

Peptides can be identified within the spectra by two approaches: peptide mass fingerprinting and *de novo* peptide sequencing. In the peptide mass fingerprinting method (protein fingerprinting) the peptide masses are compared with the database of the known protein sequences masses in order to identify the origin of the peptide. The advantage of this method is that it requires only the mass of the peptide to define the protein of origin. Similarly the downgrade is that peptides originating from unknown proteins will not be matched with any proteins from the database. *De novo* peptide sequencing analysis involves more time and effort but is capable of describing any analysed peptides. It is typically performed without any prior knowledge about the origin of the peptide, and its sequence is usually determined from tandem mass spectrometry. Peptide sequencing assigns fragment ions from a mass spectrum to particular amino acids. There are multiple algorithms that accommodate the spectra identification and they usually try to find the best fitting peptide to the spectrum. This can be achieved by subsequencing (Biemann et al. [23]) or different graph theory algorithms (Fernández-de Cossio et al. [77], Taylor and Johnson [278]). In order to identify phosphopeptides similar, computational approaches have been developed. Most popular site localization tools include MaxQuant PTM score (Cox and Mann [50]), Mascot Delta score (Savitski et al. [245]), and PhosphoRS localization probability (Dorfer et al. [67]). These tools match the spectra with an internal protein sequence database and probabilistic models to predict potential phosphorylation sites. Because it is difficult to certify the matching results, a large

library of phosphopeptides with known phosphorylation sites (Marx et al. [183]) is a popular resource to further examine and improve the accuracy of localization algorithms. In order to re-estimate and reanalyze the spectra from different experiments, public databases are available for deposition of raw data from MS-based phosphoproteomics studies. The most popular one is ProteomeXchange database (Vizcaíno et al. [293]).

1.4.1 Kinase substrates identification with use of Mass Spectrometry

Mass spectrometry experiments largely replaced metabolic labeling with radioisotopes and Edman sequencing to identify endogenous phosphoproteins and phosphorylation sites (Cañas et al. [34]). A variety of methods and protocols for LC-MS/MS setup have been successfully employed to discover and quantify various PTMs, especially phosphorylations (Witze et al. [300]). Kinase-substrate analysis benefits from all the technological progress and numerous applications of mass spectrometry (Kosako and Nagano [150]). Mass Spectrometry can be also used in the final step of almost any other biochemical analysis to identify the phosphosites. Below I am introducing the most popular methods used for kinase specificity discoveries.

Phosphoproteome profiling

In phosphoproteome profiling, protocols of identification of kinase substrates usually involve three major steps. First, the target kinase in samples is inactivated, using pharmacological inhibition, antibody injection, knockout technology or RNA interference. Because of this step, effective and highly specific technique of inactivation is required, which is a major limitation in this strategy. The second step involves measurement of the phosphorylation change due to the kinase perturbation. Unlike in the *in vitro* methods, it is difficult to determine whether a phosphorylation change is due to the target kinase inhibition, or because it is a consequence of inhibition of a kinase upstream in the cascade. In the third step, phosphorylation events are tracked and matched to its kinase targets using bioinformatics tools (Kettenbach et al. [141]). The main disadvantage of phosphoproteome profiling is that it typically does not reveal a precise relationship between the kinase and the substrate, however it narrows down the list of proteins for further explorations (Manning and Cantley [180]).

Synthetic peptide library screening

Mass spectrometry drastically improved accuracy and scanning speed of peptide library screening (Songyang et al. [268], Mah et al. [175]). Solution-phase peptide library screening

is the most popular method to study kinase substrate specificity. The general idea of this method is that the peptide library includes peptides where one fixed position of S/T/Y (designated phosphorylation) is surrounded by random amino acids. Because of exponentially growing combinatorial possibilities of the surrounding amino acid sequences, the detection and analysis of such libraries can be efficiently done using MS experiments. Another technique involves a polymer-bound peptide library, that is pre-constructed on a resin bead and connected by a photolabile linker (PLL). The use of peptide libraries also contributed to the discovery of "anti-motifs", the combination of residues that kinase strongly discriminates (Hutti et al. [119]).

In vitro kinase assay

Instead of using artificial sequence libraries, it is possible to take advantage of natural sequence diversity in a cell lysate. Enzymatically digested peptidome from cell extracts can indicate kinase preferences in the same way as the peptide library. The direct utilization of protein extracts from cells assures natural folding and preexisting post-translational modification in the possible substrate candidates. Combined with the high throughput identification by MS, the cost and labor of such experiments is significantly lower than array-based assays. Similarly to other methods, the challenge remains in distinguishing the direct kinase that caused the phosphorylation of the substrate. Some optimized procedures involve high concentration of purified kinase (Knebel et al. [144], Cohen and Knebel [45]), pulse heating to inactivate endogenous kinase activities (Troiani et al. [287]), and quantification (Kettenbach et al. [141]). To overcome difficulties that each of those enhancements might bring, a technology called Analogue-Sensitive Kinase Allele (ASKA) was developed (Shah et al. [254]). In this elegant approach, ATP is replaced by bulky analoges, that can only be accepted by the mutant kinase of interest. Other kinases are not able to use the ATP analogues, hence cannot transfer the phosphate. This method, successfully eliminates some false positives and has been used to identify novel substrates of some kinases (Habelhah et al. [94]). The main disadvantage of this method is that it requires the generation of the analogue sensitive mutant kinase by mutation and not all kinases can accommodate the required mutation(s) without a loss in activity.

1.5 Current knowledge of the human phosphoproteome

Phosphoproteomics is a relatively new, rapidly growing field of science. Research on discovering the phosphoproteome has started only within the last decade, hence it is not yet fully complete (Ficarro et al. [79], Beausoleil et al. [11], Gruhler et al. [91], Ballif et al. [9], Blagoev et al. [25]). There are few databases containing outcomes of large-scale MS phosphoproteomics experiments. The largest of them is PhosphoSitePlus (Hornbeck et al. [106]), currently containing 233,295 distinct human phosphorylation sites, along with other PTMs. Only 29% of these phosphorylation sites have been identified in more than 1 MS experiment (Needham et al. [205]). Because of the limited ways of studying kinase targets in large scale only 20% of human kinases account for 87% of all currently annotated kinase substrates (Edwards et al. [71]). The fact that around 80% of human kinases have fewer than 20 assigned substrates, while around 30% have no known substrate (Needham et al. [205]) underscores the scale of the knowledge gap. Main limitations in describing entire human phosphoproteome are: lack of exact measurements of total count of phosphosites within the proteome and lack of the estimates of how many phosphosites are statistically possible, annotation of kinases and phosphatases substrates, and unknown functionality of individual phosphosites. The estimates about the phosphorylation site occupancy and its functionality has been presented in several research studies considering the stoichiometry of phosphorylations (Sharma et al. [256], Olsen et al. [212], Wu et al. [303]). Considerations of stoichiometry alone are not enough to estimate phosphorylation turn-over on the global scale. Mathematical methods capable of modeling kinases and phosphatases activities, and providing the energy demands of maintaining cellular balance are still necessary to develop. Establishing kinase-substrate relations is a very important, laborious research task. Knowledge of the cellular context of an active kinase — like presence of a specific stress or reaction to growth factor or cytokine, provides vital information that can assist in follow up experiments that connect phosphosites with their function. At the signal-processing level it is still needed to explore the relationship between the kinases or phosphatases and their substrates, in order to understand basic architecture and direction of information flow in signalling networks. The functional outcomes as a consequence of a protein modification are parts of higher level circuitry that yet have to be fully discovered. Only about 5% of all known phosphosites have annotated function, and although there are around 500 human kinases, 90% of known kinase target sites are annotated to only 20% of the most popular kinases. 150 human kinases, up to this date, still do not have a single known substrate. Establishing substrates of all the kinases is a serious and urging research task, as it provides the first

step into other research problems. There are few reasons why the functionality might be so poorly defined. Firstly, the dynamic changes in phosphoproteome successfully constrict each experiment. Secondly, sample preparation is causing disturbance on its own — phosphatases are usually inhibited, to preserve as many phosphorylations as possible, which might highlight the phosphorylations that should be erased, and hence have no function. This state called hyperphosphorylation may indicate a lot of sites that are not normally active even in a disease state. Another, technical aspect, is the analysis of MS data. It is a recent policy, to always supply the raw data of an experiment, which makes it available for other research groups for reanalysis. Lastly, the scale and impact of non canonical phosphorylations (of eg. Histidine) remains without conclusion. There are existing examples of mechanisms controlled by such phosphorylations, however our understanding of breadth of these signalling pathways is insignificant. Because of the sample preparation and peptide digestion we have no available information about the ubiquitin chains structure when performing MS analysis.

1.5.1 Unknown phosphoproteome

The imbalance of the knowledge distribution about particular human proteins and kinases is hard to ignore. The concept of the "dark phosphoproteome" has been proposed to raise awareness to the fact that despite large phosphoproteomics datasets, we are still unable to understand all the dependencies within signalling networks. Incompleteness of phosphodata as well as the complexity of phosphosignals still prevents us from addressing upstream kinases or downstream functions to most of the known phosphosites. However, even with the very well-studied kinases new studies keep showing new important substrates and functions, and the amount of newly discovered phosphosites grows every day. The depiction of overwhelming quantity of knowledge we already obtained from the few kinases that have been analysed for years can be dwarfed with the realisation that there are still fewer than 20 substrates assigned to around 400 remaining kinases. The opportunities of discovering crucial knowledge within unknown pathways regulated by the "dark kinases" can reveal major insights into cell function. This expectation is supported by the fact that the number of known substrates does not correlate with the number of phenotypes reported for genomic lesions of kinases (Hornbeck et al. [106], Koscielny et al. [151]).

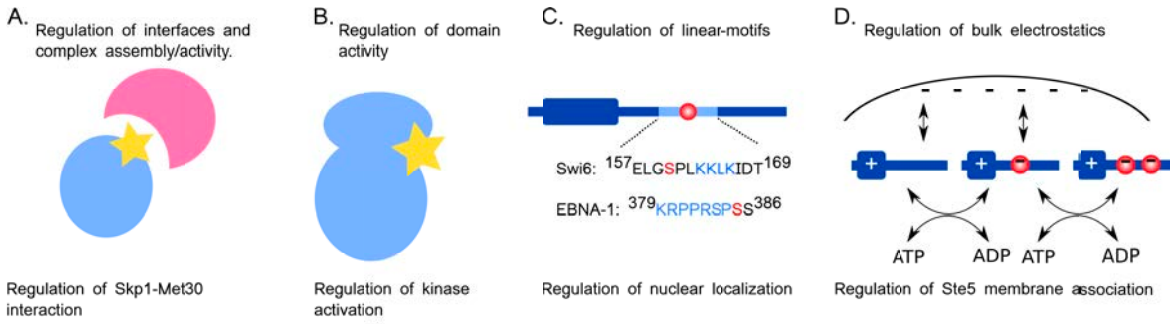


Fig. 1.10 Phosphorylations can regulate protein activity through different mechanisms.
A. Phosphorylation can inhibit or enhance complex assembly and hence its activity.
B. Phosphorylation can induce allosteric changes in a domain that can cause an activation
C. Linear motifs regulating the activity can be created or enhanced by phosphorylations
D. Phosphorylation can alternate protein function by changing its overall charge.

1.6 Functional relevance of phosphorylation

Functional phosphorylation can be characterised by changing the activity of the protein through different mechanisms. Common examples of such molecular switches can be described as regulating interfaces of the protein, directly regulating protein activity, regulating linear-motifs or regulating the bulk electrostatics of the protein. Regulation of the interface of the protein can influence complex assembly by activating or inhibiting binding to other protein, or regulating protein activity. Example presented in Figure 1.10.A shows how phosphorylation enhances protein binding. Phosphorylation of a residue in isocitrate dehydrogenase inactivates the enzyme by blocking the substrate binding (Hurley et al. [118]). Activation of the kinase is achieved by phosphorylation within the activation loop which changes its conformation (Figure 1.10.B). The other classic example is glycogen phosphorylase which when phosphorylated undergoes reorganization that allows access of the substrate to the catalytic site (Barford et al. [10]). Phosphorylation can enhance or be a part of sequential motifs, e.g. the ones that regulate the nuclear localization (la Cour et al. [156]) (Figure 1.10.C). Phosphorylation can also change protein function without the conformational change, but through the bulk electrostatics (Strickfaden et al. [273]) (Figure 1.10.D). One of such examples is MAPK cascade scaffold protein Ste5, which regulates the mating pheromone responses in S. cerevisiae. Multisite phosphorylation of Ste5 near the polybasic membrane-binding domain prevents binding to the inner leaflet of the plasma membrane.

A lot of phosphosites might not be sufficient on their own, but be a part of a logical gate — multiple phosphorylations acting as an organised system (Ferrell and Ha [78], Deshaies and Ferrell [63]). The cross-talk between different types of PTMs, can greatly disturb a simple

signal analysis, which can produce outputs more complex than the sum of its ingredients (Markevich et al. [182], Huang and Ferrell [110]). The fact that only a small percent (5%) of all known phosphosites has an assigned function can be linked to few factors. Despite the claim that a portion of phosphoproteome might be "silent", there is enough premises to assume that most of the phosphosites should be functional. If the phosphorylation causes any evolutionary advantage (in protein's function or interaction) then the phosphorylation will be conserved along with the function.

Understanding the structural mechanisms that ensure sufficient specificity of the kinases is essential for investigation of signal transduction integrity. Experimental evidence suggests that despite all of the controlling mechanisms, kinases are to some degree nonspecific and may phosphorylate substrates without a biological function (Landry et al. [159], Beltrao et al. [15]). Given that some phosphosites may have no relevance to fitness, devising ways to rank sites according to functional importance is a crucial research question (Needham et al. [205]).

1.6.1 Functional annotations of phosphorylations

One approach to determine the functional relevance of protein phosphorylation is to study their regulation across specific conditions. Mass spectrometry approaches can measure changes in PTM abundance, especially in phosphorylation, under different conditions (Choudhary and Mann [40]). Such quantitative experiments analyse the role of the modification in a specific condition, placing the phosphorylation in the context of a specific stress or other cellular stimulation conditions. Quantitative approaches have been used in many research projects to find PTMs involved in different cellular processes such as DNA damage (Matsuoka et al. [185], Bennetzen et al. [17], Bensimon et al. [18], Beli et al. [13]), cell cycle (Olsen et al. [212]), and stem-cell differentiation (Rigbolt et al. [235]). Experimental analysis not only reveals unknown kinases and networks, but also provides the in-depth characterization of well-studied kinases. An example of such re-discovering has been shown in a wide PTMs study of human skeletal muscles under the influence of high-intensity exercise (Hoffman et al. [101]). Another studies by Yang and colleagues uncovered the long sought-after mechanism by which growth factors can increase mTORC2 activity (Yang et al. [308], Humphrey et al. [115]). An *in vivo* knockout of NUAK1 coupled with phosphoproteomics has shown that the NUAK family SNF1-like kinase 1 in skeletal muscle inhibits insulin signaling (Inazuka et al. [125]). Similarly, a knockdown of a kinase TBK1 in lung cancer cells, shows its prosurvival signalling by direct phosphorylation and activation of Polo-like kinase 1 (PLK1) (Kim et al.

[143]). In the follow up study, involving MS-based quantification of Rab immunoprecipitation, functional consequences of site-specific phosphorylations have been revealed (Steger et al. [272]). Studies usually include mutations, knockdowns or knockouts of particular kinases, and phenotypic assays. New, emerging techniques, like the high throughput application of CRISPR/CAS9 system (Cong et al. [46]), simplified large-scale knockout studies (Bi et al. [22], Mali et al. [176]). Due to the fact that the use of multiple clonal cell lines for each CRISPR knockout changes the global protein abundances, suitable controls are necessary. One of the possibilities are rescue experiments, where the deleted protein is restored to provide the convincing control. Such experiments demonstrate that any effects induced by the generation of the cell line are adequately corrected for. Functionality of CRISPR/CAS9 system together with proteomics has been proven in multiple experiments, also in discovering previously unknown substrates of kinases (Schmid-Burgk et al. [248]).

1.6.2 Functional predictions of phosphorylations

To explore the functions of phosphosites multiple biological approaches has been engaged, including mutation studies and experimental phenotypic assays. Biological approaches can only be used in limited cases, and cannot be extended to study each possible phosphorylation. To help overcome this bottleneck, many computational approaches for prioritizing phosphosites has been designed and they include e.g. identification of phosphosites that are highly conserved (Studer et al. [275]) or positioned at the interfaces (Šoštarić et al. [270], Betts et al. [21], Nishi et al. [207]). Other computational approaches include machine learning algorithms, that can include gene essentiality (Seringhaus et al. [252]) amongst other features. Several unique features of the phosphosites have been identified as good predictors of biological impact, and studies (Ochoa et al. [211], Xiao et al. [304]) have shown that combining features into integrated prediction can improve functional prioritization. Ochoa and colleagues integrated 59 features into a functional score but some of them proved more relevant than the others. Amongst the ones with the most predictive power were protein abundance, residue conservation score (SIFT), phosphosite age and protein length. Similarly in Predict Functional Phosphosites (PFP) project (Xiao et al. [304]) features like evolutionary conservation, association with the kinase and the structural surroundings proved to be good predictors. Neural network models like in SAPH-ire NN (Structural Analysis of PTM Hotspots) (Torres et al. [283]) are also proven to be very effective. Studies that show predictions based on sequential information and additional discriminants have proven very successful. Additional to protein sequence information, the protein-protein interaction information (Linding et al.

[169], Wagih et al. [295]), drug treatment studies (Kanshin et al. [136], Imamura et al. [124]), or phosphoproteomics time series (Yang et al. [309]) are particularly powerful in identifying high quality substrates.

1.7 Cross talk

Although PTMs are often known and thought as acting in an independent manner, there are many examples of combinatorial outcomes of multiple PTMs of same or different types. Cross talk can occur when PTMs promote or inhibit modification of other sites (positive/negative cross talk), either in close proximity or at distal locations on the protein, or through competition for the same residue/binding space. Many proteins can be simultaneously modified by multiple events of the same PTM or its combination. It has been observed that PTMs can act combinatorially when coexisting on a single molecule, or cooperate and coordinate protein interactions. There are few known, general rules of how PTMs can mediate crosstalk, however this area of research remains deeply understudied. Many of the known examples originated while studying histone tails in Histone 2A (Goldknopf et al. [89]). Because of the density and variety of PTMs, including acetylation, methylation, biotinylation, ubiquitination, NEDDylation, SUMOylation and phosphorylation, histone tails were the first proteins with characterized PTM crosstalk events. This remarkable variety of PTMs makes it a perfect, yet very complicated object of study, that can translate into other proteins. However, the extent to which the principles that apply to histone PTM crosstalk are utilized by other proteins remains to be determined. Cross talk is usually described with the help of assigning proteins specific roles of a reader, a writer and an eraser (Pawson and Nash [219], Creixell and Linding [52]). A reader is a domain capable of recognizing the PTM, writer is responsible for transferring (creating) the PTM and eraser can remove the PTM. In particular, the distinction between the readers, writers and erasers depends on existing PTM pattern of the target protein (Figure 1.11). PTM cross talk can generally be divided into positive and negative forms (Figure 1.11) (Hunter [117]).

General tendencies help to distinguish positive cross talk, which is a PTM event that can serve as a signal for the addition of the second modification. The simplest and most well-known examples are e.g. phosphorylations requiring initial phosphorylations on a nearby positions, phosphorylation-dependent ubiquitination, or phosphorylation-dependent SUMOylation. Negative crosstalk usually means a direct competition for modification either of a single residue in a protein, or different residues which results exclude each other's consequences (two different modifications can be exclusive). This can mean that a single

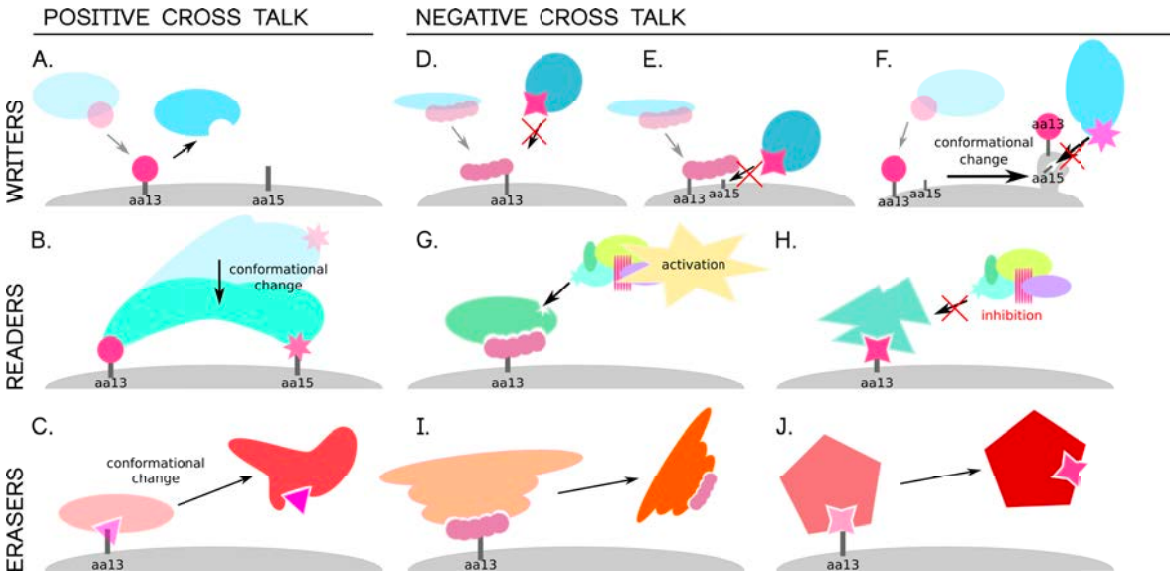


Fig. 1.11 Classification of PTM crosstalk. Depending on their function, regulatory domains can be classified as writers, readers and erasers. Positive cross talk is depicted in A, B and C. A. A writer (blue) attaches a PTM (pink) to an amino acid (aa13) on a target protein (gray). B. A reader (green) binds to PTM on aa13, which induces a conformational change in the domain allowing for further interaction with another PTM on aa15. C. Eraser (red) protein removes a PTM form the original protein, which induces its conformational change. Negative cross talk, described in D-J can be subdivided into direct competition and indirect effects. D. Two different PTM can compete for the same residue (aa13). E. Two PTM are not competing for the same residue, but the attachment of a PTM to aa13 is masking the second PTM binding site (aa15) from its writer. F. PTM on aa13 induces a conformational change that conceals aa15 from its reader. G. Different downstream events can be initiated depending on the initial PTM attachment - after PTM on the target protein the reader triggers an activation pathway. H. PTM can block the target site inhibiting the downstream pathway. I,J. Attached PTM can be removed by the eraser.

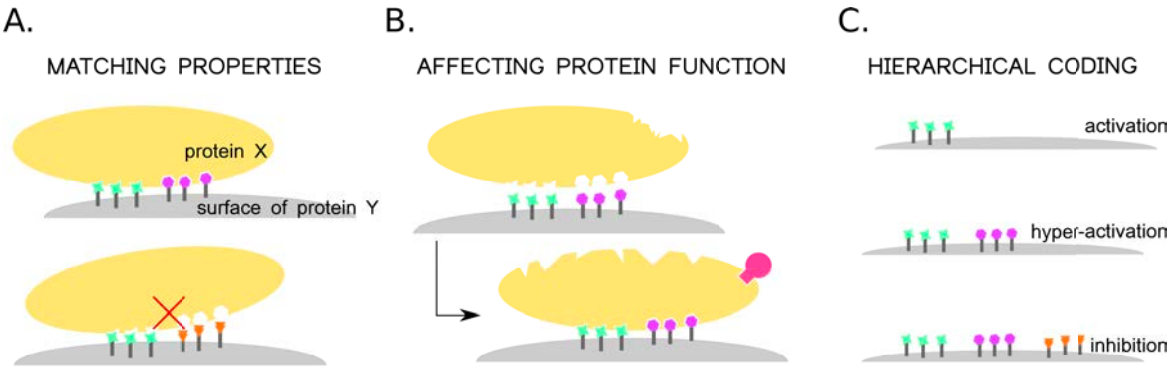


Fig. 1.12 Hypothetical example of combinatorial cross talk of PTMs in proteins. A. Distinct pattern of binding matches only one particular kind of protein, and prevents binding of other substrates. B. PTM induced protein binding can induce a conformational change in the one of the proteins, allowing for further interactions and altering its functions. C. Hierarchical ranking of PTM code can modulate the final function of independently of individual functions of PTMs.

lysine can be a target for ubiquitination or SUMOylation, or acetylation, or methylation. In case of ST phosphorylations, kinases can compete with e.g.O-linked N-acetylglucosamine residues, that attach to specific ST in many types of proteins. Definition of a general, universal PTM code is still under debate (Benayoun and Veitia [16]). PTM patterns have a strong context dependence and poor predictability, which still presents a great obstacle in forming general rules of readable PTM code. The combination of different PTMs on the protein surface could, in principle, form a highly regulated interface that could be dedicated to specific effectors to recognize, and initiate downstream pathways (Sims and Reinberg [264]). In consequence, PTMs can form interfaces that intentionally block the downstream response (Gu and Zhu [93]). This strategy allows a versatile regulation of cellular mechanisms from relatively limited number of genes and molecules, extending the size of the pathways. Examples of such combinatorial crosstalk are shown in Figure 1.12.

1.8 Ubiquitination and phospho-ubi cross talks

Ubiquitin is a 76 amino acids long protein, and the process of its attachment to proteolytic substrate is called ubiquitination. Historically, the attachment of ubiquitin to a Lysin in a target protein is recognised as a marker for degradation by the proteasome. Since the classic work from Ciechanover and colleagues in 1984 (Ciechanover et al. [41]), describing the first

observations of ubiquitin tagged degradation, it has been proven that the post-translational addition of ubiquitin to another protein has many other functions. The process requires 3 steps that include 3 enzymes (in contrast to phosphorylation, which only needs a kinase), known as E1s, E2s and E3s: ubiquitin-activating enzymes, ubiquitin-conjugating enzymes, and ubiquitin ligases, respectively (Figure 1.13). Both, the first and the second step of ubiquitination do not require nor determine the target protein. The first step is activation, where in a two-step reaction ATP dependent E1 ubiquitin-activating enzyme binds both ATP and ubiquitin, and produces an intermediate. The intermediate is transferred to an active site cysteine residue of E1 with release of AMP. There are two enzymes capable of activating ubiquitin in human: UBA1 and UBA6. In the second, conjugation step, E2 ubiquitin-conjugating enzyme catalyses the transfer of activated ubiquitin to the active site cysteine of E2. This step requires E2 binding to both activated ubiquitin and the E1 enzyme, and it is yet completely independent and undetermined what substrate will get ubiquitinated. There are 35 currently known human E2 enzymes, and they all can bind both of the E1 versions. In the final step E3 ubiquitin ligase creates an isopeptide binding between a lysine of the target protein and the C-terminal glycine of ubiquitin. There are 500-1000 estimated E3s in human, which have some substrate specificity towards the E1 and E2. E3s ligases are classified into four families: RING-finger, closely related U-box, HECT, and PHD-finger. The most populated one is the RING-finger family. E3 enzymes function as the substrate recognition modules, and interact with both E2 and the substrate. The specificity is thought to be determined by the proteins interaction and their 3D structure more than through its short sequential motifs (different than in phosphorylation). Mutations and anomalies in E3 ligases are often attributed to many different types of cancer.

The substrate protein can also be a ubiquitin itself, which results in formation of a ubiquitin-linkage or di-ubiquitin. Ubiquitin has a few lysins (and initial Methionine) that allow additional attachment of another ubiquitin. Polyubiquitination associated with K48 and K29 is known to be related to degradation, while other residues associated polyubiquitinations (e.g. on M1, K63, K11, K6) and monoubiquitinations can regulate multiple other processes, such as translation, trafficking or DNA repair. During MS experiments, the structural information about ubiquitin is lost, and only the knowledge about which ubiquitinated Lysine in the target protein remains. Once attached to a substrate, ubiquitin can be a subject to further modifications, including multiple ubiquitinations as well as phosphorylations and other PTMs, which in general is referred to as the “ubiquitin code”. The multitude of distinct signals that are a possible when the attached ubiquitin is modified further is mind-boggling, however it is not yet established how long these chains can grow. Ubiquitination is considered

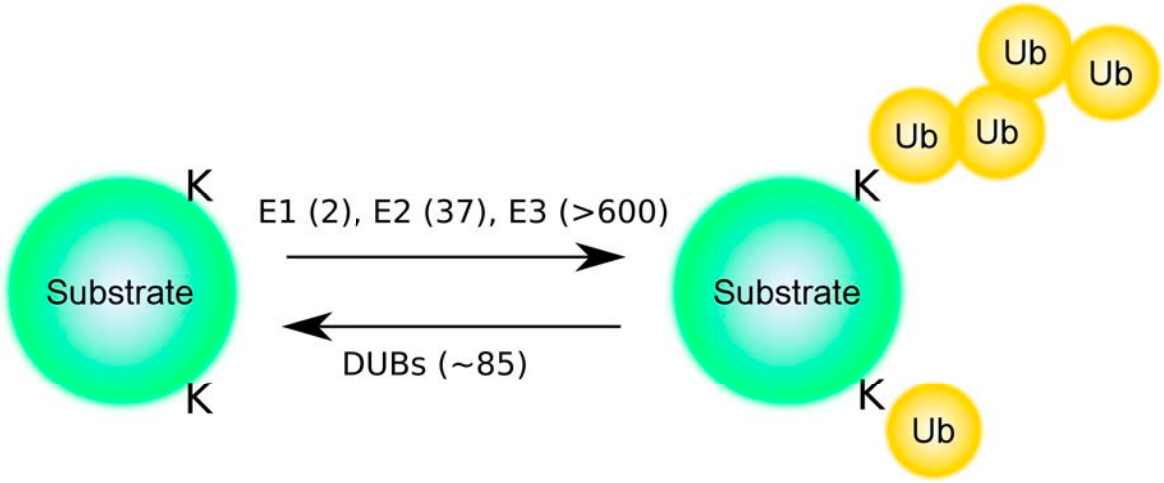


Fig. 1.13 Ubiquitination requires three enzymes (E1, E2, E3), substrate can be mono- or poly-ubiquitinated on Lysine residue. Deubiquitinating enzymes are responsible for the ubiquitin recycling.

one of the most abundant PTMs and its outcomes can influence virtually every protein in every cell. Recent studies confirm the general dogma, that the major task of ubiquitination is tagging for proteasome degradation. However the minimal threshold of (poly)ubiquitination that decides the substrate destiny is not yet fully determined. PTMs that are related to ubiquitination, like SUMO- and NEDDylations are simpler (but still laborious) to study because of the smaller number of dedicated enzymes. Chains constituted from different ubi-like proteins are known to exist, which adds significant complexity to the system and indicates crosstalk between a great number of modification types. Removal and recyclement of ubiquitin from the substrate protein is a very efficient mechanism performed by DUBs (deubiquitinating enzymes). There are more than 100 DUBs in human, however they all have only a few substrates assigned per enzyme (out of hundreds possible).

General subsets of phosp-ubi cross talk can be divided into 3 main categories: phosphorylation can positively or negatively regulate the activity of an E3 ligase (responsible for Ub transfer), phosphorylation can promote substrate recognition by an E3 ligase (phosphodegrons), phosphorylation can influence ubiquitination by regulating subcellular localization of the substrate (through phosphorylation-dependent transport). Phosphodegrons are short motifs containing phosphorylation that mediate phosphorylation-dependent recognition by an E3 ligase. One of the main ubiquitin ligases that can recognise phosphodegrons is the SCF (Feldman et al. [76], Skowyra et al. [265]). Other categories can include the regulation of E2 activity by phosphorylation (both activation and inhibition), regulation of DUB activity by phosphorylation, and phosphorylation of ubiquitin itself.

1.9 Disruption of signalling in disease

Disrupted phosphorylation mechanisms can be an indication of many diseases, including cancer (Blume-Jensen and Hunter [27], Rikova et al. [236], Zanivan et al. [318]). There is a number of clinically approved drugs targeting protein kinases e.g. in cancer treatment (Wu et al. [302]), however the number of researched, druggable kinases is embarrassingly small. Dysregulated phosphorylation signalling has also been linked to Alzheimer's disease (Grundke-Iqbal et al. [92], Eidenmüller et al. [72]), and diabetes (Meyerovitch et al. [192], Danielsson et al. [55]) amongst others. Integration of data from ClinVar database (Landrum et al. [158]) with the phosphosites by Needham and colleagues (Needham et al. [205]) shown that 762 pathogenic human mutations, that are associated with 383 diseases lie on known phosphorylation sites, and only 25% of these have a known upstream kinase. Understanding the effects and regulatory networks of these and other phosphosites will provide better druggable targets.

Missense mutations in cancer

Many efforts have been spent on elucidating the genetic basis of cancer. Large-scale sequencing research such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium, as well as many smaller scale projects provided the community with the sequence of exomes or genomes of many human tumours. The analyses of tumor genomes can indicate highly mutated genes, that are expected to have primary roles in the development of cancer. Millions of somatic mutations found due to TCGA and ICGC projects across thousands of tumor samples provide a perfect resource for mutations analysis, and the amount of available data from other projects is growing everyday. Mutations in cancer generally can be divided into two categories: the driver and the passenger mutations. Genes containing driver mutation(s) promote cancer development because of selective growth advantage, while passenger genes despite high mutations levels have no effect to fitness. Cancer genes are usually characterised by overall high mutational burden, clustering within the linear amino acid sequence, or mutation enrichment in evolutionary conserved sites. It has also been shown that the clustering within the linear sequence can be extended to clustering relative to the 3D structure of the protein product (Kamburov et al. [133]). Assigning individual cancer mutations to protein structures can help identifying new cancer proteins and the functional importance of mutations, based on their spatial location in the protein in relation to other PTMs. A common example of such structure application is the mutation clusters at protein interaction interfaces, disturbing molecular interactions (Kamburov et al. [133]). There are

many approaches (statistical, machine learning etc.) to distinguish driver from passenger mutations within the same gene (Merid et al. [190], Vogelstein et al. [294], Akavia et al. [2], Beroukhim et al. [20], Ciriello et al. [42]). Solving this complicated research question may have a lot of clinical implications. The structural approach has its advantages in the fact that mutations clustered in space are more likely reflect positive selection than their randomly scattered counterparts.

1.10 Aims of this thesis

In the first project, described in Chapter 2, I establish the occurrence of the structural kinase recognition motifs. Knowing that kinases usually have a preference for substrates that include specific linear motif, but the cases of such motifs existing also in 3D space has also been described, I analysed two data sets in search for 3D motifs. For this I devised a computational pipeline mapping known kinase target phosphosites to structural models. Using these I identified potential cases where the important specificity determinant residues are not observed in the contiguous sequences of the targets but may exist as a 3D epitope, and performed docking simulations to examine the possible kinase interactions. The 3D epitope examples were found to be rather exceptions than a rule, and the analysis confirms the general preference of linear motif recognition by kinases.

In the second project, described in Chapter 3, I address the problem of the unknown functionality of the phosphoproteome. Following the methodology established in (Beltrao et al. [15]) I present updated version of the algorithm predicting functional hotspots. The analysis of regions within domain families that are recurrently phosphorylated across different instances highlight regions termed hotspots. To better predict phosphorylation of high functional relevance I analysed phosphosites that are highly conserved across species within protein domains families using phosphosite data for a total of 40 eukaryotic species. A total of 241 domain regions were identified as hotspots within 162 domain families that were then mapped to proteins structures. These regions were shown to overlap with important structural features (i.e. protein interfaces and residues near or at catalytic sites).

In the third project described in Chapter 4, I analyse ubiquitination and mutation hotspots in order to find cross talk examples with the phosphorylation hotspots. To further study the regulatory regions of protein domains I searched for regions of conserved ubiquitination and/or a high degree of recurrent mutations found in cancer. Of 68 domains that had enough data for analysis of all 3 types of hotspots I present the analysis of interesting cases and domains containing overlapping PTM and/or mutational hotspots.

Chapter 2

Kinase specificity motifs in 3D

2.1 Introduction

Kinase specificity of substrate recognition is determined by structural and chemical characteristic of both, substrate and the kinase (Ubersax and Ferrell [289]). Across 518 known human kinases (Manning et al. [181]) its general fold is quite similar, and the substrate binding specificity is often determined by small changes in the binding pocket. The kinase is thought to recognise a sequential, contiguous motif around the phosphorylation site (P-site) (Knighton et al. [145]). The recognition usually occurs with around five amino acids on either side of the phosphosite, although only a small number of positions contributes strongly to the binding preferences (Pearson and Kemp [220], Pinna and Ruzzene [225], Amanchy et al. [4]). Some of the kinases have a well-established recognition motif, e.g. CMGC kines, which tend to phosphorylate serine and threonine residues with proline at position +1 relative to the p-site (Kannan and Neuwald [134]). As described in the introduction chapter, kinases can also recognize their substrates through additional mechanisms that include: docking motifs (binding motifs that interact with regions distal to the kinase active site); interaction with protein scaffolds; co-expression and co-localization (Biondi and Nebreda [24], Holland and Cooper [102]). Binding motifs distal from the substrate P-site may increase the affinity of the kinases for specific substrates by increasing local substrate concentration around the kinase, or causing allosteric effects that either positively or negatively regulate kinase activity. Similarly, protein scaffolds can act as organizing platforms that coordinate both, the kinase and the substrate. Interestingly, novel kinase-substrate interactions can be engineered with artificial scaffolds (Zeke et al. [319]), further emphasizing the role of scaffoldings for determining kinase-substrate interactions. Conditional docking sites provide an additional level of control, ensuring the proper timing of substrate phosphorylation. These conditions may

include previous phosphorylation of other phosphosites close or distant from the required p-site (Elia et al. [73]). Finally, kinase specificity can be easily enhanced by localization. Distinct localization in subcellular compartments or structures can increase concentrations of the correct substrates, as well as separate kinases with overlapping recognition motifs (Shirakata et al. [261]).

Various experimental approaches have been developed to identify substrates of protein kinases, however knowledge coming from such experiments covers only a fraction of all substrate information in a few model organisms (roughly 5% of known human phosphosites are connected to one or more kinases). This is why the prediction of kinase specificity can play a major role in studying signalling relationships. Common computational approaches such as scan-x (Chou and Schwartz [38]), Scansite (Obenauer et al. [210]), NetPhorest (Mok et al. [197]), KinasePhos (Huang et al. [111]) and others, use phosphorylation data from curated kinase targets found in databases e.g. PhosphoSitePlus (Hornbeck et al. [104]), Phospho.ELM (Biondi and Nebreda [24]) etc. to predict kinase specificity. The specificity inferred by these methods do not fully account for all of the known target phosphosites of a given kinase. There are several known target sites that do not conform to the main linear sequence determinants as modelled by these methods. This could suggest that there might be folded epitopes that are contributing to binding to the active site or that the recognition for these sites occurs mostly with other regions of the kinase.

Although most interactions between the active site and the P-sites are not thought to require a folded epitope, new structural mechanisms of achieving high kinase-substrate fidelity are still being discovered. In a recent study Duarte and colleagues demonstrated a noncontiguous recognition motif for the PKC kinase (Duarte et al. [69]). Although PKC prefers a linear motif of a basic amino acid at position -3 or -2 relative to the P-site, it has been shown that a basic amino acid is able to form a noncontiguous target motif recognized by the kinase. Another example of the recognition of a folded epitope is the phosphorylation in cofilin/actin-depolymerizing factor (ADF), where the target P-site is directly oriented to the correct position for phosphorylation by an additional interaction between the LIM kinase C-lobe and cofilin helix (Hamill et al. [95]). These examples suggest that the recognition of folded epitopes or “3D motifs” can occur more frequently than indicated by current models of kinase target recognition. However, the extent by which such non-contiguous recognition can occur is not known. Following this study I decided to systematically investigate the frequency of such 3D motifs among known kinase target sites.

2.2 Results

2.2.1 Potential non-contiguous 3D motifs found in *in vitro* kinase target sites

In order to study the relevance of 3D motifs for kinase target recognition I compiled lists of experimentally validated kinase targets sites. The first source of kinase targets information that I analysed has been described by Imamura and colleagues (Imamura et al. [123]). In this study the combination strategy of *in vitro* kinase reaction followed by LC-MS/MS analysis has been applied to identify substrates of three kinases: AKT1, ERK1 and PKA. A total of 9710 phosphosites were identified as potential targets of one of these kinases. Although several of these target sites may not constitute relevant *in vivo* targets, they should still represent the possible regions in proteins that can be recognized and phosphorylated by these kinases. For each of the kinases a sequence motif was defined using the R version of motif-x (Chou and Schwartz [39]), using all *in vitro* sequential data provided. Since motif-x returns all possible motifs with a score, I chose the most common ones (Methods).

The sequence recognition motif for each of the kinase is different, as shown in Figure 2.1. The serine and threonine phosphosites were analysed separately, however I present here the results for the serine phosphosites. The serine phosphosites set was larger and the results were very similar to those of threonine phosphosites.

To find the structural models that can be used to measure the distance between the serines and motif determinants and establishing 3D motifs, I devised a computational pipeline addressing those questions (Figure 2.2). The pipeline uses a structural database I created, which includes known structures and predicted homology models, on which I mapped the kinase target phosphosites. The structural database includes 12162 proteins, where both solved PDB structures along with the ModPipe (Pieper et al. [222]) predictions summed up to 290,009 models (Methods).

Finally, I mapped the 9710 experimentally determined *in vitro* phosphosites on available structures, defining a total set of 1275 phosphosites with a known kinase regulator and structural information (224, 434, 617 for AKT1, ERK1 and PKA respectively).

Having mapped the kinase target sites to structural models and defined their linear motif preferences I then calculated the frequency of the structurally mapped phosphosites having each of the individual major determinants defined by the motif enrichment (Figure 2.3). As expected by the motif enrichment analysis, AKT1 target sites are more likely to have the AKT1 determinant of R-5 (19.6%) than the PKA or ERK1 specific determinants (5% to

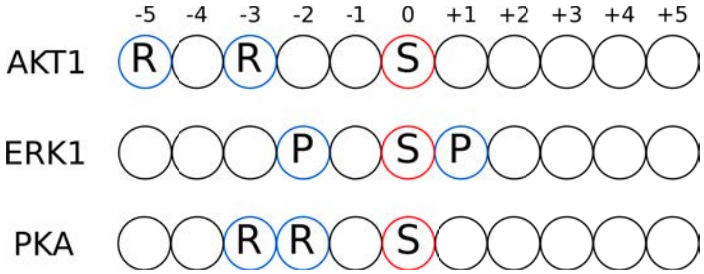


Fig. 2.1 Sequence motifs obtained with the motif-x algorithm with the data from the in vitro set for AKT1 (R-5, R-3), ERK1 (P-2,P+1), PKA (R-2,R-3). The determinant for R-3 overlaps for AKT1 and PKA.

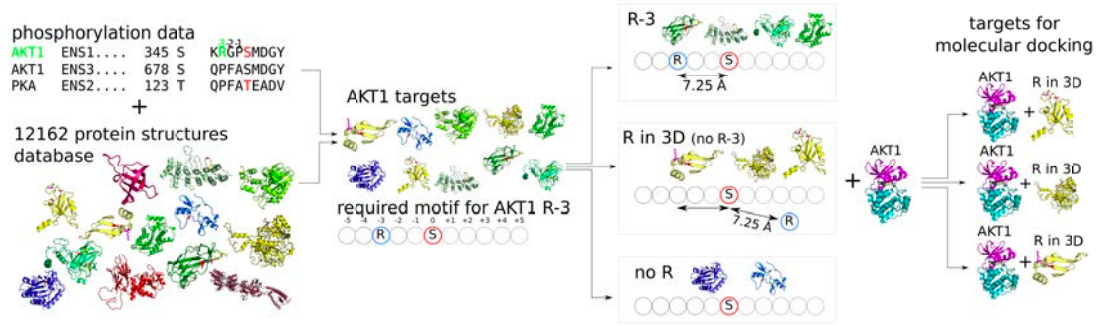


Fig. 2.2 Pipeline for the identification of potential cases of 3D recognition of kinase target sites: 1. Connection of sequential and structural data into one database. 2. Establishing sequential target motifs for all kinases, e.g. AKT1 was found to prefer R at position -3 relative to the phosphosite. 3. Measuring distances between the phosphosite and the R on position -3 in structures with that sequential motif (set "-3R"). Selection of those non sequential proteins which have R in the particular distance (set "R in 3D"). 4. Validation of proposed 3D motifs — molecular docking with a kinase structure.

	pS total:224		pS total:434		pS total:617	
LINEAR	AKT1 (R-5, R-3)		ERK1 (P-2, P+1)		PKA (R-2, R-3)	
R-5	44	19.64%	22	5.07%	81	13.13%
R-3	85	37.95%	28	6.45%	190	30.79%
R-2	35	15.63%	12	2.76%	133	21.56%
P-2	7	3.13%	50	11.52%	14	2.27%
P+1	9	4.02%	151	34.79%	5	0.81%

Fig. 2.3 Data for phosphorylated serines containing one of the single amino acid determinants identified by motif enrichment ("LINEAR"). Each determinant is considered individually. "pS total" is the total number of phosphosites that detected in the in vitro kinase experiments that could be mapped to a structural model phosphorylation data was available for (1275 out of initial 9710 phosphosites). The determinants from the sequence motifs are listed in the first column, as well as mentioned next to its kinase in row 2.

13.13%). The AKT1 determinant of R-3 is shared with PKA and as such it occurs at similar percentages (37.95%) also in PKA (30.79%), while for ERK1 it only occurs in 6.45% of the sites. Similar results are observed for ERK1 and PKA (Figure 2.3). The percentage confirms the dominance of defined sequence motifs, however it is still not a certain feature to determine a kinase specificity. Overall, nearly half of kinase target sites miss both of the determinants (42.41% for AKT1, 53.69% for ERK1, and 47.65% for PKA).

Having established that a substantial fraction of the examined kinase target sites does not have the expected sequence determinants I then considered the possibility that the acceptor may be found instead close in the 3D space. To perform this analysis I first calculated the expected 3D distances for each of the residues that matched the important determinants described above. Figure 2.4 shows the distribution of 3D distances for the determinants if the sequence motifs of each kinase. For example, the R at position minus 5 relative to the phosphosites targeted by AKT1 tend to be at 11.2 ± 2.6 Å away from the phosphosite. These distances were used in the next step to find potential 3D motifs.

For each of the kinase sites that did not match the linear sequence preferences the distances listed in the table in Figure 2.3 were used to search for residues that could be at an equivalent distance and could replace the linear sequence determinant. I hypothesized that it should be more probable to find a dedicated amino acid from the sequential motif in the specific distance of the dedicated kinase than in the substrates of other kinases. For example, it should be more probable to find an R in the distance of 11 Å from the phosphosite for the substrates of AKT1, than for the substrates of ERK1. The frequency of appearance of potential 3D motifs are presented in table in Figure 2.5.

	average distance [Å]	standard deviation
AKT1 (R-5)	11.28	2.63
AKT1 (R-3)	7.8	1.8
PKA (R-3)	7.45	1.88
PKA (R-2)	5.97	0.64
ERK1 (P-2)	6.07	0.65
ERK1 (P+1)	3.71	0.14

Fig. 2.4 Distances defined from phosphorylated serines to amino acid determinants of sequence motifs of each kinase. The distances have been counted between the Cα of the amino acids and summarized as average and standard deviation.

3D	Expected distance Å	pS total: 224		pS total: 434		pS total: 617	
		AKT1 (R-5, R-3)		ERK1 (P-2, P+1)		PKA (R-2, R-3)	
R-5	11.28	107	47.77%	212	48.85%	301	48.78%
R-3	7.8	29	12.95%	48	11.06%	77	12.48%
R-2	5.97	5	2.23%	10	2.30%	19	3.08%
P-2	6.07	4	1.79%	6	1.38%	8	1.30%
P+1	3.71	0	0.00%	0	0.00%	0	0.00%

Fig. 2.5 Data for phosphorylated serines containing at least one amino acid determinant within a defined expected distance suggestive of a potential 3D motif. Each amino acid determinant of each motif was considered individually. The determinants of each motif are listed in the first column, as well as mentioned next to its kinase in the first row. pS total is the total number of phosphosites mapped to structures.

The results are summarized in table in Figure 2.5 showing that there are many Arginines in the R-5 and R-3 distances in all of the substrates, however there are no Prolines in the P+1 distances in any substrates. Overall, the frequency of matching determinants in 3D relates primarily to the distance used for the search with few differences in matching determinants across the 3 kinases. To establish whether the frequency of determinants at the distances defined by each motif depends on the substrates or is just a general feature of the structural distances searched I repeated the procedure for surface accessible non phosphorylated serines.

Unphosphorylated serines were used as controls, coming from the same structural models as the phosphorylated ones and restricting only to surface accessible sites (>20% relative surface accessibility). Figure 2.6.A. shows the ratio between the percentage of phosphoserines that have one of the linear motif determinants and the percentage of unphosphorylated serines having the same determinant. As expected by the motif enrichment analysis the determinants for each kinase occur at a higher than random expectation. This ratio identifies the expected ERK1 preference for proline and AKT1 and PKA preferences for arginine. I then calculated the ratio between the fraction of phosphosites with a potential 3D motif in phosphorylated serines with the same fraction in unphosphorylated serine in Figure 2.6.B. Unlike the previous outcome Figure 2.6.B. does not show different tendencies for different kinases. Most ratios are close to 1, which suggests that the probability of finding such 3D motifs for a phosphosite is similar to those randomly found in unphosphorylated serines. There is a modest but not significant enrichment for Arginine found at approximately 6Å for PKA targets which matches the PKA R-2 preference. There is a similar enrichment for Proline at approximately 6Å for AKT1 but this does not match the preference of AKT1 which does not favour P-2. No 3D motifs have been found for motif P+1 and a significant amount has been found for R-5. It is possible that the distance equivalent to R-5 may be too wide and generate too many false positives while distance for P+1 (3 Å) may be too narrow to find other amino acids within it.

Because of the concerns that the distance for the R-5 motif is too big, and hence generating too many false positives I measured the probability of finding any amino acid in AKT1 substrates in the distances of R-5 and R-3 motifs from the phosphosites. I analysed all the phosphosites of AKT1 substrates and unphosphorylated Serines and measured the frequency of amino acids within two distances 11.28 Å ± 2.63 (which is an average distance of R from the phosphosite in the linear motif R-5) and 7.8 Å ± 1.8 (average distance of R from the phosphosite in the linear motif R-3). The results are shown in Figure 2.7. The probabilities of finding any amino acid are quite similar, and the expected 3D motifs are not distinguishable.

Despite the expectation I did not observe higher probability of finding the amino acid from the dedicated motif (e.g. Arginine) in the particular distance (e.g. of the R-5/R-3 motif

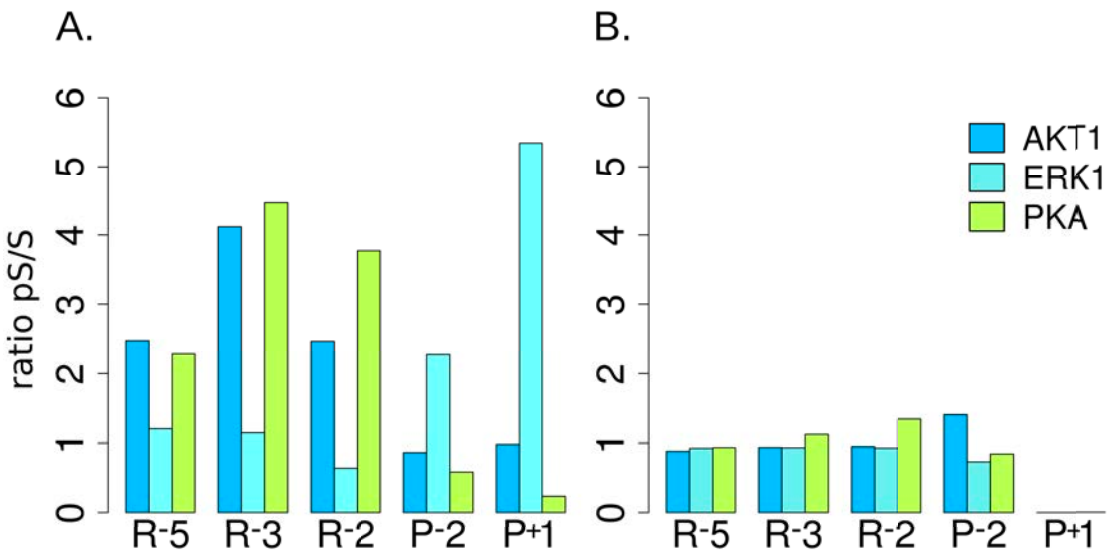


Fig. 2.6 Proportion of phosphosites with possible 3D motifs over unphosphorylated serines with possible 3D motifs. A) sequence motifs are visibly expressed in comparison to unphosphorylated serines. B) 3D motifs are not visibly expressed in comparison to unphosphorylated serines.

AKT1, R-5 motif, 11.28 Å ± 2.63 distance

aa	phosph S	p1	unphosph S	p2	odds ratio
A	105	5.49%	1110	5.55%	0.99
C	10	0.52%	161	0.81%	0.65
D	109	5.70%	1372	6.86%	0.82
E	161	8.42%	1813	9.07%	0.92
F	62	3.24%	541	2.71%	1.20
G	176	9.20%	1566	7.83%	1.19
H	49	2.56%	498	2.49%	1.03
I	54	2.82%	589	2.95%	0.96
K	215	11.24%	1949	9.75%	1.17
L	84	4.39%	1025	5.13%	0.85
M	31	1.62%	284	1.42%	1.14
N	91	4.76%	959	4.80%	0.99
P	92	4.81%	1042	5.21%	0.92
Q	106	5.54%	1037	5.19%	1.07
R	161	8.42%	1727	8.64%	0.97
S	141	7.37%	1638	8.19%	0.89
T	97	5.07%	1110	5.55%	0.91
V	94	4.91%	821	4.11%	1.21
W	12	0.63%	153	0.77%	0.82
Y	63	3.29%	595	2.98%	1.11
total	1913		19990		

AKT1, R-3 motif, 7.8 Å ± 1.8 distance

aa	phosph S	p1	unphosph S	p2	odds ratio
A	27	7.36%	218	5.07%	1.49
C	3	0.82%	42	0.98%	0.84
D	16	4.36%	280	6.51%	0.65
E	18	4.90%	274	6.37%	0.76
F	6	1.63%	88	2.05%	0.80
G	47	12.81%	336	7.81%	1.73
H	10	2.72%	91	2.11%	1.30
I	11	3.00%	104	2.42%	1.25
K	38	10.35%	336	7.81%	1.36
L	14	3.81%	193	4.49%	0.84
M	10	2.72%	52	1.21%	2.29
N	27	7.36%	207	4.81%	1.57
P	19	5.18%	348	8.09%	0.62
Q	17	4.63%	189	4.39%	1.06
R	28	7.63%	680	15.80%	0.44
S	29	7.90%	334	7.76%	1.02
T	19	5.18%	209	4.86%	1.07
V	16	4.36%	162	3.76%	1.17
W	1	0.27%	23	0.53%	0.51
Y	11	3.00%	137	3.18%	0.94
total	367		4303		

Fig. 2.7 Probabilities of finding particular amino acids within distances for R-5 motif and R-3 motif.

motif	kinases	distance [Å]	phosp in models	matching sequence motif	3D motifs
D/E-X-X-pY	SRC, LCK	7.21±1.73	127	39 (30.71%)	28 (22.05%)
D/E-X-pS/pT	PLK1	5.57±0.09	16	3 (18.75%)	0
pS/pT-P	CDK1, CDK2, CDK4, CDK6, CDK9, MAPK3, MAPK, MAPK9, MTOR	3.68±0.21	171	110 (64.33%)	0
pS/pT-X-X-P	ABL1, SYK	9.09±1.3	38	14 (36.84%)	12 (31.58%)
R/K-R/K-X-pS/pT	AKT1, AURKB, AURKA, CAMK2A, CHEK1, CHEK2, DAPK3, MAPKAPK2, PJN1, PRKAA1, PRKCE, PRKCQ, ROCK1, RPS6KA3, RPS6KA5, RPS6KB1, PRKACA, PRKCD	7.25±1.79 6.1±0.69	555	353 (63.60%)	92 (16.58%)

Fig. 2.8 Kinase target sites with available structural models and a known kinase with a determined target motif. In the "distance" column all distances between motif and phosphosite are shown, if multiple then starting from the N-terminus consequently. Number of available phosphosites in structural models are in the column —"phosp in models", phosphosites fully matching the sequence pattern listed in "Matching sequence motif", possible 3D motifs listed in "3D motifs".

of AKT1). I next extended these results to a set of kinase target sites compiled from the literature and of a higher in vivo relevance.

2.2.2 Potential non-contiguous 3D motifs found in *in vivo* kinase target sites

To extend the analysis of 3D motifs to in vivo target sites, human phosphorylation and kinase-substrate relationships have been compiled using PhosphositePlus (Hornbeck et al. [104]), HPRD (Prasad et al. [228]) and Phospho.ELM (Diella et al. [65]) for 2165 unique proteins. All sites were mapped to Ensembl v73 (Herrero et al. [99]), and a single representative or canonical transcript was selected for each gene to remove redundancy. Additionally I examined a phosphorylation dataset from yeast, in order to obtain more working examples, however it did not extend the amount of data, because of the lack of structural models. From the human data I investigated 100 kinases with the larger number of known target sites, for which I could obtain structural models. As previously, for each kinase a sequence motif was defined using the R version of motif-x (Chou and Schwartz [38]), using all *in vivo* data compiled. I have found clear sequence motifs for 32 human kinases for which I could also analyse structural models. The 32 kinases were grouped according to 5 well established recognition motifs that were obtained from the motif enrichment analysis and are presented in table in Figure 2.8. The target motif with largest number of target sites with available structural information is the preference for R/K at positions -2 and -3.

Similarly as in the analysis for the in vitro data, I used the structural information to calculate the average distance of the amino acid determinants in the sequence motif from the phosphosite (table in Figure 2.8) The sequence motifs are defined by motif-x program and described in Methods. These distances were then used to search for the same amino acid within the maximum acceptable 3D distance in models that did not have the corresponding amino-acid determinant within their phospho-peptide sequence. As for the *in vitro* data there are several target sites that do not fully match the expected motif for the respective kinases (388 out of 907 - 43%). A fraction of sites (132 out of 907 - 15%) that lack one of the major sequence determinants have the lacking residue at an equivalent 3D distance (Figure 2.8 "3D motifs"). These cases are potential in vivo human kinase target sites that may be recognized by kinases as a 3D motif.

The previous analysis with the in vitro set suggests that the potential 3D epitopes do not exist at a frequency above random expectation. Using the in vivo set of kinase targets I next tested if target sites fully matching the sequence determinants showed higher evidence of functional relevance compared with phosphosites classified as potential 3D motifs, as defined above. For this I quantified a metric of sequence constraint by predicting the impact of mutating each phosphosite using SIFT (Ng and Henikoff [206]). This tool predicts whether an amino acid substitution affects protein function based on sequence conservation and the physical properties of the amino acids. I could only study phosphosites and determinant residues of 3 of the motifs, since 2 of them had no predicted 3D motif phosphosite. The outcome of this sequence constraint analysis is presented in Figure 2.9. The score provided by SIFT is the probability that a specific mutation (e.g. serine to alanine) is not deleterious, with scores below 0.05 indicating a potential deleterious mutations. For this analysis I calculated for each phosphosite position or determinant residue the average SIFT score for all possible mutations at each of the phospho-acceptor residues. The lower this average score is the more important the position should be. Using this average score I compared the importance of the determinant residues around phosphosites that match the linear motif with the importance of the determinants that could be potentially recognized as a 3D epitope (3D). In addition I also tested a set of amino acid residues that are not part of a linear or 3D motif but come from the same protein and are equally surface accessible. In addition, I also compared the importance of phosphorylated residues with the importance of equivalent acceptor residues. In Figure 2.9, represented with red boxplots, the mutations of phosphorylated serines, threonines or tyrosines (S/T,Y phosp) show, on average, a higher probability of being deleterious than mutations of serines, threonines or tyrosines that are not phosphosites (S/T,Y nonphosp). However, the difference found was not statistically significant. Nevertheless I still tried to

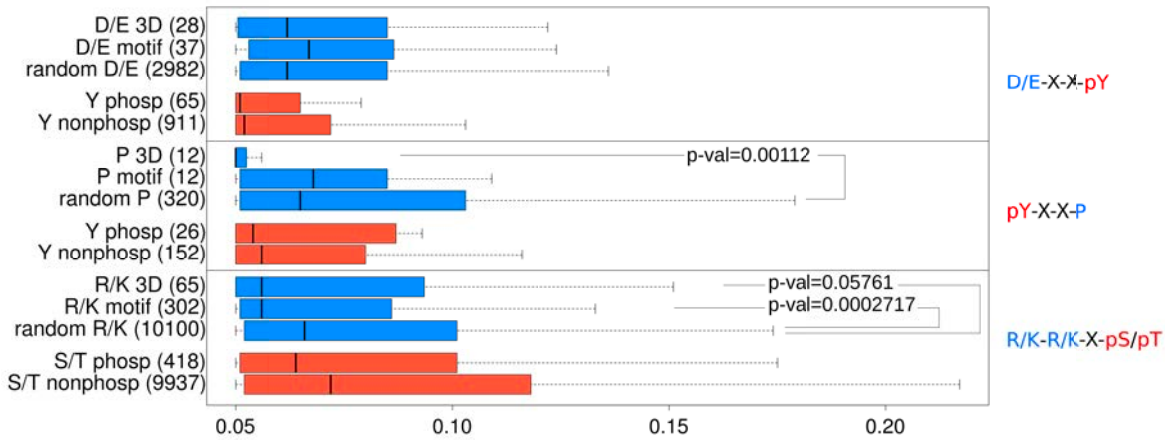


Fig. 2.9 Average SIFT values for phosphosites matching the 3 most represented motifs. Motif residues (D,E,P,R,K) are drawn in blue, phosphosites (S,T,Y) in red. Motif for each plot is presented on the right. In each plot, from the top in blue: "X 3D", where X is the amino acid from the motif — average SIFT score values for the amino acid in the predicted 3D motif; "X motif" — values for amino acids from the motif in the sequential motif, "random X" — values for random amino acids of the same type as for the motif, coming from the same models. In red: "X phos" — average SIFT score values for the phosphosites, "X nonphos" — values for the same amino acids as the phosphosite, from the same model, but not phosphorylated. P-values were obtained using a Wilcoxon test.

measure if there is a significant difference in the conservation of the amino acids that make up the determinant residues.

In all cases analysed the functional importance, as measured by SIFT, of the determinant residues that may form potential 3D motifs were either higher (i.e. lower p-values) or equivalent to that of the residues that are determinants in the linear motifs. In two of the cases, the level of constraint is lower than for a random sample of equivalent amino acids (pY-X-X-P and R/K-R/K-X-pS/pT) but only one of these showed a significant difference (pY-X-X-P). Overall, these results suggest that there is a small trend for the amino acids that I predicted to be part of 3D motifs to be constrained in evolution and therefore could, in some of these cases, represent true determinants for the kinase recognition.

Additional evaluation of whether the predicted 3D motifs are connected to functional phosphosites was performed by looking at phosphosite coregulation with their regulatory kinases. Kinase activity is likely to be increased if its known target sites tend to be up-regulated and vice-versa. My group has shown that the phosphorylation levels of kinase substrates across multiple conditions can describe kinase regulation under different perturbations (Ochoa et al. [211]). Coregulation data coming from an in-house compilation of 32 quantitative studies reporting the relative changes in phosphopeptide abundance after

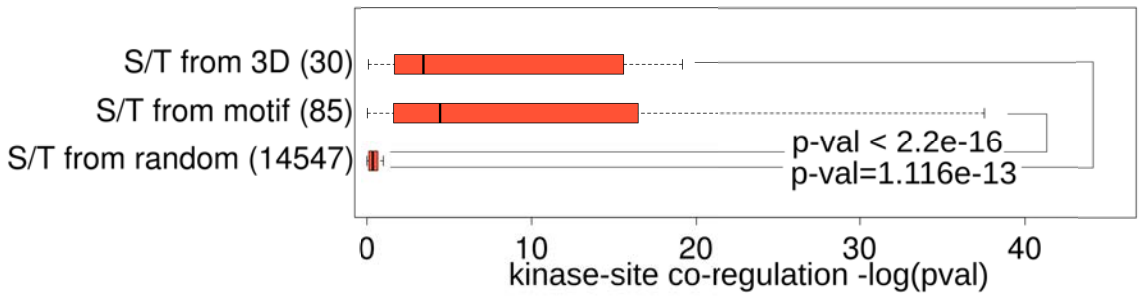


Fig. 2.10 Co-regulation between phosphosites and kinases recognizing the sequence motif R-3,R-2. “S/T from 3D” - phosphorylated serines that have predicted 3D motif, “S/T from motif” - phosphorylated serines that have sequential motif (R-2,R-3), “S/T from random” - phosphorylated serines that have either 3D or sequential motif (sample of 200) correlated with random kinases not recognizing R-2/3. Co-regulation is measured as the negative logarithm of the correlation p-value between the KSEA activity and phosphosites fold-change.

different perturbations was used. From the detected peptides, 113,565 phosphorylated amino acids in 10,868 canonical proteins have been collected. For these sites, 136,464 quantitative changes in phosphopeptide abundance in a panel of 150 biological conditions have been collected. Enrichment of reported substrates on differentially regulated phosphosites has been tested and changes in the activity of each kinase has been predicted for each condition using a Gene Set Enrichment Analysis. The changes in the activity of the kinases of interest were then correlated with the changes in phosphosite abundance for phosphosites matching linear motifs as well as the proposed 3D motifs. I could only obtain sufficient data for the study of the R-2,R-3 motif (presented in Figure 6.).

Phosphosites explained by a predicted 3D motif (S/T from 3D) and by the R-2/3 sequential motif (S/T from motif) present a similar coregulation with the catalytic kinase that is significantly higher than with random kinases. This outcome suggests that the phosphosites predicted to be a part of 3D motifs are equally functionally important as those forming linear motifs.

Docking Results

To further test the predicted 3D motifs I ran docking simulations that explored the capacity of the structures to accommodate the conformation where the 3D epitope binds the active site. The simulations have been performed for chosen kinase-substrate pairs where the substrate has not shown a sequential motif, but a 3D motif within the required spatial distance has been

predicted. Two motifs has been tested - R-2/R-3 in AKT1, CAMK2A, CHEK1, DAPK3, PRKACA, and ROCK1 and D/E-3 in SRC and LCK.

For the motif R-2/R-3 both of the possibilities has been examined - that the R/K in close distance may bind like a R-2 motif or like a R-3 motif. Although this motif might not represent the ideal sequence recognition scenario, it contains minimal sequence motif detected by motif-x (Methods). I decided to include also Lysine predictions to contribute to a bigger number of simulations. The tyrosine kinases SRC and LCK with the D/E-3 motif have only a few structural models available.

Preparation of all the kinase structures has been possible thanks to another PhD student from the lab David Bradley, who predicted the residues that are responsible for binding to the motifs (Bradley et al. [30]). The definition of the amino acids binding the peptides within the motif is based only on structures available in the PDB database. The kinases analysed in his work are slightly different than the ones that I used in all my simulations — to establish the homological binding residues I aligned both of the kinases and found reciprocating residues of binding. To set up the docking simulations HADDOCK protocol allows to choose the predefined virtual springs that are bringing together the amino acids. These springs have been chosen on the basis of the residues that are predicted to bind the structural motifs and are presented for each kinase in table in Figure 2.11. Structures of activated kinases were available for all simulated proteins, however they did not contain the ATP. I aligned them with PKCA structure (4wb8) containing the ATP and copied the ATP into the used models. Initial simulation of minimizing the energy and removing sterical clashes has been performed with the kinase and the aligned ATP in HADDOCK After minimizing the kinases in HADDOCK, standard docking protocol has been performed.

To better understand the spatial localisation of where the peptides bind inside the catalytic pocket, I highlighted the determinants in Figure 2.12. The amino acids described in Table 8 have been set up to have the affinity (through the virtual spring) towards the motif and the phosphosite. Depending on the motif, the number of springs was 5 (for R/K-2 and D/E-3) or 6 (for R/K-3).

I considered a simulation successful if all of the distances defined by the strings have been smaller than the threshold. To establish the threshold of possible binding I used data published in (Ching, 1989), which demonstrates the length of linearly stretched amino acids. For the threshold I used the sum of both stretched out amino acids involved in the binding. The distances between the amino acids after the simulation have been measured between the alpha carbons. This allows for an objective judgment without considering the rotation of the sidechain. Each HADDOCK simulation produced 200 models, from which I arbitrarily

kinase	pdb code	ATP, P0 (aa in pdb)	R/K-2 (aa in pdb)	R/K-3 (aa in pdb)
AKT1	3qkl	499, 274, 276	278, 341	234, 236, 278
CAMK2A	2vz6	401, 135, 137	139, 205	96, 98, 139
CHEK1	2yex	401, 130, 132	134, 200	91, 93, 134
DAPK3	3bhy	401, 139, 141	143, 209	100, 102, 143
PRKACA	4wb8	401, 166, 168	170, 230	127, 129, 170
ROCK1	3v8s	499, 198, 200	202, 270	160, 162, 202
kinase	pdb code	ATP, P0 (aa in pdb)	x	D/E-3 (aa in pdb)
SRC	1y57	601, 386, 388	x	351, 460
LCK	1qpc	601, 364, 366	x	329, 438

Fig. 2.11 In each row representing the kinase, its pdb code has been shown along with the residues from the kinase that should bind in the ATP, with the phosphosites and arginines from the motif. In the lower part, for the tyrosine kinases, its pdb code has been shown along with the residues that should bind the substrate from the ATP, and the residues from the kinase pocket that should bind the motif.

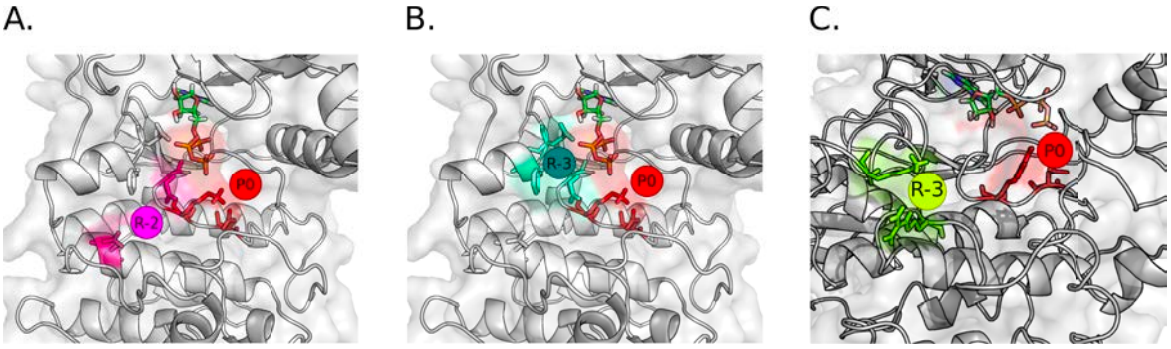


Fig. 2.12 A. A R-2 motif binds to the residues colored in pink, with the phoshphosite binding to the red ones including part of the ATP. B. R-3 motif binds within residues colored with green and the phosphosite with the ones in red including part of the ATP. C. In LCK and SRC the motif D/E -3 is recognised by the residues colored in green and the phosphosite is accepted by the ones colored in red (including part of ATP).

kinase	substrate id	phosp	aa phosp	3D motif	aa 3D	R/K-2 (max 5)	R/K-3 (max 6)
AKT1	ENSP00000361626	102	S	93	K	3	5
AKT1	ENSP00000312455	273	S	432	R	4	5
CAMK2A	ENSP00000035383	1439	S	1453	K	5	6
CAMK2A	ENSP00000428994	577	S	569	R	4	4
CHEK1	ENSP00000346694	159	T	137	K	5	6
CHEK1	ENSP00000216122	633	T	573	R	5	5
CHEK1	ENSP00000216122	633	T	581	K	5	4
PRKACA	ENSP00000420588	56	S	96	K	1	3
PRKACA	ENSP00000420588	56	S	97	K	4	3
ROCK1	ENSP00000361021	229	S	260	K	3	5
ROCK1	ENSP00000384136	573	T	263	K	3	5

Fig. 2.13 List of kinases and their substrates that has been prepared for docking for motif R-2 and R-3. Columns phosp and 3D motif show the number of the residue in substrate that is providing the motif. Since each of kinases have the available motifs R-2/R-3 both of the possibilities has been simulated independently. The number of maximum bindings for each motif is different, because motifs bind to different residues — for the motif R-2 the maximum amount of bindings is 5, and for R-3 it is 6. The rows containing the maximum amount of available bindings for both of the motifs are indicated with green highlight.

chose one of the lowest energy as the representative. The total number of simulations is 28 (including all the kinases), and 10 are considered successful - 2 for CAMK2A, 4 for CHECK1, 3 for SRC, and 1 for LCK.

Serine/Threonine kinases

Figure 2.13 presents outcomes of simulations for the kinases with a motif R/K-3 or R/K-2, which are AKT1, CAMK2A, CHEK1, PRKACA and ROCK1. Because the motif does not strictly identify whether the Arginine/Lysine has to be in the -2 or -3 position, for each kinase substrate pair I ran two individual docking simulations and their outcomes are presented in table in Figure 2.13. Motif preferences are exactly listed in Methods. Some of the kinases have stronger preferences than the others. Since Arginine and Lysine are both positive amino acids, I decided to accept both amino acids in potential 3D motifs. For two kinase-substrate pairs both of the simulations has been successful (CAMK2A and CHEK1), and for two more (CHECK1) the motif R/K-2 has been docked.

Tyrosine Kinases

The set of Tyrosine kinases (SRC and LCK) with the motif D/E-3 had 5 string constraints within the simulation. The scores for the motifs presented in Methods may seem lower than those of Serine kinases, but they were still the highest on the list. The lower score is caused mostly by the lower number of substrates available for the analysis. At the time of setting up

kinase	substrate	phosp	aa phosp	3D motif	aa 3D motif	D/E-3 (max 5)
SRC	ENSP00000381854	1173	Y	1198	D	4
SRC	ENSP00000381854	1206	Y	1240	E	4
SRC	ENSP00000225655	129	Y	83	E	4
SRC	ENSP00000327251	151	Y	161	E	3
SRC	ENSP00000362014	231	Y	221	E	3
SRC	ENSP00000426909	246	Y	260	D	4
SRC	ENSP00000426909	246	Y	262	E	5
SRC	ENSP00000457230	33	Y	23	E	4
SRC	ENSP00000457230	33	Y	24	E	4
SRC	ENSP00000346032	42	Y	315	E	4
SRC	ENSP00000248444	60	Y	86	D	3
SRC	ENSP00000344115	685	Y	721	D	4
SRC	ENSP00000365439	72	Y	93	D	5
SRC	ENSP00000365439	72	Y	95	E	4
SRC	ENSP00000365439	72	Y	99	E	5
LCK	ENSP00000046794	426	Y	449	D	5
LCK	ENSP00000315768	690	Y	680	E	3

Fig. 2.14 The results of docking simulations as D/E-3 motifs. Columns phosp and 3D motif show the number of the residue in substrate that is providing the motif.The number of available bindings is 5. The rows containing the maximum amount of available bindings for the motifs are indicated with the green highlight.

the experiment PDB database contained only one solved LCK and one SRC structure with a docked peptide, that could have been used to establish where the peptide should be binding inside the catalytic cleft. Four of the 17 kinase-substrate pairs showed a positive result of docking, fulfilling the 5 bindings threshold. Nine of all SRC targets had 4 bindings which can be considered as a satisfying indication for positive binding.

Three examples of successful docking are presented in Figure 2.15. Structural representations show how a kinase (colored in blue) interactions with the substrate (white). Kinase is presented as a surface, with ATP colored orange. Phosphosite from the substrate is presented as red sticks, and is in close distance to the ATP. Residue from the 3D motif is shown in yellow stick. Figure 2.15.A. shows CAMK2A (pdb: 2VZ6) interacting with ENSP0000035383 (structure prediction (Methods)), where the 3D motif is bound as R-2 motif. Figure 2.15.B shows 3D motif bound as R-3 in CHEK1 (pdb: 2YEX) interacting with ENSP00000216122 (structure prediction (Methods)). In Figure 2.15.C SRC kinase (pdb: 1Y57) interacts with ENSP00000365439 (pdb: 1J5K). Details of the interactions can be also found in Table in Figure 2.13. (A. and B.) and Table in Figure 2.14. (C.).

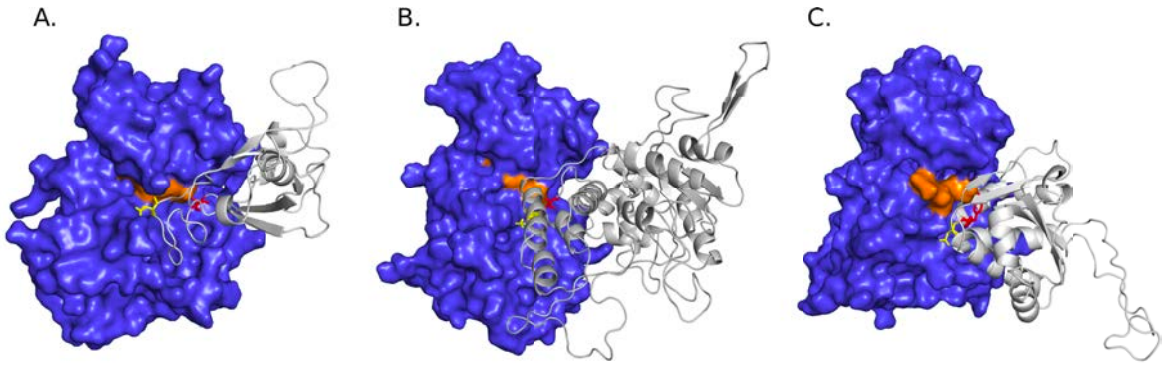


Fig. 2.15 Successfully docked examples of 3D motifs. Kinases represented in blue surface, with orange highlighted ATP. Substrates are white, with the phosphosites in red stick representation and the aa from the 3D motif in yellow sticks. A. CAMK2A with ENSP0000035383. B. CHEK1 with ENSP00000216122 C. SRC with ENSP00000365439.

2.3 Methods

2.3.1 Sequential motifs

In vitro

For each of the kinases a sequential motif has been defined using the R version of motif-x (Chou and Schwartz 2011a), including all the available sequential data. Since the kinases and motifs are different for the in vitro and in vivo sets, corresponding sequential motifs are presented in the results section. Sequence motifs for the in vitro set are presented in Table in Figure 2.16. The "motif-x score" is calculated by taking the sum of the negative log probabilities used to fix each position of the motif. As shown in Figure 1.5. I took the consensus motifs from the subsets obtained from motif-x results.

In vivo

For the *in vivo* data I grouped the kinases according to their preferences (Figure 2.17). In the first group I explored the motifs R/K -3 and R/K -2, however not each kinase had exact preference for both. For example CHEK1 has strong preference for Arginine in position -3, however Lysine in that position is also possible.

2.3.2 Structural motifs

The structural database was established around the sequences of human, canonical proteins. For each of the sequences, if available, the longest pdb structure was assigned as a model and then relaxed and minimised using ModPipe (Pieper et al. [222]). If the structure was not available, the homological prediction using ModPipe was performed. The percentage of pdb

	AKT1	Motif-x score
1	R.R..[ST]..S..	627.12
2	R.R..[ST].....	615.31
3	..RR.[ST].....	318.84
4	..R..[ST].....	307.65
	ERK1	
1	...P.[ST]PP...	623.34
2	...P.[ST]P....	615.31
3	N....[ST]P....	313.74
4	.....[ST]P....	307.65
	PKA	
1	..RR.[ST]...L.	922.96
2	P.RR.[ST].....	922.96
3	G.RR.[ST].....	922.96
4	..RR.[ST]..S..	622.21
5	..RR.[ST].....	615.31
6	..RK.[ST].....	615.31
7	..KR.[ST].D...	922.96

Fig. 2.16 Motif-x results for the *in vitro* kinases set.

motif	AKT1	CAMK2A	PRKACA	CHEK1	ROCK1
..R..S…..	627.12	307.65	922.96	307.65	922.96
...R.S…..	318.84	8.27	922.96	x	12.59
..K..S…..	217.1	x	615.3	9.7	16.65
...K.S…..	x	x	319.9	x	13.95
motif		LCK		SRC	
..D..Y…..		9.4		10.62	
..E..Y….		6.54		7.15	

Fig. 2.17 Motif-x results for the *in vivo* kinases set. The motif-x score is calculated by taking the sum of the negative log probabilities used to fix each position of the motif.

structures to models in the database is 40%. The database includes 12.162 proteins, where the predictions were filtered so that the model was considered acceptable if the template sequence identity was at least 30% and met at least one following criteria: GA341>0.7, e-value <0.0001 or average dope score <0. When more than one model was available for a protein, the model with the highest dope score was chosen as the representative.

2.3.3 Docking

The docking simulations have been performed with HADDOCK (Dominguez et al. [66]) standard protocol, which allows for the selection of residues likely to be the part of the interface. Simulations were performed based on the examples chosen from the *in vivo* data set.

2.4 Discussion

In vitro

There are a number of target peptides that have important determinants absent in linear sequence space but have a potential "replacement" residue found at an equivalent 3D distance. These determinants are observed at a higher frequency in the true targets when compared to random non phosphorylated residues. Most of the target sites are missing at least one of the important determinants and a large number of these have a potential replacement at an equivalent 3D distance. This suggests that there may be a significant number of sites that are recognized by kinases 3D epitopes. However, the occurrence of 3D determinants is not higher in the targets of a kinase than random expectation. The spatial distribution of the residues might influence their detection — distance equivalent to R-5 may be too wide and generate too many false positives while distance for P+1 (3 Å) may be too narrow to find other amino acids within it.

In vivo

The analysis suggests that while kinase target sites display a clear enrichment for specific linear sequence motifs, the majority of the target sites miss one or more of the important determinants. These target sites may be recognized by the kinases using a residue of an equivalent 3D distance, forming a 3D motif. However, the occurrence of 3D motifs is not higher than expected from sampling random serines or threonines. Predicted in vivo cases of

3D motifs have residues that tend to be more conserved than random and show correlated changes with the changes in the respective kinase activity across a panel of conditions. 3D recognition of kinase target sites may be a significant factor for kinase target specificity. The predicted cases of in vivo 3D motifs has been selected for docking studies to analyse the structural properties of the predicted interactions. A number of docking simulations show that such associations are possible, however its existence should be validated in the wet lab.

Keeping in mind that the docking simulation is not a wet lab experiment, it is still possible that actual binding can be executed in the wet lab settings. Out of 11 tested 3D motifs, two has shown double possibilities of binding (CAMK2A with ENSP00000035383 and CHECK1 with ENSP00000346694). Two more pairs (for CHECK1) showed a possible R-2 motif binding, and three more pairs have 4 out 5 possible bindings — this can be considered as a positive result. Five pairs from the R-3 have 5 out of 6 bindings, which also could be further evaluated.

The uncertainty of the simulation can be explained on a few preparation levels. Firstly it is the choice of the structures representing both the kinase and the substrate and the choice of its binding amino acids. There are only few activated kinase structures known, and I defined the binding residues through homological comparison with other kinase models. For best simulations in HADDOCK, the more of the binding surface is defined, the more confident the simulation outcome is. While setting up the simulation I only had a few amino acids on each side to define the connection, which does not take into account all the possible distal docking mechanisms. Secondly it is hard to validate how well did the kinase and the substrate connected. Choosing the model of the lowest available energy to represent the complex, and checking how many of initially suggested bindings the simulation managed to fulfill was the arbitrary criteria that I chose when selecting the best simulations. Those requirements are reasonable enough to validate the further examination in the wet lab experiment. I cannot take into the account of the simulation the conformational changes and all the loop conformations that might exist in the lifetime of the targets or the kinases. The value of the simulation itself is the fact that those two proteins are physically able to allow the motif to fit inside the kinase cleft. This is not a confirmation that such 3D motif exists, but it is a good premise worth further testing.

Chapter 3

Conserved phosphorylation hotspots in eukaryotic protein domain families

3.1 Introduction

Mass spectrometry and biochemical enrichment methods allow for the study of PTM regulation at very large scale (Choudhary and Mann [40]) and such approaches have been extensively applied to study protein phosphorylation (Mann et al. [179]). In the most comprehensive single study to date estimated order of 75% of the detected proteome was found to be phosphorylated (Sharma et al. [256]) and approximately 160,000 non-redundant human phosphosites are listed in public repositories (Hornbeck et al. [106]). Although the regulation of protein functions by phosphorylation has been under study for over 60 years (Fischer and Krebs [81]) the recent wealth of knowledge regarding protein phosphorylation generated by mass-spectrometry remains mostly uncharacterized. Only around 5% of human phosphosites have an annotated regulatory role (Hornbeck et al. [106]). Evolutionary studies have suggested that some degree of protein phosphorylation, and other PTM sites, may have little to no biological function (Landry et al. [159], Beltrao et al. [15]). Given that some phosphosites may have no relevance to fitness, devising ways to rank sites according to functional importance is a crucial research question. Functionally important phosphosites have been shown to be more likely conserved across species and across protein domains of the same type (Studer et al. [275], Beltrao et al. [15]). For protein domains, conserved phosphorylation within a specific region, termed phosphorylation hotspot, tends to identify regions with regulatory potential (Beltrao et al. [15]). Since each domain family is represented by multiple copies within each species, these domain centric analyses have the advantage of increased statistical power when compared to the study of conservation of orthologous genes. Domain centric

analyses have also been used with success to study recurrence of mutations in cancer samples (Miller et al. [194]).

3.2 Results

3.2.1 Identification of eukaryotic phosphorylation hotspot domain regions

In order to study the conservation of protein phosphorylation within protein domain families I collected protein phosphosite data from publicly available sources for a total of 40 eukaryotic species, including 11 animals, 19 fungi, 7 plants and 3 apicomplexa species (Figure 3.1A and Methods). A total of 537,321 phosphosites were compiled and mapped to reference proteomes and protein domain regions which were identified using the Pfam domain models (Finn et al. [80]) across all species (Methods). Of all phosphosites, 83,359 phosphosites occur within Pfam domain regions (Figure 3.1A). As most phosphosites tend to occur in disordered regions (Iakoucheva et al. [120]) it is not unexpected that the majority of sites are not found within protein domains. The ranked list of ten most commonly modified domains is shown in Table in Figure 3.2. In line with previous findings, the most commonly regulated domains included many involved in cell signaling (e.g. protein kinase, Ras), chaperone function (e.g. HSP70, TCP, HSP90), and cytoskeleton (e.g. Actin, Myosin).

In order to statistically identify domain regions that are regulated by phosphorylation above random expectation I selected 344 domain families that are represented by at least 10 different instances and contained a total of 50 or more phosphosites. For these domain families, the protein sequences containing phosphosites were aligned and an enrichment score was calculated using a rolling window approach, with a fixed length of 5 positions, to identify regions with an above average degree of phosphorylation as illustrated in Figure 3.1B. The random expectation was calculated by permutation testing where phosphosites were randomly re-assigned within each protein sequence to equivalent phospho-acceptor residues (Methods). A rolling window approach was used to take into account alignment uncertainty and errors in assignment of the phosphosite position within the phosphopeptide as identified in the mass spectrometry studies. For each position within the domain alignments a p-value was calculated and after Bonferroni multiple testing correction a corrected cut-off p-value<0.01 (uncorrected p-value 6.70x10-8) was used globally to identify domain regulatory hotspots. A cut-off of an average of 2 phosphosites per position was also used to avoid significant positions with a low effect size difference. Contiguous positions were merged

to identify domain regions of interest that were defined as phosphorylation hotspot regions (Methods). Using this procedure a total of 1999 positions corresponding to 241 domain regions were identified as hotspots within 162 Pfam domain families. Full list of hotspots is presented in Appendix A., along with plots and structures for remaining, analysed domains.

3.2.2 Benchmarking results

Under the assumption that strong conservation of protein phosphorylation within a region of a domain is predictive of functional relevance I expected to find an enrichment of phosphosites with known functions at regions defined as regulatory by the hotspot identification approach. In order to analyse this hypothesis I compared the functional predictions with human phosphosites that are known to have a regulatory role obtained from the PhosphositePlus database. For each human phosphosite within the protein domains analysed I assigned the hotspot p-value. I considered only Pfam domains having more than 1 known human regulatory phosphosites and I analyzed separately serine/threonine (S/T) sites from tyrosines (Y). I was able to analyse a set of 983 S/T and 317 Y phosphosites with known regulatory functions in human, as defined in PhosphositePlus, out of a total of 8270 S/T and 1395 Y human phosphorylated positions within the same domains.

The capacity to discriminate the human phosphosites with known regulatory roles from other human phosphosites was tested using the receiver operating characteristic (ROC) curve (Figure 3.1C) and summarized as the area under the ROC (AUC) curve. The regulatory hotspot p-value was a strong predictor of know regulatory phosphosites (AUC=0.76 for S/T and 0.64 for Y). In line with this, the defined hotspot regions show significant enrichment over random for human phosphosites of known function (Figure 3.1.D). Overall, these results show that the regulatory hotspots identified here are enriched in previously known regulatory phosphosites.

To better test the capacity of hotspot method to identify the regulatory regions I used another set of known, regulatory human phosphosites. I mapped these human phosphosites with known functions to the full domain alignments (removing the regions that contained more than 30% of gaps) and assigned a regulatory p-value to each alignment position based on my calculation. Then, I marked a domain alignment position as a regulatory region if it mapped to at least 1 human regulatory phosphosite. Some positions were marked as regulatory by more than 1 human regulatory phosphosite, with the highest value for a kinase position — 85 human regulatory phosphosites, and 556 phosphosites over all species.

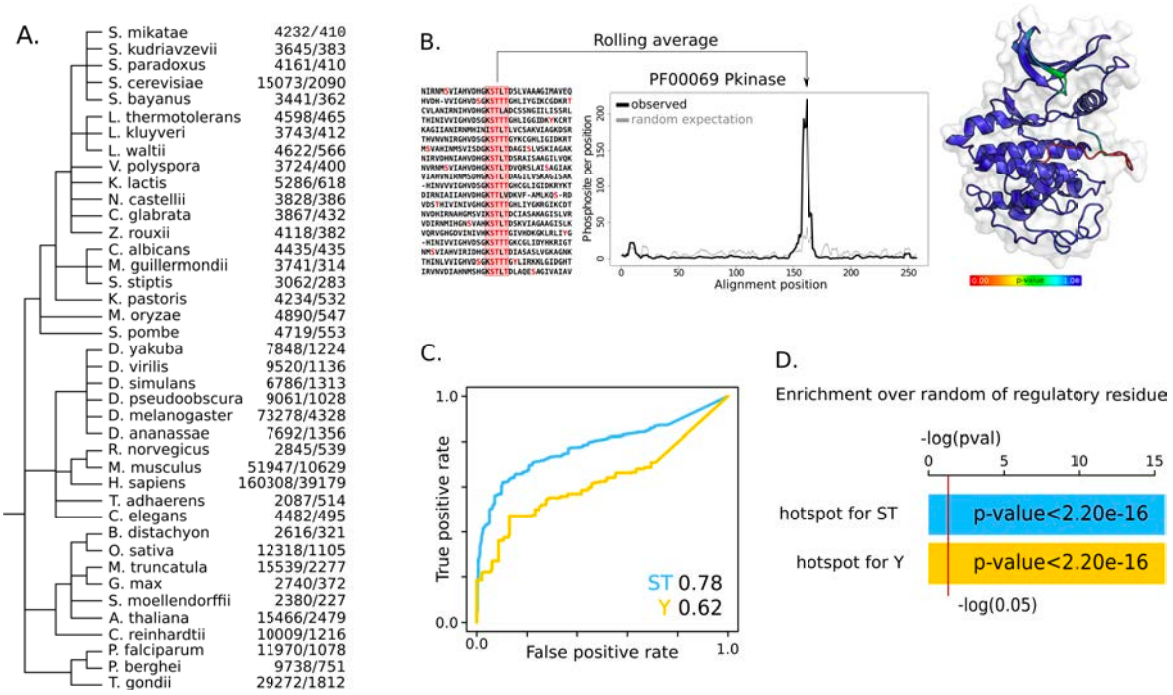


Fig. 3.1 Prediction of phosphorylation hotspots regions for eukaryotic domain families. A) Phylogenetic tree of the species from which phosphorylation data has been obtained. The numbers in the left column correspond to the phosphosites per species obtained and the the right column the phosphosites found within Pfam domains. B) Hotspot regions are defined as those having higher than randomly expected number of phosphorylation. A rolling window is used to count the observed average number of phosphosites in the alignment (black line) and a background expectation is calculated from random sampling (gray line and gray band for standard deviation). A p-value is calculated for the enrichment of phosphorylation at each position and projected onto structural models C) The capacity to discriminate between phosphosites of known function from other phosphosites was tested using a ROC curve. Comparison of the discrimination power of the hotspot p-value (blue line for ST and purple line for Y). D) Enrichment over random of human phosphosites with known functions for residues predicted as a hotspot region when compared with the rest of the domain (blue for ST and purple for Y - p-values for fisher test).

PFAM id	Domain name	Domain count	Phosphosites in domain
PF00069.23	Pkinase	9083	2671
PF00443.27	UCH	911	764
PF00076.20	RRM_1	7002	764
PF07714.15	Pkinase_Tyr	2362	716
PF00012.18	HSP70	592	598
PF00071.20	Ras	1806	404
PF15440.4	THRAP3_BCLAF1	10	397
PF00038.19	Filament	256	395
PF00118.22	Cpn60_TCP1	535	373
PF07690.14	MFS_1	2415	361
PF00022.17	Actin	615	355
PF00183.16	HSP90	152	350
PF00630.17	Filamin	1051	330
PF00063.19	Myosin_head	502	320
PF00270.27	DEAD	2101	312
PF08065.10	K167R	49	290
PF07679.14	I-set	6145	286
PF00041.19	fn3	4589	283
PF00125.22	Histone	953	276
PF00435.19	Spectrin	2204	270
PF00400.30	WD40	13824	251
PF00428.17	Ribosomal_60s	226	248
PF01576.17	Myosin_tail_1	155	246
PF00001.19	7tm_1	2063	244
PF00169.27	PH	1509	243

Fig. 3.2 List of 25 most phosphorylated domains. The ranking is not normalized by the total number of domains present in the genomes nor the total number of corresponding amino-acids. This ranking allows us to select the list of domains with a large number of experimentally identified phosphosites for analysis.

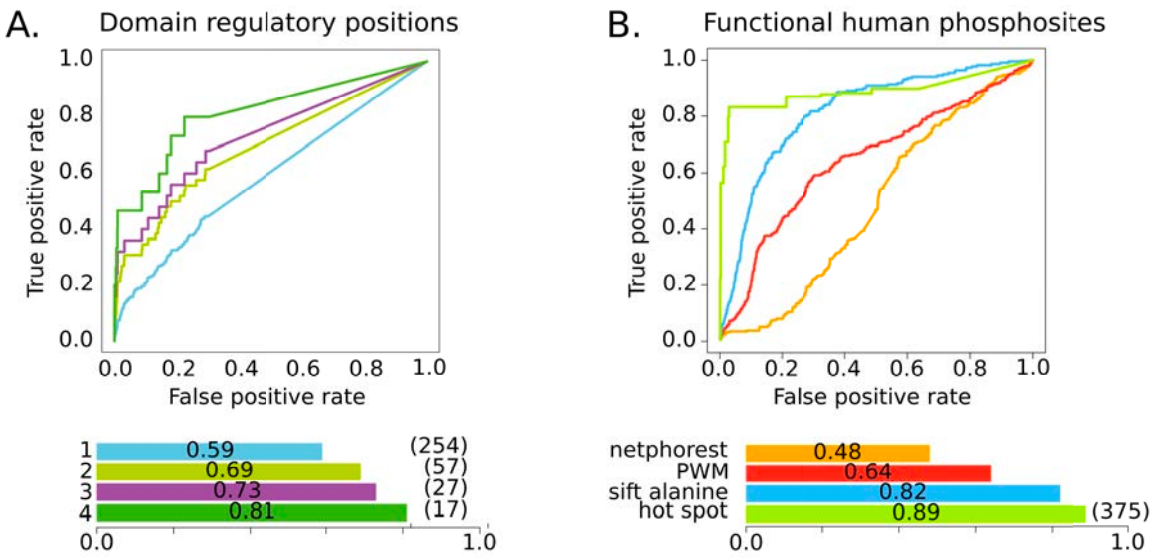


Fig. 3.3 A) ROC curve for domain regulatory positions, for 82 domains each of them having at least 1 position with a known regulatory phosphosite with a total of 17408 positions in all domains. Domain positions were considered to be true regulatory regions if they matched to 1 to 4 of human phosphosites previously known to have regulatory functions. The total amount of true positive domain positions after filtering is presented in the barplot in parenthesis. The increasing filter lowers the amount of available phosphosites. B) Human phosphosites, from 55 domains that each of them have at least 1 human phosphosite mapped to a regulatory position. Common set of 4774 human phosphosites, their functionality predicted by 4 methods (375 true-regulatory sites).

Multiple regulatory sites mapped to one domain position indicate that the position is more likely to be important in regulation not for a single protein but also different domains of the same family across different genes. I decided to evaluate whether finding positions supported by more than only one regulatory human site would improve the performance of the method. To do that I created four filters, starting from score of 1 regulatory position as true positive to 4 (all smaller were treated as negatives). The ROC curves, which show the true positive rate as function of false-positive rate are presented in Figure 3.4.A. The results show that my approach can correctly identify domain positions known to play regulatory roles, in particular for positions that are supported by multiple regulatory sites across different genes. A maximum of AUC of 0.81 was observed for positions that are supported by 4 or more regulatory sites.

The domain positions will be differentially phosphorylated, with some very rarely or never regulated. A more practical and relevant application of the regulatory hotspot score is to discriminate which of the phosphosites in a protein is more likely to be important in a given species of interest. Therefore I tested if the regulatory hotspot predictions could discriminate

between human phosphosites with known function from other human phosphosites with unknown function. For each human phosphosite within the PFAM domains analysed I assigned a regulatory p-value based on my calculation. Considering only PFAM domains having more than 1 known human regulatory phosphosite I was able to analyse 55 domains, which when creating a common set provides 375 true positive regulatory phosphosites out of a total of 4774 human phosphorylated STY positions within the same domains. For all these phosphosites I was able to obtain other scores that could be indicative of functional relevance including netphorest (Horn et al. [103]), Position Weight Matrix score (using the maximum as the metric) and SIFT score for Alanine (Sim et al. [263]).

The SIFT algorithm predicts whether a substitution of amino acid affects protein function. The score is based on conservation of residues in sequence alignments of closely related proteins, obtained from PSI-BLAST. The Alanine score is predicting whether the mutation of the amino-acid to Alanine is deleterious. The NetPhorest predictions are built from an atlas of linear motifs recognized by specific protein kinases and domains that bind to phosphorylated residues. The comparison allows to quantify the reliability of hotspot predictions as well as to determine whether my method can outperform existing ones (Figure 3.3.B). Using the regulatory hotspot p-value as a predictor I could discriminate the known regulatory phosphosites from the others (area under the ROC curve of 0.89). I compared this with SIFT predictions, using the Alanine mutation score (also shown in Figure 3.3.B) (AUC=0.82), the Netphorest prediction score (AUC=0.48), and the kinase position weight matrices (AUC was 0.64). These different scores are not redundant and a postdoc in the group (David Ochoa) has shown that they can be combined to improve the prediction of regulatory sites. However, the focus of my project is to identify the domain regions that are likely to be important to regulate protein function.

3.2.3 Mapping of regulatory hotspots to representative structural models

The phosphorylation hotspots are of functional and structural interest as these are very likely to have regulatory potential that should often be a general property of the domain family. To study these regions in the context of protein structures I have collected structures available for Pfam domains in PDB (Berman et al. [19]). For each domain family I discarded structures with short sequences, performed structural clustering and selected a representative from the most populated structural cluster (Methods). The protein sequences from the selected structural models were added to the alignments and a total of 116 hotspots regions were

mapped to a 3D model for 85 domains. I present the structural representation of all these hotspot regions and enrichment plots in Appendix A.

To gain a better understanding of how these regions may control domain functions, I studied in more detail some regions that overlap with human phosphosites of known function (Figure 2).

The protein kinase activation loop is the prototypical example of a regulatory hotspot (Figure 3.4, top). Over 50% of phosphosites and 74% (128 out 174) of known human regulatory sites (red dots) found within protein kinases are in this loop that follows the $\beta 9$ sheet near the active site. Phosphorylation of this loop is typically required to achieve full activation of kinases by positioning the activation segment in order to allow for substrate recognition (Johnson et al. [130]). Another well characterized example is the regulation of the pyruvate dehydrogenase complex (PDC) which is primarily composed of multiple copies of pyruvate dehydrogenase (E1), dihydrolipoamide acetyltransferase (E2) and dihydrolipamide dehydrogenase (E3). PDC activity is regulated by phosphorylation of E1 in positions that overlap with our identified phosphorylation hotspot (Figure 3.4). The phospho-regulation of this domain is well characterized with 3 described regulatory phosphosites (Korotchkina and Patel [149], Patel et al. [218]). Two of these positions fall within what is termed the phosphorylation loop A (Ph-loop A) region, which overlaps directly with the hotspot region. Phosphorylation of this loop region is known to induce a conformational change in the loop that causes enzyme inhibition (Kato et al. [138]). I expect that the identified hotspots from other domain families will be of regulatory importance in analogy to the activation loop of kinases and the phosphorylation loops of pyruvate dehydrogenase.

A clear phosphorylation hotspot was found for the Ras domain family (Figure 3.4). This small GTPase superfamily is known to change in conformation depending on the GTP versus GDP bound state with two loop regions, called switch 1 and switch 2, being particularly sensitive to the nucleotide binding.

The major Ras phosphorylation hotspot occurs just after the switch 2 region at the start of $\alpha 2$. This region is also known to often form contacts with effector molecules (Mott and Owen [201]) implicating the phosphorylation of this region in the regulation of Ras-effector interactions. Supporting this hypothesis, the phosphorylation of human KRas in this region at Y64 regulates the interaction between KRas and AGO2 (Shankar et al. [255]). Similarly, phosphorylation of the same region in Rab7 (S72) and in RAC1 (Y64) is associated with changes in interaction partners (Shinde and Maddika [260], Chang et al. [36]). This indicates that this a commonly used regulatory feature of Ras domains.

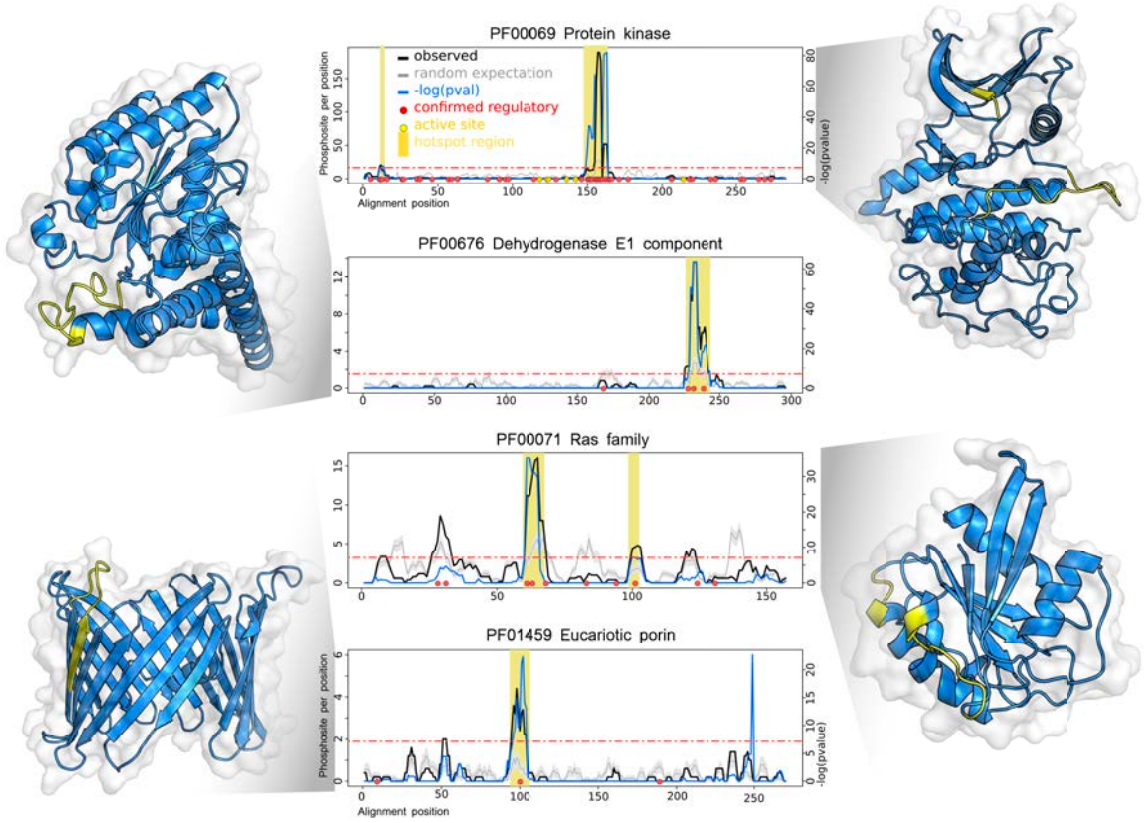


Fig. 3.4 Examples of phosphorylation hotspots containing human phosphosites with known regulatory roles. For 4 protein domain families it shows the enrichment over random of protein phosphorylation along the domain sequence. The average number of phosphosites observed per rolling window is plotted in a solid black line (observed). The background level of expected phosphorylation calculated from random sampling is shown in gray line, with standard deviations as gray band. The blue line represents the negative logarithm of p-value at each position (right y axis). A horizontal red line indicates a cut-off of the Bonferroni corrected p-value of 0.01. Positions with a -log(p-value) above this cut-off and average phosphosites per window higher than 2 are considered putative regulatory regions and highlighted under a vertical yellow bar. Red circles indicate human phosphosite positions with known regulatory function. In the structural representations the predicted hotspot regions are highlighted in yellow.

A different mode of phospho-regulation has been observed in the voltage-dependent anion channel (VDAC). This 19 beta sheet beta barrel domain is a class of porin ion channel spanning the outer mitochondrial membrane. The major phosphorylation hotspot in the VDAC domain is predicted for a loop region between $\beta 6$ and $\beta 7$ (Figure 3.4). The regulation of the human VDAC1 at position S104 within this region controls VDAC1 protein levels by inhibiting proteasome mediated degradation (Yuan et al. [316]). The orientation of VDAC domains in the membrane is contentious with conflicting reports suggesting that the C-terminal may point towards the cytoplasm (Tomasello et al. [282]), the mitochondrial intermembrane space (IMS) or that is may occur in both orientations (McDonald et al. [188]). The phosphorylation hotspot between $\beta 6$ and $\beta 7$ is on the opposite side of C-terminal region, suggesting that this loop most likely or most often will face the cytoplasm placing the C-terminal towards the IMS.

These examples further illustrate how hotspot analysis recovers well known regulatory regions as well as shows some ways in which domain function is regulated by protein phosphorylation (e.g. changing activity, conformation, interactions, degradation rates). Following examples were aimed to test if some of the regulatory mechanisms can be generalized to other domain families.

3.2.4 Phosphorylation hotspots are enriched for positions at protein interfaces and near catalytic residues

Regulation of interactions and catalytic activities may be general mechanisms of domain regulation. To study this across all domains I used annotations for interface residues from the 3DID database (Mosca et al. [199]) and for catalytic residues based on Uniprot annotations (UniProt Consortium [291]) (Methods). I also analysed surface accessibility, defined as >20 relative surface accessibility (RSA), and predicted disorder (DISOPRED Ward et al. [298]) (Methods). As expected from prior studies of protein phosphorylation (Johnson et al. [129]) the hotspot positions are more likely to be surface exposed (Figure 3.5.A, p-value=1.66x10-8) and within disordered elements (Figure 3.5.A, p-value<2.20x10-16) when compared to other residues. The next thing I wanted to examine were distances between hotspot positions and catalytic residues. Compared to other residues within enzymes, hotspot positions are 3 times more likely to be catalytic residues (Figure 3.5.A, catalytic residues, p-value=0.03); 3.4 times more likely to be within 5 amino-acid distance (Figure 3.5.A p-value=1x10-8) and 5 times more likely to be within 5Å distance (Figure 3.5.A, <5Å p-value=1.5x10-8) to catalytic

residues. For enzyme domains 3.3% of hotspot residues are within 5Å distance of catalytic residues compared with 0.97% for other residues.

For each domain position I identified interface contacts found in 3D structures with any other protein domain based on 3DID, excluding intra-domain contacts. For the interface residue enrichment test we considered only surface accessible residues (>20% RSA) to avoid an enrichment simply due to accessibility of both interface and hotspot positions. Controlling for surface accessibility hotspots are 1.8 more likely to be interface positions (Figure 3.5.A, p-value<2.4x10-9). 39% of accessible hotspot residues are interface positions compared with 26% for other accessible residues. Some hotspot regions overlap with interaction regions that can make contacts with a large number of different types of domains as determined by crystal structures. For example, the hotspot region in Ras described above contacts 42 other domain types, some of which are illustrated in Figure 3.5.B. There are 13 domain families with hotspot regions contacting with more than 3 other domain families including the SH2 domain and RNA recognition domain families shown in Figure 3.5.B. This suggests that protein phosphorylation of such regions may be important for switching the interaction partners of these domain families.

These results indicate that regulation of protein interactions and catalytic activities may be a recurrent feature of domain regions regulated by protein phosphorylation. Of the 116 hotspot regions mapped to a structural models, 97 overlap with interface residues and of 32 hotspot regions with putative catalytic residues, 23 are within 15Å and 5 are within 5Å to a catalytic residue. The list of hotspot regions which is presented in Appendix 1 contains information regarding interface positions and proximity to catalytic residues.

3.2.5 Phosphorylation hotspot regions near catalytic residues

Hotspot regions in enzyme domains are often found at or near catalytic residues and could, in these cases, play a role in regulating catalytic activities. From the 23 hotspots found within 15Å of a catalytic residue I illustrate 4 examples in more detail (Figure 3.6). Protein phosphorylation is typically catalyzed by protein kinases, however I found examples of hotspot regions explained as reaction intermediates or auto-phosphorylation not catalyzed by kinase. For example, the hotspot region of Alkaline Phosphatase (ALP) overlaps directly a catalytic residue (Figure 3.6). The hydrolysis and transphosphorylation of monoesters reaction takes place in the active pocket of the enzyme which contains a catalytic serine that forms a covalent serine-phosphate intermediate. This catalytic serine, located at the N-end of α5, is found phosphorylated in different species explaining the identified hotspot. This

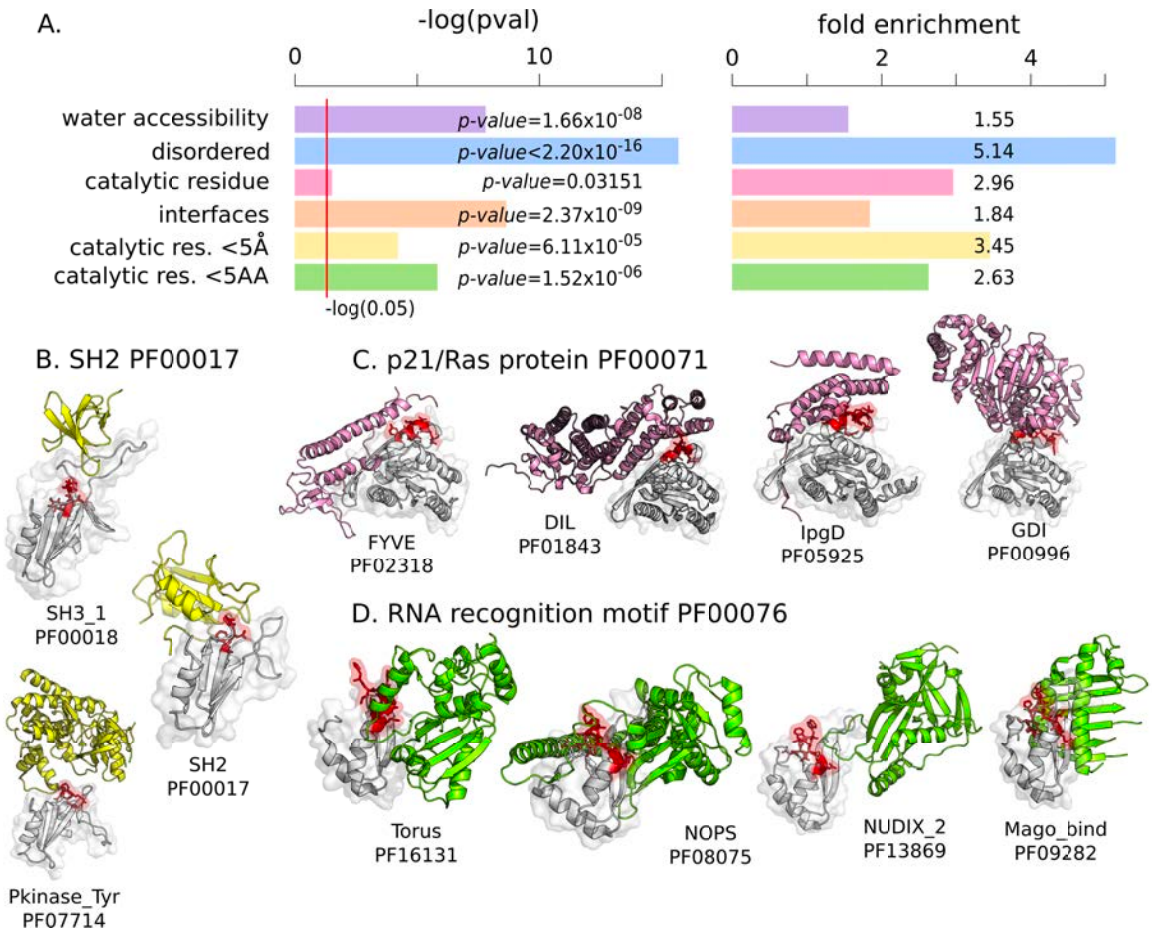


Fig. 3.5 Structural features of phosphorylation hotspots. A) Enrichment over random of structural features comparing hotspots with other residues within the same domains. The tested features include water accessibility — residues with>20% RSA; protein disorder as predicted by DISOPRED; catalytic residues; residues within 5 amino-acid distance to a catalytic residue; residues within 5 angstroms of a catalytic residue; residues at an interface based on 3DID. For each feature we report the -log(p-value). B) Examples of hotspot regions at interfaces where the hotspot region (red) from a domain (grey) has been observed contacting many other types of domains (other colours) in empirical structures.

hotspot is therefore the result of a reaction intermediate and not regulated by protein kinases. A hotspot for the phosphoglucomutase/phosphomannomutase Pfam domain (PF02878) is identical to this in that a catalytic serine is often found phosphorylated and known to be a reaction intermediate not catalyzed by kinases (Figure 3.8).

The Nucleoside-diphosphate kinases (NDK) catalyze the exchange of terminal phosphate between different nucleoside diphosphates (NDP) and nucleoside triphosphates (NTP). A NTP serves as a donor and the reaction proceeds via a phospho-histidine intermediate in the NDK active site. The main hotspot in this domain occurs just next to this active site histidine (Figure 3.6). The phospho-histidine is not detected as phosphorylated in the proteomics data, most likely due to phospho-histidine being very labile (Sickmann and Meyer [262]) and not usually searched for during the mass spectrometry data processing steps. Phosphorylation of these nearby serines has been suggested to be the result of a transfer of phosphate between the histidine and nearby serines which may be important for the enzyme activity (Dar and Chakraborti [56], Mocan et al. [195]). In addition I found a second hotspot in the loop between α7 and α8 with an unknown function. Given that this loop partially covers the catalytic centre the phosphorylation of this loop likely regulates substrate accessibility.

The next two domain families I studied are examples of conserved phosphorylation regions distant in sequence but close in 3D space to catalytic residues (within 15Å). The IMP dehydrogenase (IMPDH) catalyzes the oxidation of IMP to XMP with the concomitant reduction of NAD+. In human cells Akt has been shown to interact with IMPDH and phosphorylate the protein *in vitro* (Ingley and Hemmings [126]) but the position or functional role of IMPDH phosphorylation has not been established. In structures of this domain a serine residue can be found in this hotspot (Figure 3.7.A) pointing towards the substrate binding pocket and its phosphorylation may sterically impact on substrate binding. A loop next to this hotspot changes in conformation during the catalytic cycle (Hedstrom [98]) (Figure 3.7.A, Open to Closed) so the phosphorylation of the hotspot could also have an impact on these dynamics.

Similarly to IMPDH, the hotspot region of transaldolase (TAL) is in a position that could influence the access to the catalytic centre. TAL is an enzyme of the nonoxidative part of the pentose phosphate pathway (PPP). The active site, located in the center of the barrel is formed of a lysine, that holds the sugar in place and a glutamate and aspartate that act as proton donors and acceptors. There is evidence that TAL activity can be regulated by phosphorylation (Lachaise et al. [157]) but the position or mechanism of this regulation has not been determined. The hotspot identified for TAL is very likely to alter the accessibility of the substrate to the active site. In structures of this domain a serine residue within this hotpot

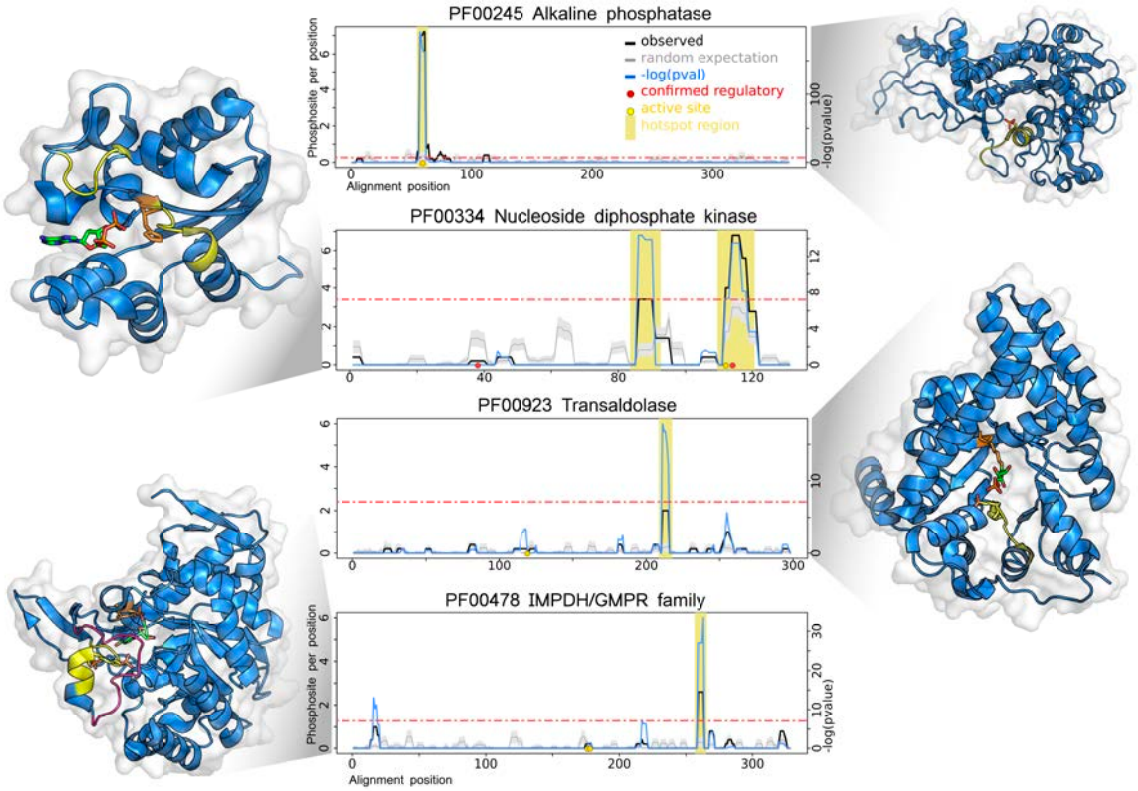


Fig. 3.6 Examples of putative regulatory hotspots at or near catalytic residues. The average number of phosphosites observed per rolling window is plotted in a solid black line (observed). The background level of randomly expected phosphorylation is shown in gray line, with standard deviations as gray band. The blue line represents the negative logarithm of p-value at each position (right y axis). A horizontal red line indicates a cut-off of the Bonferroni corrected p-value of 0.05. Positions with a -log(p-value) above this cut-off and average phosphosites per window higher than 2 are considered putative regulatory regions and highlighted under the yellow bar. Red circles indicate human phosphosite positions with known regulatory function and yellow circles represent catalytic residue positions. In the structural representations the predicted hotspot regions are highlighted in yellow. The catalytic residues have been represented as orange sticks and in red stick representations are substrates or products.

can be found just at the entrance to the substrate pocket (Figure 3.7.B) and phosphorylation of this residue may control access to the cavity.

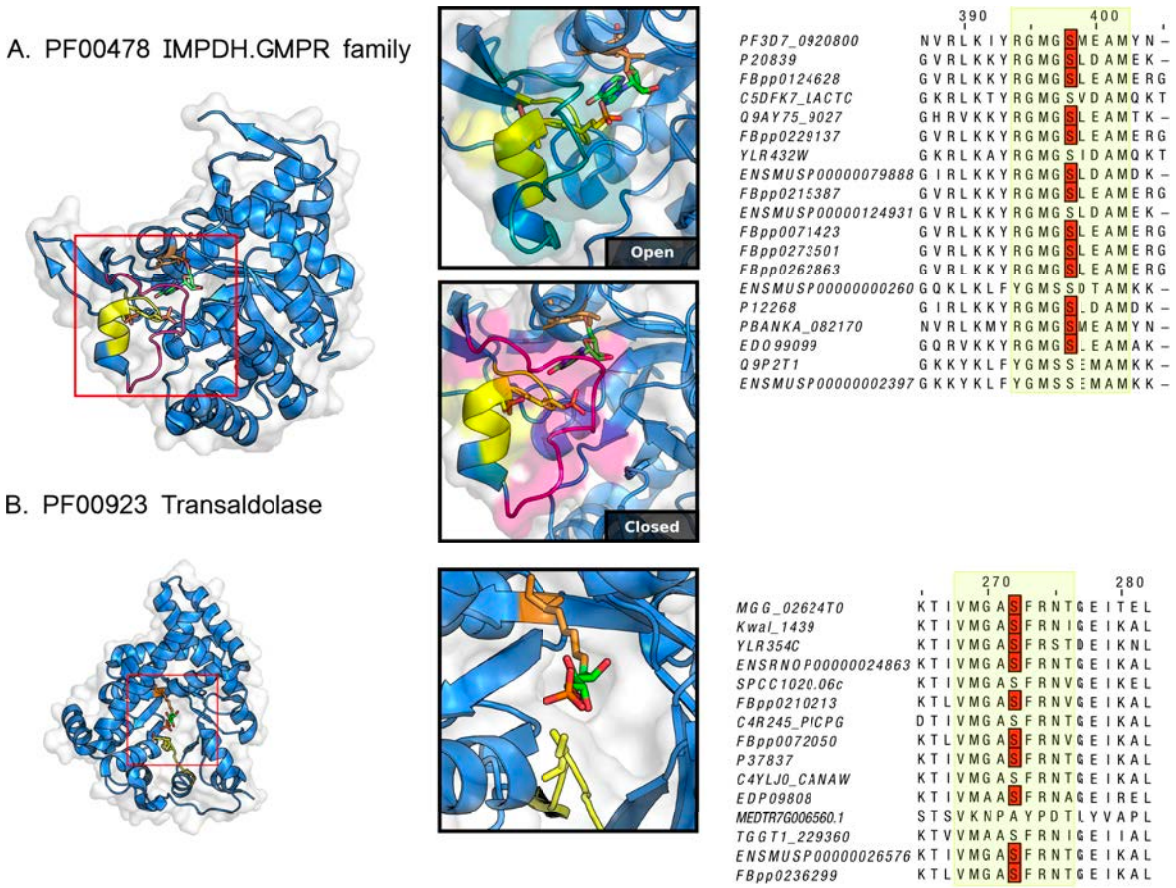


Fig. 3.7 Hotspot regions of near catalytic residues that are distal in protein sequence. A) The IMPDH hotspot region is represented in yellow segment. In the insets, the loop near the hotspot region is shown changing from an open conformation (blue volumes) to closed conformation (magenta volumes). A serine residue within the hotspot region (yellow sticks) points to substrate binding pocket and is often found phosphorylated across species (see alignment). B) The transaldolase hotspot region is show in yellow. In the structural inset a serine within the hotspot region (yellow sticks) is found just at the entrance of the substrate cavity. The identified phosphorylation sites contributing to the identification of the hotspot region are shown in the alignments in red.

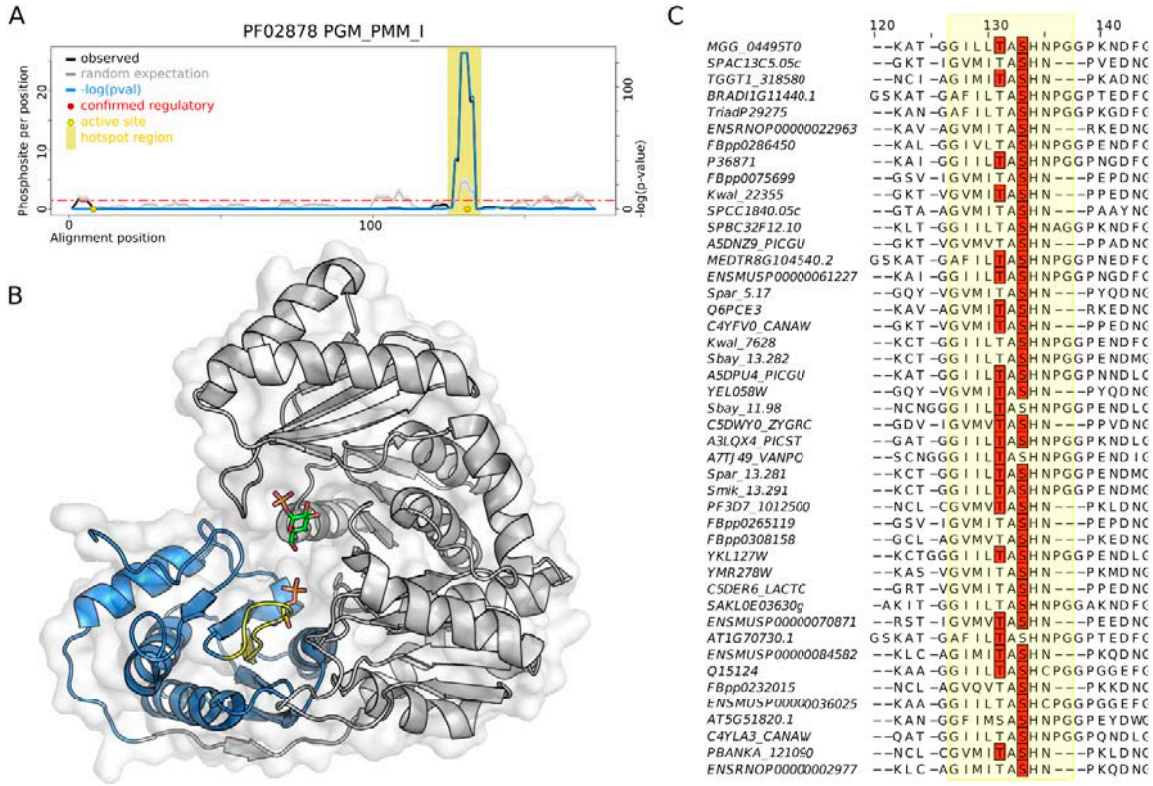


Fig. 3.8 Hotspot regions at a catalytic serine of the phosphoglucomutase/ phosphomannomutase (PGM/PMM) domain (PF02878). A) The hotspot region is represented in yellow segment and the catalytic serine residue within the hotspot region (yellow sticks) points is shown in the structure (B). This phospho-serine catalytic intermediary is capture in the MS phosphoproteomics data as shown in the alignment (C). There are often phosphorylated threonines next to the serine residue which could be miss-assigned phosphosites, or potentially these could be targeted by kinases.

3.3 Functional relevance of phosphorylation within the C-terminal hotspot region of the Ribosomal S11 domain

Thanks to the collaboration with Michaela Oplova and Cris Vieitez I was able to follow on the functional studies on ribosomal S11 domain family. The predicted regulatory hotspot, found within the 40S ribosomal protein uS11 has been kindly examined by Michaela and Cris and the results of this work are presented in (Strumillo et al. [274]). I did not perform any of the wet lab work, and below I am describing the outcomes that are related to the hotspot described by me in the C-terminal tail of ribosomal protein family (Figure 3.9.A). To test the functional relevance of this phosphorylation hotspot Cris Vieitez selected 2 phosphosites in the yeast protein Rps14A (uS11) that have been identified near this region (T119 and S123), marked in Figure 3.9.A and 3.9.B. The two sites are known to be phosphorylated in different species including human (Figure 3.9.E). She constructed two strains with alanine mutations at each of these positions inserted in the genome (Methods). Then these strains have been tested for growth defects under different set of stress conditions: 6-azauracil (6AU), cycloheximide (CHX) and cold shock. There was no effect under 6AU, a weak growth defect for S123A under CHX and a robust cold shock phenotype for the T119A mutant (Figure 3.9.D). Interestingly, RPS14A has a paralog - RPS14B that was not deleted or mutated for these studies, meaning that rps14a T119A mutant might act in a dominant negative manner. Based on the initial growth defect Cris tested but saw no phenotype in the early steps of ribosome assembly using a uS5-GFP reporter assay. A previous report has indicated a role for Rps14A C-terminal region in activating the ATPase Fap7. Failure to activate Fap7 impaired downstream cytoplasmic 20S pre-rRNA processing (Peña et al. [221]) which we assayed for in the T119A mutant. Michaela Oplova tested the capacity of the mutant to convert 20S pre-rRNA to 18S rRNA by determining the subcellular localization of 20S pre-rRNA by fluorescence *in situ* hybridization (Methods). Michaela observed a cold shock dependent 20S pre-rRNA accumulation in the cytoplasm, indicative of a processing defect (Figure 3.9.F). This tail region was shown to make contacts with the ATPase domain of Fap7 (Figure 3.9.C) and the C-terminal region of uS11 was demonstrated to activate the ATPase Fap7, a critical step to release and deposit uS11 and its interacting partner eS26 into its rRNA binding site (Peña et al. [221]). It seems likely that the phospho-mutant rps14a T119A may not be able to activate Fap7.

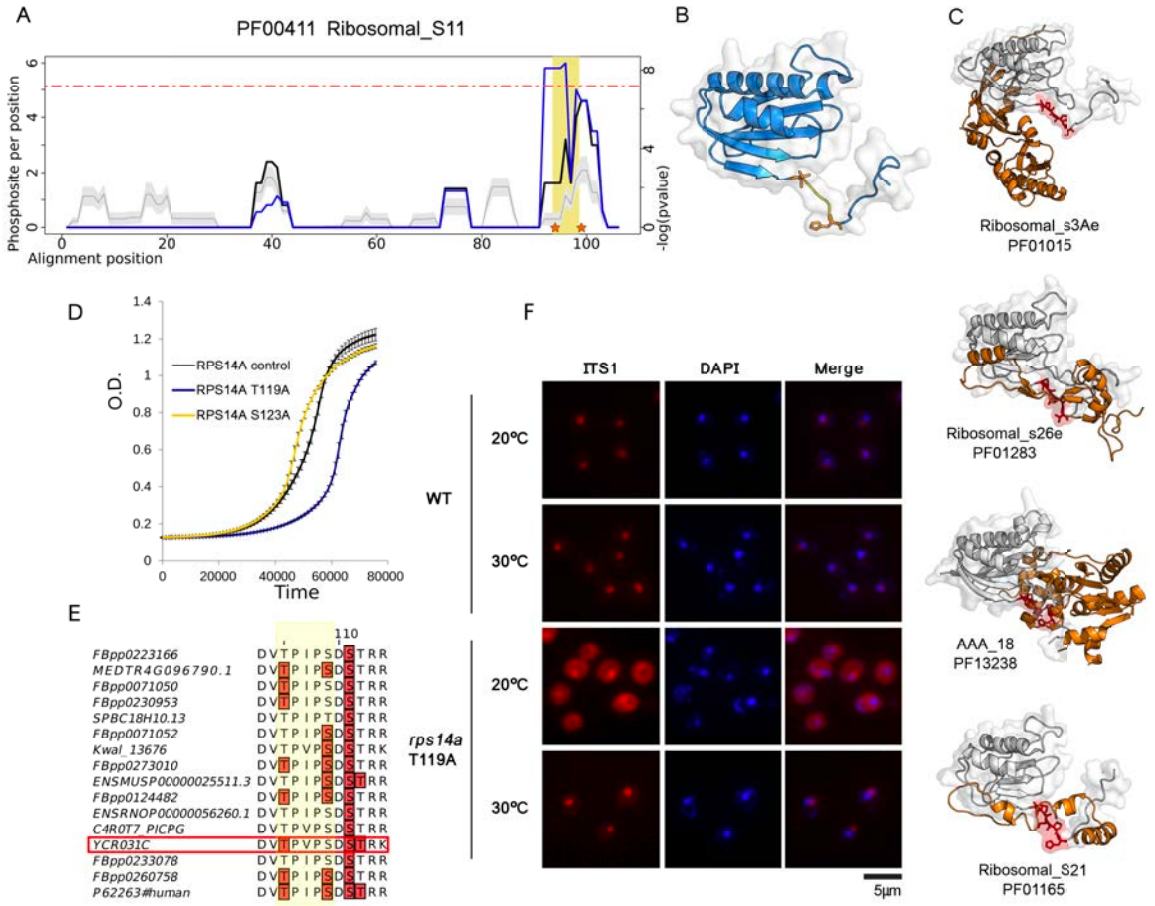


Fig. 3.9 RPS14A phospho-site deficient mutant T119A shows cold shock dependent growth defects and impairment of 20S processing. A) The phosphorylation enrichment over random for the Ribosomal S11 domain (PFAM:PF00411) is plotted in a solid black line. The background expectation is shown in gray line, with standard deviations as gray band. The blue line represents the negative logarithm of p-value (Y axis on the right side). A horizontal red line indicates a cut-off equivalent to a Bonferroni corrected p-value of 0.05. Mutated residues in RPS14A (T119 and S123) are indicated by orange stars in the plot and shown in B) as orange stick representations (PDB:5wntK). C) Structural representation of contacts between the hotspot region of RPS14A (represented in gray) and the ATPase domain of Fap7 (represented in orange). D) Growth curve for RPS14A T119A and S123A mutants in cold shock (25° C) in SC media. E) Conservation of phosphorylation sites in this region across species. F) in situ hybridization with a Cy3-labeled oligonucleotide complementary to the 5' sequence portion of ITS1 was assayed in 30°C and 20° C (cold shock).

3.4 Materials and Methods

Phosphorylation data sources and compilation

Phosphorylated residues *H. sapiens, M. musculus* and *R. norvegicus* were obtained from the Phosphositeplus database (Hornbeck et al. [107]). Phosphorylation data for 6 Drosophila species (*D. ananassae, D. melanogaster, D. pseudoobscura, D. simulans, D. virilis* and *D. yakuba*) was obtained from the iProteinDB in FlyDB database (Hu et al. [109]). Two additional metazoan phosphoproteomes were obtained from published studies for *C. elegans* (Rhoads et al. [234]) and *T. adhaerens* (Ringrose et al. [237]). Phosphosites for 18 fungal species (*S. cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii, S. bayanus, N. castellii, C. glabrata, V. polyspora, Z. rouxii, K. lactis, L. kluyveri, L. waltii, L. thermotolerans, K. pastoris, M. guilliermondii, C. albicans, S. stipitis, S. pombe*) were obtained from Studer and colleagues (Studer et al. [276]). An additional fungal phosphoproteome was added for *M. oryzae* (Franck et al. [84]). Plant phosphosites for 4 species (*A. thaliana, G. max, M. truncatula, O. sativa*) was retrieved from the P3DB database (Yao et al. [312]) and additional plant species were retrieve from selected publications for *B. distachyon* (Lv et al. [172]), *C. reinhardtii* (Wang et al. [297]) and *S. moellendorffii* (Chen et al. [37]). Information for 3 Apicomplexa species (*P. falciparum, P. berghei, T. gondii*) were compiled from phosphoproteomic studies for these species (Treeck et al. [286], Invergo et al. [127]). For all species I removed potential redundancies to avoid assigning the same phosphosites to multiple sequences which could cause false enrichments in proteins rich in isoforms. For *H. sapiens, M. musculus* and *R. norvegicus* retrieved from Phosphositeplus, I removed isoform redundancy by using only the canonical sets of proteins as defined by Uniprot. For other species I filtered out redundant peptides by removing identical 11 amino acid peptides centered around reported phosphosite positions. The total list of phosphosites compiled for this project is shown in Figure 3.1.

Domain mapping, alignment and hotspot predictions

For all of the analysed protein sequences I used PfamScan to predict Pfam domains. The PfamScan option for predicting catalytic residues was used to retrieve annotations on these types of residues. For each Pfam domain, all corresponding sequences having at least 1 phosphorylation site mapped to them were selected and aligned using MAFFT (version 7, using the G-INS-i option) (Katoh et al. [139]). In order to identify alignment regions containing more phosphosites than expected by chance I used a permutation strategy to generate a null background. I first count the observed phosphosites for a given region of a

Pfam domain using a rolling window with a fixed size of 5 positions. I chose to use a window instead of individual positions due to: the uncertainty in phosphosite localization within the phosphopeptides; evolutionary drift whereby the phosphorylation of nearby residues could have the same outcome; potential uncertainty in the alignment. To generate a null background model I randomly select phospho-acceptor residues (serine, threonine and tyrosine) within the alignment respecting the total number of acceptor residues for the 3 amino-acid types. Permutations were repeated 100 times and for each position in the alignment an expected median and standard deviation of phosphorylation is calculated. The observed values were converted to z-scores using the permutation information and then to p-values using the survival function of the normal distribution from scipy (scipy.stats.norm.sf) (Meurer et al. [191]). Only enrichment over random was considered, not depletion and the Bonferroni correction was used to account for multiple testing globally. In addition, to avoid the identification of positions with a low effect size a cut-off of an average of 2 phosphosites per position was used. Finally, contiguous positions were merged to identify domain regions of interest with added +/-2 positions on either side that were defined as phosphorylation hotspot regions.

Identification of representative structures

For each Pfam domain having a significant hotspot region I obtained 3D structures available in the PDB. Then I identified and selected the corresponding Pfam domain regions within each structure excluding the remainder. The structures were filtered to exclude those with gaps larger than 1 amino acid and those shorter than 70% of the longest structural model for a given domain family. Each set of structures representing a domain were structurally clustered with MaxCluster with single linkage (http://www.sbg.bio.ic.ac.uk/ maxcluster) and for each domain one structure was selected from the cluster representing the most common conformation. In order to map the hotspot information to structures, the sequence of the representative structure of each domain was aligned to the corresponding Pfam domain sequences using MAFFT. In some instances the structural model did not cover a predicted hotspot region.

Annotation of interface, catalytic, regulatory, accessibly and disordered residues

For the selected structural models I obtained water accessibility using naccess, disordered prediction with DISOPRED (Ward et al. [298]) and catalytic residues from Pfam (PfamScan). For interface contact information I used data from 3DID (Mosca et al. 2014). For a given

Pfam domain in 3DID I calculated, for each residue, the number of times this residue is found in contact of another protein chain (intra-chain contacts were ignored). I considered a position within a Pfam domain family to be at an interface if it was found forming contacts with other proteins in 10 or more structures. This cut-off was used to obtain a high confidence list of Pfam domain interface positions but similar results were observed considering more lenient definitions of interface positions. A set of human phosphorylation sites known to have regulatory roles were obtained from Phosphositeplus (Hornbeck et al. [105]).

Code deposition and data resources

Alignments and code are provided online on (https://github.com/evocellnet/ptm_hotspots), all the predictions has been shared in a publication (Strumillo et al. [274]) and provided in the supplementaries.

S. cerevisiae phospho-deficient strain construction

Phospho-deficient mutants RPS14A T119A and S123A were constructed using the Y8205 background strain (MATα, his3Δ1; leu2Δ0; ura3Δ0; MET15+; LYS2+; can1Δ::STE2pr-SpHIS5; lyp1Δ::STE3pr-LEU2). SceI endonuclease (from pND32 plasmid) was integrated at the mutated leu2Δ0 locus (Y8205 + leu2Δ0:: natNT2-Gal1pr-I-SceI). The point mutations T119 or S123 were introduced into the RPS14A endogenous locus and the URA3 marker after the stop codon, as described in (Toulmay and Schneiter [284]). The URA3 marker was flanked by SceI recognition sites to enable its removal by the Galactose inducible SceI endonuclease (Khmelinskii et al. [142]). Point mutations were verified by sequencing.

Serial spot dilution assay in *S. cerevisiae*

Yeast strains were grown on agar plates and individual colonies of each strain were picked and arrayed in 96 well plates containing the synthetic SC medium and incubated overnight. The strains were then serially diluted four times at one in 20 dilutions in 96 well plates filled with 160 μl sterile ddH20, the dilutions were performed using a Bio mek FXp liquid handler. The diluted cells were then immediately spotted onto SC + condition agar plates using a VP scientific (VP 405) 96 format manual pinning tool. The agar plates were incubated for 48 and 72 hours and imaged.

Fluorescence *in situ* hybridization and microscopy

Cy3-labeled oligonucleotide probe (5'-Cy3-ATG CTC TTG CCA AAA CAA AAA AAT CCA TTT TCA AAA TTA TTA AAT TTC TT-3') that is complementary to the 5' portion of ITS1 was used to localize 20S pre-rRNA as previously described (Faza et al. [75]). Early biogenesis defect of small ribosomal subunit was determined by localization of uS5-GFP as previously described (Faza et al. [75], Altvater et al. [3]). Cells were visualized using a DMI6000 microscope (Leica, Germany) equipped with a HCX PL Fluotar 63×/1.25 NA oil immersion objective (Leica, Germany). Images were acquired with a fitted digital camera (ORCA-ER; Hamamatsu Photonics, Japan) and Openlab software (Perkin-Elmer, USA).

Plasmids used

pRS316-*RPS2-GFP (uS5)*	*RPS2-GFP CEN URA3 AMP*	(Faza et al. [75])
pRS316-*RPL5-GFP (uL18)*	*RPL5-GFP CEN URA3 AMP*	(Faza et al. [75])

3.5 Discussion

In this project I have identified regions within domain families that are recurrently phosphorylated across different instances of a protein domain family in different proteins and species. Given that often there are multiple copies of the same domain within each genome, a domain centric conservation analysis has increased statistical power over studying conservation across orthologs. However, domain centric approaches will tend to identify only features that are conserved across members of the domain family and will tend to miss gene specific features. This limitation is well illustrated with the protein kinase family. While 75% of human phosphosites with known regulatory roles in kinases are found in the activation loop, several other regulatory sites are found across most of the kinase domain sequence(Figure 2). In addition, the domain models only cover a fraction of the proteome and most phosphorylation occurs outside these regions. Therefore domain hotspots are useful to identify functional important phosphosites but will tend to miss gene specific regulation and cannot provide information for phosphosites outside domain regions.

Although phosphorylation hotspots cover only a fraction of phosphosites, the annotation of these regions, in the context of protein structures, allows us to study how protein domain functions can be regulated. I observed that these hotspots are enriched in positions that are at interfaces or near catalytic residues. This suggests that controlling interactions and regulating access to the catalytic centre may be common mechanisms by which phosphorylation can tune the function of protein domains. For some enzymes that catalyze phosphotransfer

reactions via phosphoenzyme intermediates (i.e. ALP, NDK, PGM/PMM) I observed recurrent detection of phosphorylation of the catalytic residue or neighbouring amino acids. These cases may often represent phosphorylation that is not catalyzed by protein kinases but instead an intermediate enzymatic step or autophosphorylation. This suggests also that mass spectrometry based approaches can be used to track such enzyme reactions via their intermediate phosphorylation states.

With this analysis I was able to identify in total 241 domain phosphorylation hotspots. The identification of phosphorylation hotspots across a diverse set of domains suggests a widespread ancient role for control of protein domains by phosphorylation in eukaryotic species. It remains to be studied how these regulatory regions arise during evolution. These hotspot regions and annotations are provided in Supplementary Table 2and can be used as a resource for future studies. Jointly analysing phosphoproteomics and structural data has allowed me to study the potential functions for the phosphorylation of different regions. However, a structural characterization of the role of such phosphosites will require experimentally determining structures in the phosphorylated form. These studies are typically difficult to perform since it is not straightforward to obtain large amounts of purified phosphoprotein, in particular in a residue specific manner. However, recent progress in genetically encoded phosphorylated residues in protein expression systems (Rogerson et al. [240], Zhang et al. [320]) should make these studies more feasible. Such studies can in turn can spur the rational design of novel phosphorylation switches.

Chapter 4

Overlap of phosphorylation hotspots with conserved ubiquitination sites and recurrent mutations in cancer

4.1 Introduction

In Chapter 3 I have used large scale phosphoproteomic data to identify regions within protein domains that are likely to be regulated by phosphorylation. Compared to other positions in the same domain, I have shown that these phosphorylation hotspot regions tend to be interface residues, close to the catalytic residues and enriched in phosphosites of known function. However, the potential functional importance of the majority of these hotspot regions remains to be studied. The functional importance of these regions can be further studied by analyzing regulation by other PTMs or mutational data. Phosphorylation often acts to either promote or inhibit other PTMs such as ubiquitination that in turn can have functional consequences such as causing the degradation of proteins.

Proteomic analysis of PTMs other than phosphorylation is generally less advanced, however ubiquitination is one of the most prominently studied PTMs with developed mass spectrometry assays (Hunter [117]). The most well characterized role of ubiquitination is marking the protein for degradation via the proteasome by the formation of a poly-ubiquitin chain. However, single and poly ubiquitination can have many other regulatory consequences that have been now described for multiple proteins (Komander and Rape [148]). Increase in regulating information content can be developed when the PTMs act combinatorially (cross talk) (Lonard and O'malley [170]), however the general lack of data prevents thorough

analysis of PTMs interaction. Direct information regarding whether different amino acids are simultaneously modified on a protein molecule or whether they form logical gates could greatly improve the knowledge about the signalling networks. Devising methods to determine which PTMs are functional and which cooperate in cross talk is an urging research task. By describing the ubiquitination hotspots this chapter aims to highlight and analyse some of the functional ubiquitinations that are common across different proteins sharing the same structural domain. Mapping the ubiquitination and phosphorylation hotspots together allows for simultaneous analysis of both PTMs.

Genomes of many human tumors are sequenced in search of the genetic basis of cancer. Large-scale tumor sequencing projects enable the identification of many new cancer gene candidates through computational approaches. These efforts, such as the TCGA and ICGC consortia found millions of somatic mutations in most of human genes across thousands of human samples. One of the current challenges of oncogenomics is to distinguish the so called driver mutations that are involved in tumorigenesis, from those that occur randomly and are neutral to cancer cells (passenger mutations). Computational methods have been developed to address this challenge, most often based on a significant overall burden of mutations or on significant positional clustering of mutations in the gene sequences, corresponding to mutational hotspots. Recurrent mutations of genes has been used in the past as a signal to identify cancer driver genes (Youn and Simon [315]) and cancer driver mutations (Youn and Simon [315], Creixell et al. [54]). Recurrence of mutations within domain families has also already been used to find regions within globular domains that are often mutated in cancer across different proteins and cancer samples (Kamburov et al. [133]).

In this chapter I am mapping mutation hotspots along with the ubiquitination and phosphorylation hotspots to further help discovering regions of these domain families that are most likely to be functionally important. I show with some examples how looking at the overlap of mutations and different PTM types informs on the importance of the PTMs and/or potential mechanism of mutations in cancer.

4.2 Results

4.2.1 Ubiquitination hotspots

In order to study the relationship between conservation of protein phosphorylation, ubiquitination and cancer mutations within protein domain families, I integrated data available for ubiquitin and phosphorylation from different species and cancer mutation data from

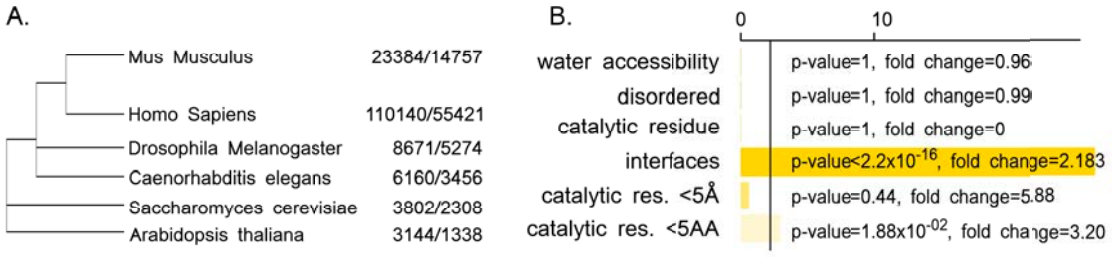


Fig. 4.1 A) Phylogenetic tree of the species containing ubiquitination data. The numbers in the left column correspond to the ubisites per species obtained and the right column the ubisites found within Pfam domains. B) Enrichment over random of structural features comparing ubiquitination hotspots with other residues within the same domains. The tested features include water accessibility — residues with>20% RSA; protein disorder as predicted by DISOPRED; catalytic residues; residues at an interface based on 3DID; residues within 5 amino-acid distance to a catalytic residue; residues within 5 angstroms of a catalytic residue. For each feature I report the -log(p-value).

different tumours types. The outcome is a collection of domains described with hotspots in different combinations of PTMs and recurrent mutations in cancer (described below). In order to identify ubiquitination hotspots I compiled 155,301 ubiquitination sites from 6 species (*H.sapiens, M. musculus, A.thaliana, C. elegans, D. melanogaster* and *S. cerevisiae*) from publicly available data (PhosphoSitePlus) and unpublished results from Eric Bennett's lab (UCSD).

In this analysis (Figure 4.1.A) 82,554 of those sites were located within protein domains as defined by Pfam domain families. As in the previous chapter, I further restricted the analysis to 189 Pfam domains represented by at least 15 different instances, containing a total of 50 or more ubisites and having a representative 3D structure. To identify ubiquitination hotspot regions within these domain families I used the same method as described in the previous chapter. Briefly, the ubiquitination sites were mapped to the domain alignment; a rolling window averaging was used to count the number of observed PTMs and a permutation based strategy used to calculate the significance of the enrichment. After multiple testing correction (Bonferroni correction) and selection of significant regions, I found 46 ubiquitination hotspots in 32 Pfam domains, 41 of which mapped to a pdb structure.

In comparison with phospho-hotspots, ubi-hotspots did not tend to be more water accessible than random residues within the protein. Lysine residues are not likely to be found buried, probably explaining the lack of expected association between ubiquitination and surface accessibility. In fact, the only significant enrichment was for the ubiquitination hotspot residues to be more likely than random at interfaces (Figure 4.1.B). No enrichment

was found for disordered residues, catalytic residues or residues near the catalytic residues. The ubiquitination hotspots do not tend to be more water accessible, or more structurally ordered than other residues in protein domains. They also are not more likely to be a catalytic residue, or be in a close 3D distance to the catalytic residue. However they tend to be in close sequential distance to catalytic residues (p-value=0.02). The most significant outcome for this analysis (p-value<2.2e-16) shows that ubiquitination residues have more interface interactions than other residues in domains.

In collaboration with Inigo Barrio, postdoc from the lab, we analysed the condition specific regulation of ubiquitination hotspots in comparison with other ubiquitination sites reported. For 20 domains containing ubiquitination hotspots, Inigo obtained additional data from MS-based quantitative ubiquitination studies at the site level (Methods). This data contains the quantification of fold changes of abundance levels upon different stimuli. The changes in fold change after stimulation for sites contained within hotspots have been compared using Kolmogorov-Smirnof test with other sites coming from the dataset. After p-value correction we were able to find significant enrichments for some (6 out of 50) of the conditions. Three of the conditions with significant regulation relate to proteasome inhibition (Figure 4.2). MG-132 and Epoxomicin are proteasome inhibitors and PR619 is DUBs inhibitor (which also indirectly inhibits the proteasome). During a proteasomal inhibition experiment, proteins that were targeted for degradation start to accumulate in the cytoplasm. When measuring the ubiquitome, the number of ubiquitination sites responsible for degradation should significantly rise. Functional ubisites that do not lead to protein degradation are not expected to increase after proteasomal inhibition. Unlike ubiquitination sites that target proteins for degradation by the proteasome, the ubisites found within these hotspots are, on average, showing a decrease in levels after proteasome inhibition. This observation suggests that these hotspots ubisites are likely to have non-degreation related functions. Proteins targeted for proteasomal degradation do not have to be polyubiquitinated on a particular residue. Polyubiquitin chain on any Lysine residue is usually enough to trigger the degradation. This analysis suggests ubisites with non-degradation functions may show a higher than average positional conservation, as is the case for the hotspot.

The results that I obtained for phosphorylation strongly indicate that ubisites found in hotspots should more likely be functionally important. Based on the results above these should tend to have non-degradative functions on average but not exclusively. Unlike for phosphorylation where a large set of phosphosites with well characterized functions, no such benchmark set exists for ubisites. Instead of a global benchmark I present two examples of known and well described ubiquitination hotspots.

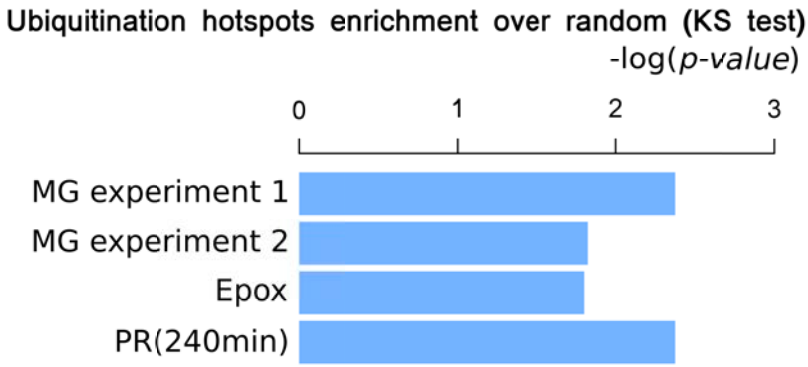


Fig. 4.2 Condition specific regulation of ubisites in ubiquitination hotspots, The regulation of ubisites in hotspots was compared with all other ubisites using a Kolmogorov-Smirnof test. MG experiment 1 and 2, are experiments with MG-132, a proteasome inhibitor. Epox (Epoximicin) is another type of proteasome inhibitor, and PR is DUBs inhibitor. For these conditions, ubisites found in hotspots were found to be significantly down-regulated compared to all other ubisites.

ABC transporters are ATP-binding cassette transporters constituting one of the largest and possibly the oldest gene superfamily (Jones and George [132]). ABC transporters consist of multiple subunits, mainly ABC transporter transmembrane domain (shown in Figure 4.3 top panel) and ATP-binding domain. ABC transporter transmembrane domain consist of six alpha helices harboured in the plasma membrane. Many ABC-transporter family members have been reported to undergo proteasomal degradation (Nakagawa et al. [203]). This includes, for example the yeast protein Ste6. That was the first known example of a membrane protein undergoing ubiquitination contributing to degradation (Kölling and Hollenberg [147]). While the role of ubiquitin mediated degradation in controlling the transporters is well established the regions targeted by ubiquitination is not clear. This analysis identified a ubiquitination hotspot (Figure 4.3, top panel) that may represent a common mode of regulation for the proteins having this domain.

Annexins (Figure 4.3, bottom panel) play a role in protein scaffolding as well as in multiple other processes such as exocytosis and endocytosis. All proteins belonging to annexin family must bind to phospholipids of the membrane in a Calcium dependent manner. This binding results in subsequent events, such as membrane trafficking, signal transduction and exocytosis. Annexins are a known target for PTMs — degradational ubiquitinations (Nasu et al. [204]), sumoylations (Hirata et al. [100]) and phosphorylations (Dorovkov and Ryazanov [68]). The ubiquitination hotspot is located on a different residue than Lysine targeted for degradation, which suggests it may have a non-degradation related function. The

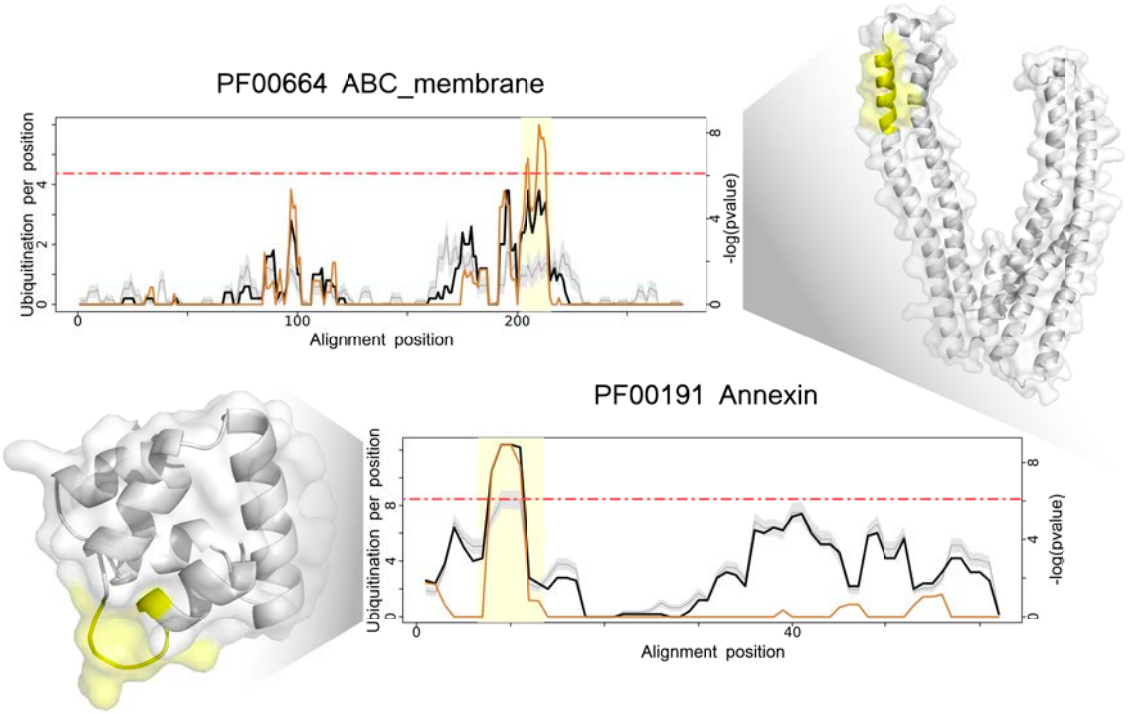


Fig. 4.3 Enrichment over random of protein ubiquitination along the domain sequence for 2 examples of ubiquitination hotspots. The average number of ubiquitinations observed per rolling window is plotted in a solid black line (observed). The background level of expected ubiquitination calculated from random sampling is shown in gray line, with standard deviations as gray band. The yellow line represents the negative logarithm of p-value at each position (right y axis). A horizontal red line indicates a cut-off of the Bonferroni corrected p-value of 0.05. Positions with a -log(p-value) above this cut-off and average ubisites per window higher than 2 are considered putative regulatory regions and highlighted under a vertical yellow bar. In the structural representations the predicted hotspot regions are highlighted in yellow.

lysine often modified in this hotspot is very close in sequence and space to glycines that are involved in the binding of Ca^{2+}. The adjacent location can suggest a possible controlling mechanism of binding.

Based on the phosphorylation analysis, I expect that the ubisite hotspot regions I have identified should be highly enriched in functionally important regulatory regions. These regions are provided in Appendix B.

4.2.2 Mutation hotspots

In addition to the ubisite hotspots I next used mutational data from cancer samples to predict additional regulatory regions in the same domain families. Mutation data from the TCGA pan-cancer atlas dataset included 1,820,460 missense variants from 33 human tumour types (presented in Figure 4.4.A). I analysed 141 domains, from which 77 also had phosphorylation data assigned. Additionally I added mutations from ClinVar database (Landrum et al. [158]) and Uniprot annotations of mutations causing changes in protein function. This simple approach does not take into account that different samples have different background mutation rates, and tumors differ substantially with regard to the number of accumulated mutations. As above, I used the same analysis approach to find regions in domain families that have a higher than randomly expected number of missense mutations. In structural analysis, similarly to phosphorylation hotspots, mutation hotspots show enrichment over random in every examined feature (Figure 4.4.B). Compared to other protein regions, recurrent mutations within the same domain region are more likely to occur in accessible regions but also at interfaces or near catalytic residues. These observations suggest that recurrent mutations in specific domain positions are more likely to have functions other that distabilizing the mutated protein.

Using recurrent mutations within protein and domain families have already been previously shown to identify functionally important regions (Kamburov et al. [133]), I provide here an illustration of two examples of well described mutations contained within mutational hotspots (Figure 4.5). I focus further below on the intersection between the different types of hotspot regions.

The Src Homology 2 (SH2) domain is contained in many intracellular signal-transducing proteins, however its name originates from its existence in Src oncoprotein. SH2 domains are about 100 amino acids long, and associate almost invariably with phosphorylated tyrosine residues. Because of this typical association they play an important role in protein tyrosine kinase (PTK) pathways. They are usually found in adaptor proteins that aid in the signal transduction of receptor tyrosine kinase pathways as well as adaptor domains within tyrosine kinase proteins. Genome wide studies of SH2 domains have revealed many missense mutations that can be connected to human diseases, like Noonan syndrome, diabetes and different types of cancer (Calpe et al. [33], Friedman [86], Kurosaki [154], Shepherd et al. [258], Mattsson et al. [186]). Mutations affecting SH2 domains can be divided into those affecting important functions and into those affecting the structure (Lappalainen et al. [161]). Those affecting functionally important amino acids are involved in ligand binding or interactions with other domains. Some mutations can alter the electrostatic surface potential

A.

total number	in domain	TCGA study name
6470	624	Adrenocortical carcinoma
90037	8874	Bladder Urothelial Carcinoma
68804	6702	Breast invasive carcinoma
50035	4768	Cervical squamous cell carcinoma and endocervical adenocarcinoma
2150	245	Cholangiocarcinoma
142885	14637	Colon adenocarcinoma
4268	383	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
23196	2198	Esophageal carcinoma
40919	4076	Glioblastoma multiforme
72525	7217	Head and Neck squamous cell carcinoma
1856	185	Kidney Chromophobe
16905	1619	Kidney renal clear cell carcinoma
18339	1677	Kidney renal papillary cell carcinoma
5688	442	Acute Myeloid Leukemia
23525	2486	Brain Lower Grade Glioma
33226	3204	Liver hepatocellular carcinoma
147518	14008	Lung adenocarcinoma
119097	11192	Lung squamous cell carcinoma
2149	219	Mesothelioma
30179	3010	Ovarian serous cystadenocarcinoma
18892	2057	Pancreatic adenocarcinoma
1613	182	Pheochromocytoma and Paraganglioma
20234	2187	Prostate adenocarcinoma
37803	3842	Rectum adenocarcinoma
15972	1507	Sarcoma
310981	28421	Skin Cutaneous Melanoma
131905	13870	Stomach adenocarcinoma
1834	202	Testicular Germ Cell Tumors
6962	936	Thyroid carcinoma
2336	231	Thymoma
477141	47200	Uterine Corpus Endometrial Carcinoma
6232	641	Uterine Carcinosarcoma
1211	128	Uveal Melanoma

B.

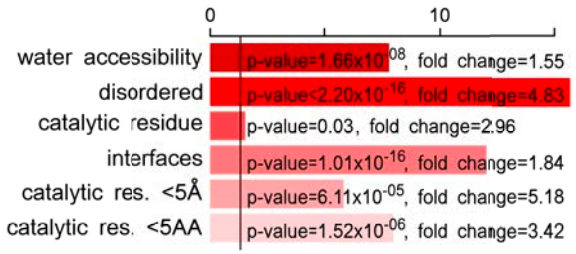


Fig. 4.4 A.) List of missense mutation data coming from TCGA atlas, across 33 cancer types. B.) Enrichment over random of structural features comparing mutation hotspots with other residues within the same domains. The tested features include water accessibility - residues with>20% RSA; protein disorder as predicted by DISOPRED; catalytic residues; residues at an interface based on 3DID; residues within 5 amino-acid distance to a catalytic residue; residues within 5 angstroms of a catalytic residue. For each feature I report the -log(p-value).

of the protein, especially at the pocket that binds the phosphotyrosine (Figure 4.5, top panel). Although mutations within the entire SH2 domain have been reported (Lappalainen et al. [161]), the hotspot analysis identified recurrent mutations in the binding pocket (Figure 4.5, top panel).

Tyrosine kinases are a subfamily of protein kinases, that can distinguish between Serine and Threonine and phosphorylate only Tyrosine. Similarly to Ser and Thr kinases, Tyr kinases are often activated by phosphorylation at the active loop to change conformation and bind substrates (Hubbard and Miller [114]). I identified several mutation hotspots in this domain family (Figure 4.5, bottom panel), one of which coinciding with the activation loop regions (red stripe near the 170 position). It is well recognised that mutations in this region can cause particular kinases to remain constitutively active (Casado et al. [35], Yao et al. [314], Yadav et al. [306]). This constant activity has been linked to several cancers, and drugs that inhibit the catalytic cleft of activated kinases (e.g Imatinib, (Weisberg et al. [299]) have been proven to be effective in treatment.

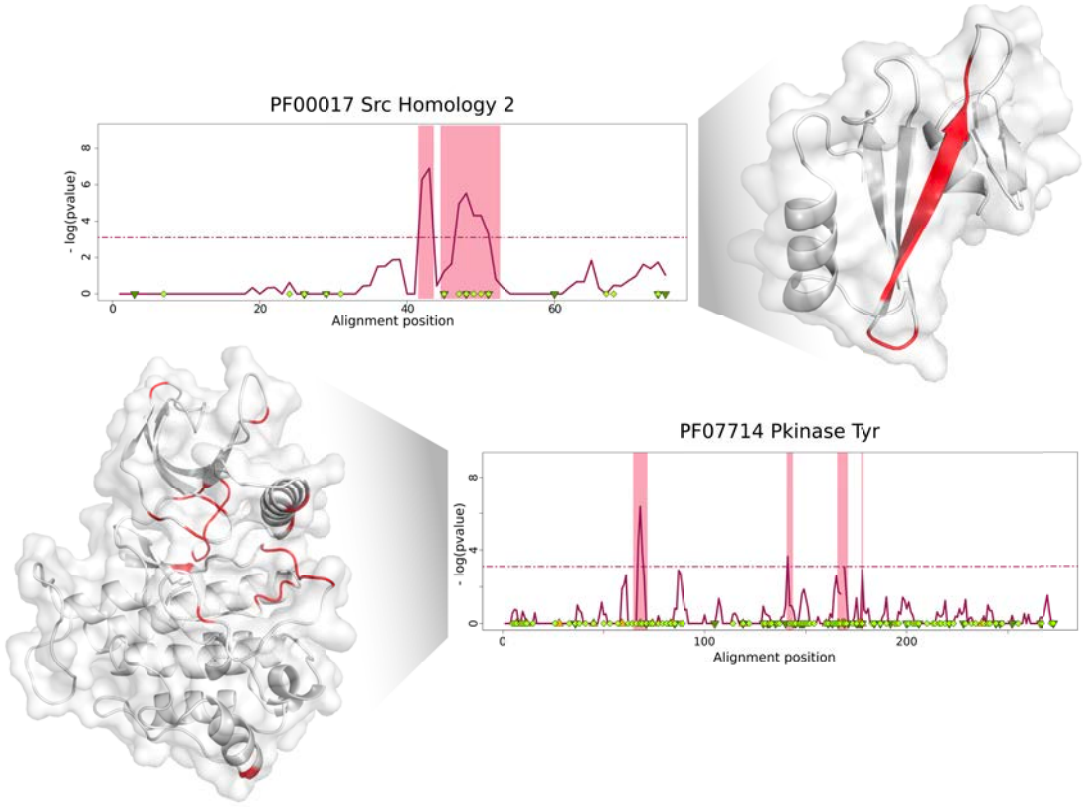


Fig. 4.5 Enrichment over random of missense mutations along the domain sequence for two examples of mutation hotspots. The red line represents the negative logarithm of p-value at each position (y axis). A horizontal red line indicates a cut-off of the Bonferroni corrected p-value of 0.05. Positions with a -log(p-value) above this cut-off and average mutations per window higher than 2 are considered putative regulatory regions and highlighted under a vertical red bar. In the structural representations the predicted hotspot regions are highlighted in red.

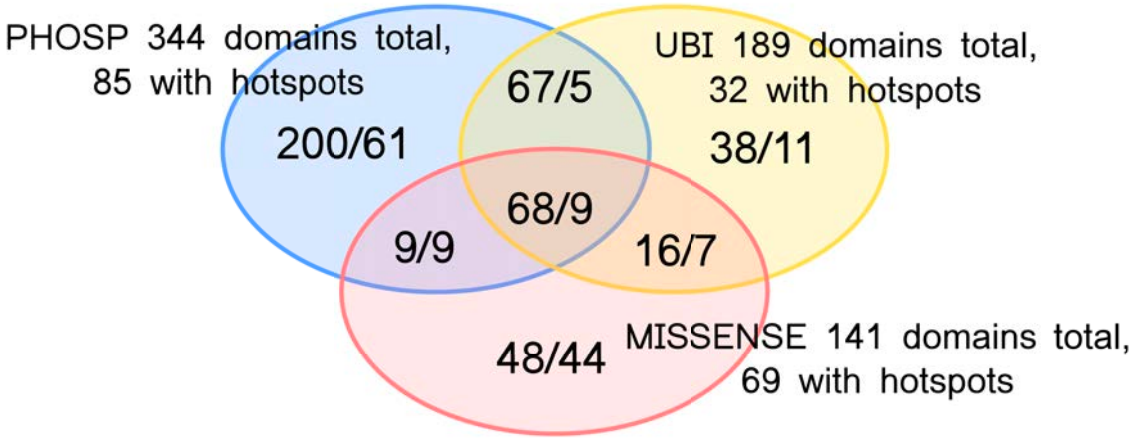


Fig. 4.6 Venn diagram showing the distribution of the domains and hotspots. From 344 domains that were available to analyse for phosphorylation, 144 have been analysed for ubiquitination or missense mutations. Each number is followed by the number of hotspots that were found in each subset along the PDB structure. 68 domains had enough of each data to be analysed for all three kinds of PTMs, however only 9 had hotspots of the 3 types.

4.2.3 Analysing the overlap between phospho, ubi and mutation hotspots

The overlap of the three analysed sets has been constructed by mapping all 3 modifications to the same PDB structure positions. Naturally the common subset involves less domains (Figure 4.6), with 9 domain families having identified hotspot regions of the 3 types. In addition to these 9 cases, some domain families contain hotspots of two different types which may provide novel insight into how the functional regulation of these domains. Additional informative annotations from Clinvar have been added to the domains as well, which just per se is helpful in discovering new dependencies.

In the following sections I present some examples of domains having different types of hotspots from this analysis. I briefly describe here the plots that summarize the findings. Similarly to the hotspots plots showed so far, the x-axis represents the alignment positions for the common pdb structure representing the domain. The negative logarithm of the p-value for each modification is plotted on the y axis - blue line for phosphorylation, yellow for ubiquitination and red for mutations. Foreground and background (previously represented by black and gray lines) for each of the modifications are not shown. This simplification allowed to represent all modifications in the same plot, and highlighted the most important characteristics of each domain. Hotspots are indicated with corresponding colors as vertical rectangles. This is different color scheme than in Chapter 3 - previously, the phosphorylation hotspots were indicated by yellow rectangles, now they are blue (same as the p-value line).

Ubiquitination hotspot regions are in yellow, and mutation plots in red. In cases of overlapping hotspots, the combination of colors has been implied. Overlap of phospho- and ubi- hotspot is green and phospho- and mutation- hotspot purple. Each of these combinations is additionally explained in the figure description for clarity. The thresholds for p-values are represented with dashed lines in corresponding colors. The round red dots are PhosphositePlus annotated human regulatory phosphorylation sites. A set of functional (regulatory) ubiquitination sites has not been available. The predicted active (catalytic) sites in protein are marked with yellow dots. Clinvar mutations are presented with bright green diamonds, and Uniprot downregulating mutations are green triangles pointing down. Yellow upright triangles are the upregulating mutations.

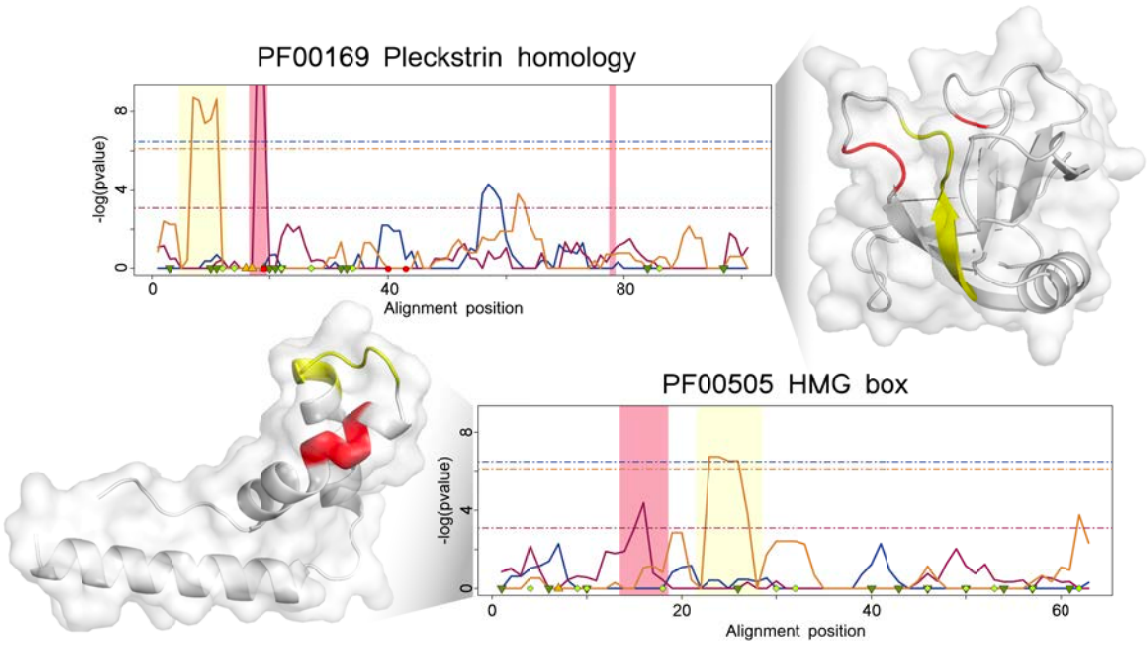


Fig. 4.7 All three (phosphorylation - blue, mutation - red and ubiquitination - yellow) hotspots illustrated for Pleckstrin and HMG box domain families. Only negative logarithms of p-value are shown, in corresponding colors. Hotspots are indicated with vertical bars in corresponding colors. Hotspots are also colored in the structural representations.

4.2.4 Example of domain families with identified ubi and mutation hotspots

Evidence for the functional relevance of hotspots identified in the Pleckstrin Homology domain (PH) family are well illustrated by the PH domain of the Akt kinase. Recruitment of Akt kinase (also known as Protein Kinase B, PKB) from the cytosol to the plasma membrane requires its binding through the PH domain to the PI(3,4,5)P3 phospholipid in the membrane (Yang et al. [310]). After further phosphorylation on Threonine and Serine residue within the kinase domain Akt becomes activated. Ubiquitination of Akt is well correlated with Akt phosphorylation and activation. The PH domain not only mediates PI(3,4,5)P3 lipid binding, but also provides a platform for recruiting critical adaptors important for Akt localization and phosphorylation. Akt undergoes K63-linked ubiquitination at K8 and K14 by TRAF6, which is critical for Akt membrane recruitment and phosphorylation. Both of these sites occur at the predicted yellow ubiquitination hotspot (Figure 4.7, top panel). Mutations on the Akt PH domain were identified in a subset of human cancers. Mutations like E17K display enhanced ubiquitination and enhanced lipid binding which leads to increased activation of Akt. This

mutation can be found in range of the first highlighted mutation hotspot (red) as a yellow triangle pointing up (Figure 4.7, top panel). These observations and the physical proximity in the structural model of the PH domain suggest that there is a functional relationship between these two regions. This result also gives further credibility for the functional importance of the ubi hotspot region found for this domain family.

High mobility group (HMG) box proteins are DNA binding proteins that are abundant and ubiquitous in the cell. They, along with the histones, fulfill global genomic functions in establishing active/inactive chromatin domains. The transcription factor Sox6 (containing HMG box domain) plays an essential role in suppressing genes of slow fiber type-specific muscles. The exact mechanism of how the activity of Sox6 in a skeletal muscle affects the fiber phenotype is unknown, but it has been confirmed that Sox6 undergoes polyubiquitination led degradation (An et al. [7]). This ubiquitination has been located within the ubiquitin hotspot (Figure 4.7, bottom panel, yellow). The E3 ubiquitin ligase that targets Sox6 is Trip12, leading to Sox6 proteasomal degradation, which affects the gene expression in muscle cells. Two more known polyubiquitinations sites are located within the same hotspot: SOX9, where E6-AP acts as a ubiquitin ligase (Hattori et al. [97]); and SOX10, un unstable protein, targeted by Fbxw71 (E3 ubiquitin ligase) for degradation (Lv et al. [173]). These three cases validate the relevance of this ubiquitination hotspot. Unlike for the PH domain, the mutational hotspot found in these domains is not well characterized but given it structural proximity to the ubiquitin hotspot I suggest that these cancer mutations may interfere with the degradation of proteins containing this domain family.

4.2.5 Example of domain families with identified phosphorylation and mutation hotspots

The K homology domain (KH) mostly functions as an RNA recognition domain. The domain can be found in several proteins, where it can exist in multiple copies. However individual domains can perform functions cooperatively or independently. There are several cases where multiple domain copies (eg. 4) are controlled through PTMs, which patterns are different on each of the copies. Due to this fact, analysis based on a single domain can be tricky and does not reveal the whole signalling landscape. There are few examples showing the PTMs influence on the structure - a single phosphorylation on the KH domain can change its conformation (Díaz-Moreno et al. [64]). This phosphorylation is indicated in Figure 4.8, top panel as a red dot within the blue phosphorylation hotspot around position 40. One of the Clinvar datapoints (bright green dot within red mutation hotspot in Figure 4.8, top panel) validates the functional importance of the mutational hotspot regions. It corresponds to the noted mutation of Isoleucine to Asparagine, which can be observed in Fragile X syndrome (Zang et al. [317]). This mutation alters protein folding and stability which has its consequences in changing its localization in Cajal bodies and reducing the association with cytoplasmic granules, polyribosome, and RNA-binding as well as impairing interactions with number of proteins. Unlike some of the previous examples, this mutational hotspot does not further help to define the role of the phosphorylation hotspots.

Another example of a phosphorylation hotspot with functional support can be found in the DEAD-box RNA helicase (DDX3), containing the DEAD domain. The red dot within the third blue phosphorylation hotspot (Figure 4.7, middle panel) represents essential phosphorylation in human DDX3 (Sekiguchi et al. [251]). This phosphorylation on T323 is conserved amongst the 36 human DEAD box type RNA helicases, and it contains the Ib motif which is important for the substrate binding function of the helicase. DDX41 directly binds to dsDNA and CDNs and associates with the adaptor STING (in viral/ bacterial response). Mutation of Y364 or T414 to Phenylalanine compromised the ability of the DDX41 to bind dsDNA, hence the two residues are individually critical for binding. The two mutation hotspots in this domain highlight two beta strands that occur next to the beta strand that contains the above described phosphorylation hotspot. I suggest that cancer mutations in this region are likely to interfere with substrate binding.

The Tyrosine phosphatase family provides a clear example of how cancer mutations overlap directly with a PTM hotspot indicating both the functional relevance of the hotspot and the regulatory impact of the cancer mutations. The phosphatase active site is marked

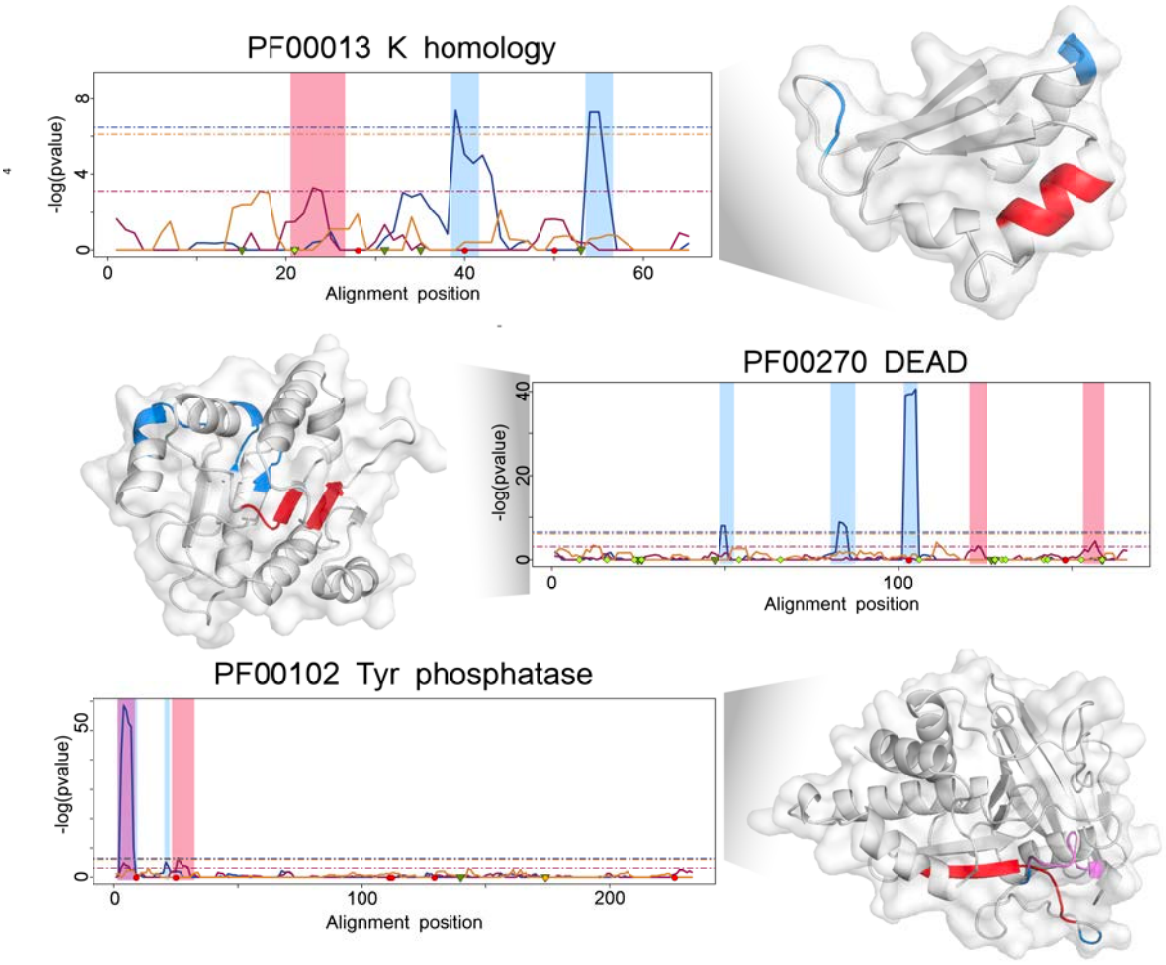


Fig. 4.8 All three (phosphorylation - blue, mutation - red and ubiquitination - yellow) hotspots illustrated for K homology domain (KH) and DEAD domain and Tyrosine phosphatase families. Only negative logarithms of p-value are shown, in corresponding colors. Hotspots are indicated with vertical bars in blue for phosphorylation and red for mutations. The overlaps are indicated by purple.

on the plot with a yellow dot near 175 position in Figure 4.8 (bottom panel). The first known regulatory site marked with the red dot around the position 10, is also within mutation and phosphorylation (purple) hotspot. One of the phosphatases containing this domain is PTPase 1B, which has been shown to undergo phosphorylation by Akt at S50 (first red dot) (Ravichandran et al. [233]). PTP1B contains well studied motif recognised by Akt, RXRXXS/T which also shows a high rate of mutations. The proceeding amino acids play important roles in stabilizing substrates in the catalytic cleft of PTP1B, and mutation hotspot in that region may indicate a potential impairment mechanism by disturbing phosphatase activation. Increased PTP1B activity can lead to insulin and leptin resistance, which can cause type 2 diabetes and many other metabolic and functional disorders (Ma et al. [174]).

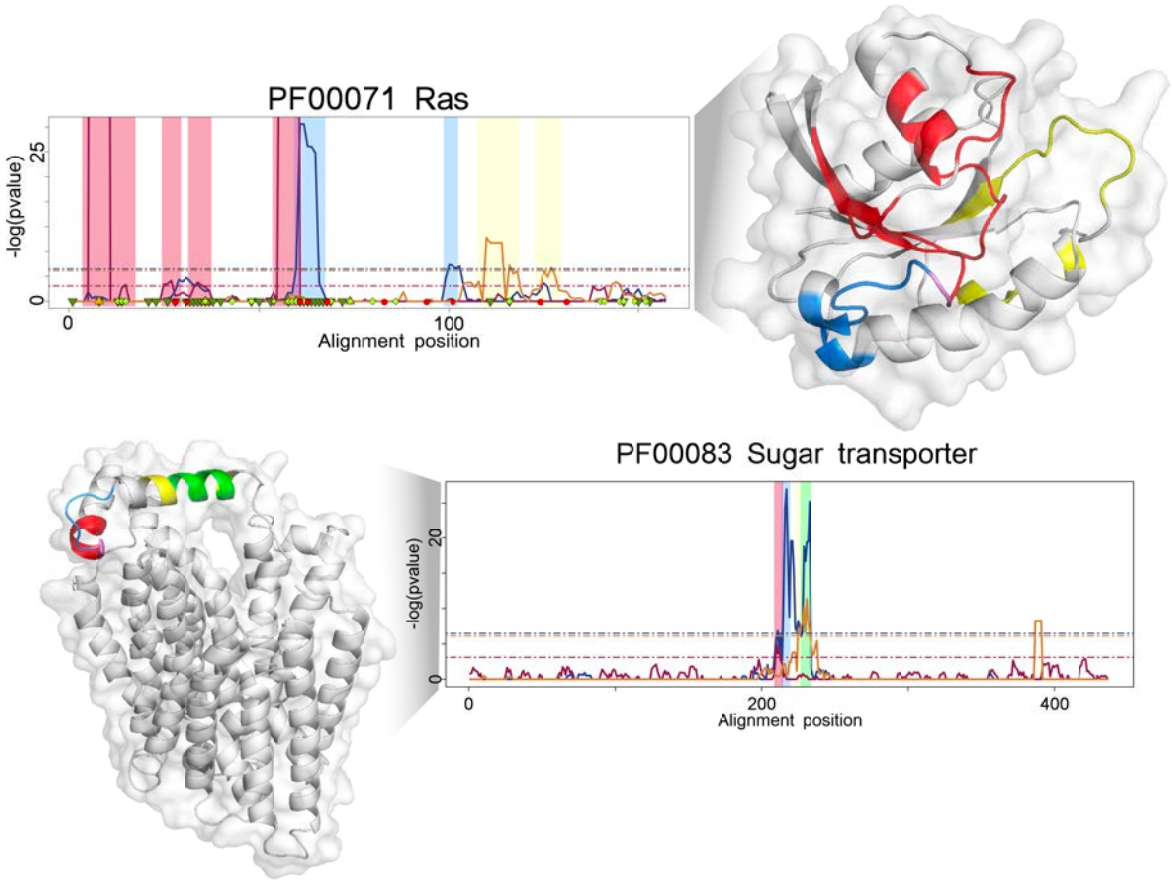


Fig. 4.9 Mapping of all three (phosphorylation - blue, mutation - red and ubiquitination - yellow) negative logarithms of p-values for RAS and Sugar transporter domains. Overlapping phosphorylation and ubiquitination hotspots are colored green..

4.2.6 Examples of domains with multiple regulatory and mutational hotspots

RAS is a small GTPase superfamily which is known to change its conformation depending on the GTP vs GDP binding, and hence transduce extracellular growth signals (Karnoub and Weinberg [137]). It is also one of the first proto-oncogenes identified (DeFeo et al. [60]), and proved to be one of the genes that show most frequent recurrent mutations in a range of human cancers (Pylayeva-Gupta et al. [230]). As described by Kamburov and colleagues (Kamburov et al. [133]) the missense hotspots cluster in the same 3D space of the substrate-binding pocket of the domain. As described in Chapter 3, the main phosphorylation hotpot (blue, Figure 4.9) around position 60 overlaps primarily with a common binding interface. This one of the mutational hotspots overlaps with this phosphorylation region

(purple) validating the importance of the phosphorylation region and the showcasing one common cancer mutation mechanism for this family.

There are also several known examples of non-degradative ubiquitinations of the Ras superfamily proteins, for example in proto-oncogenes H-Ras and N-Ras ubiquitination causes disruption in balance between plasma and endomembrane localization of Ras (Yan et al. [307]). I found two ubiquitination hotspots in the Ras domain family, the first around position 110, and second around position 125. An example of ubiquitination with known function that falls within the region of the first hotspot, is described in (Shin et al. [259]). K140 in RAB5A is found monoubiquitinated and relevant to the protein function (located witin the first, yellow ubiquitination hotspot, Figure 4.9). Ubiquitination of K140 in RAB5A downregulates the interaction with downstream effectors. Other detected but not described ubiquitination is K174 in RAB5C, which is equivalent in sequence to that of K140 in RAB5A, and is not affected by proteasomal inhibition. These and other examples suggest a general tendency of regulation by protein ubiquitination in this family.

The Sugar and other transporters domain family includes glucose transporters (GLUTs), organic anion transporters (OATs) and a few other types of transmembrane proteins. Alterations in the expression or functions of these transporters are known to play important roles in the homeostasis of the organism. Some of the common PTMs have been observed to influence the transporters like glycosylation, phosphorylation, ubiquitination, methylation, and acetylation. Phosphorylation influence has been well described in GLUT1, which is an erythrocyte/brain glucose transporter (Lee et al. [164]). Serine 226 (contained in the presented blue hotspot in Figure 11, bottom panel) in GLUT1 is a substrate of Protein Kinase C. The phosphorylation is required for the quick increase in glucose uptake by the cell. Naturally occuring, pathogenic mutations that cause GLUT1 deficiency syndrome (G1D) disrupt the PKC recognition phosphomotif, which impairs the phosphorylation and blocks the glucose uptake (Figure 4.9, bottom panel, thin purple overlap between red and blue hotspot). A well described ubiquitination example contained within hotspot (yellow) has been found in OATs family (Xu et al. [305]). Xu and coworkers showed for a few members of the family that ubiquitination mediates functional regulation of the transporters.

The examples described above provide evidence of the functional relevance of the PTM and mutational hotspot regions. In several instances the overlap of different types of hotspots gives further evidence of the functional importance of the regulatory regions or the mechanisms by which the cancer mutations are acting. I provide the full list of the PTM and mutational hotspots in supplementary information for future studies.

4.3 Methods

Ubiquitination data sources and compilation

Ubiquitinated residues data for *H. sapiens, M. musculus, A. thaliana, C.elegans, D. melanogaster* and *S. cerevisiae* comes from PhosphoSitePlus database (Hornbeck et al. [107]) and from unpublished results from Eric Bennett's lab (UCSD). For all the species I removed potential redundancies to avoid assigning the same ubisites to multiple sequences which could cause false enrichments in proteins rich in isoforms. For *H. sapiens, M. musculus* and *R. norvegicus* retrieved from Phosphositeplus, I removed isoform redundancy by using only the canonical sets of proteins as defined by Uniprot. For other species I filtered out redundant peptides by removing identical 11 amino acid peptides centered around reported phosphosite positions. The total list of ubiquitinations compiled for this study is shown in Figure 4.6.

Mutation data sources and compilation

I used annotated somatic mutation data for 10,155 tumour samples, obtained from the TCGA pan-cancer atlas dataset downloaded from cBioPortal v1.15.0 on 11/08/2018. The dataset comprises of 1,820,460 missense variants belonging to 33 tumour types. Clinvar mutations are presented with bright green diamonds, and uniprot downregulating mutations are green triangles pointing down. Yellow triangles pointing up are the upregulating mutations. For my analysis of cancer mutations I used the same approach as used for the analysis of phosphorylation and ubiquitination to search for domain regions recurrently mutated in cancer.

Domain mapping, alignment and hotspot predictions

I analysed ubiquitinations and mutations separately, but I describe the same methods for both. For all of the analysed protein sequences I used PfamScan to predict Pfam domains. The PfamScan option for predicting catalytic residues was used to retrieve annotations on these types of residues. For each Pfam domain, all corresponding sequences having at least 1 ubiquitination/mutation site mapped to them were selected and aligned using MAFFT (version 7, using the G-INS-i option) (Katoh et al. [139]). In order to identify alignment regions containing more ubiquitinations/mutations than expected by chance I used a permutation strategy to generate a null background. I first count the observed ubiquitinations/mutations for a given region of a Pfam domain using a rolling window with a fixed size of 5 positions. I chose to use a window instead of individual positions due to the uncertainty in phosphosite

localization within the ubisites and mutations and potential uncertainty in the alignment. To generate a null background model in ubiquitination analyses I randomly select lysine residues within the alignment respecting the total number of ubiquitinated residues. Similarly in the mutation analyses I sample corresponding to mutations amino acids. Permutations were repeated 100 times and for each position in the alignment an expected median and standard deviation of phosphorylation is calculated. The observed values were converted to z-scores using the permutation information and then to p-values using the survival function of the normal distribution from scipy (scipy.stats.norm.sf) (Meurer et al. [191]). Only enrichment over random was considered, and the Bonferroni correction was used to account for multiple testing globally. The Bonferroni threshold for ubiquitination was 2.10e-06 and for the mutations 8.93e-04. In addition, to avoid the identification of positions with a low effect size a cut-off of an average of 2 ubiquitinations/mutations per position was used. Finally, contiguous positions were merged to identify domain regions of interest with added +/-2 positions on either side that were defined as phosphorylation hotspot regions.

Identification of representative structures

For each of the domains that already has been analysed for phosphorylation hotspots I chose the same PDB structure to represent in ubiquitination set and missense set as for phosphorylation (144 domains). This allows for the simultaneous comparison of all 3 (or 2) kinds of hotspots, which are presented in Figure 4.6. The total numbers are not of domains with PDB structures, but domains with enough data (not necessarily with PDB or a hotspot). Each set has important features to be discovered, and however the common set of 68 domains seems to be the most exciting, the domains included are quite well known and well analysed. For domains that are not repeated in phosphorylation set, I obtained all available 3D structures from PDB. Then I identified and selected the corresponding Pfam domain regions within each structure excluding the remainings. The structures were filtered to exclude those with gaps larger than 1 amino acid and those shorter than 70% of the longest structural model for a given domain family. Each set of structures representing a domain were structurally clustered with MaxCluster with single linkage (http://www.sbg.bio.ic.ac.uk/ maxcluster) and for each domain one structure was selected from the cluster representing the most common conformation. In order to map the hotspot information to structures, the sequence of the representative structure of each domain was aligned to the corresponding Pfam domain sequences using MAFFT. In some instances the structural model did not cover a predicted hotspot region.

Annotation of interface, catalytic, regulatory, accessibly and disordered residues

53% of ubiquitination data could be mapped to a domain (82,554 out of 155,301 ubiquitinations). This is a higher percentage than phosphorylations (16%). For the selected structural models I obtained water accessibility using naccess, disordered prediction with DISOPRED (Ward et al. [298]) and catalytic residues from Pfam (PfamScan). For interface contact information I used data from 3DID. (Mosca et al. [200]). For a given Pfam domain in 3DID I calculated for each residue the number of times this residue is found in contact of another protein chain (intra-chain contacts were ignored). I considered a position within a Pfam domain family to be at an interface if it was found forming contacts with other proteins in 10 or more structures. This cut-off was used to obtain a high confidence list of Pfam domain interface positions, but similar results were observed considering more lenient definitions of interface positions. I was not able to obtain a set of ubiquitinations that would be known to have regulatory roles, as I did for phosphosites in Chapter 3.

4.4 Discussion

All plots and structures of remaining proteins are supplied with the supplementary of this chapter. I was able to analyse and address 455 protein domains, for which 68 could be assigned with all three: phosphorylation, ubiquitination and cancer mutation data. Each of the subsets provides valuable resource for future investigation. A couple presented, known examples (annexin, kinase, phosphatase) provided a method control in the absence of ubiquitin or mutational functional assignments. The experimental validation of provided analysis is difficult to perform, since it requires simultaneous examination of multiple PTMs. Addressing ubiquitination function, without provided ubi code, or degradational/non degradational information is very challenging. Experiments require well established Mass Spectrometry protocols, with advanced analysis. However challenging, exploration of multiple PTMs targeting one protein is necessary for exploring more of the signalling networks. I identified regions within domain families that are recurrently ubiquitinated and mutated across different instances of a protein domain family in different proteins and species. Given that often there are multiple copies of the same domain within each genome, a domain centric conservation analysis has increased statistical power over studying conservation across orthologs. Although this analysis covers only a fraction of available phosphorylation, ubiquitination and mutation data, it allows us to study domains regulation. Supplementary

materials cover a diverse set of domains and hotspots, and can be used as a resource for future studies and reference.

Chapter 5

Summary of the projects and future predictions

The first project I have described in Chapter 2 confirms that kinases display a clear enrichment for specific linear sequence motifs. However, the majority of the target sites (both *in vitro* and *in vivo*) miss some of the amino acids defined as important determinants in the motif. The statistics for *in vitro* and *in vivo* targets show a substantial promiscuity of kinases that are able to phosphorylate a number of peptides without the predicted motif. This suggests that the overall possible motif may be smaller than the average motif, that partial motifs can be accepted as long as there is no antimotifs within the peptide, or that the missing amino acids can be substituted with amino acids coming from different parts of the sequence that are in structural proximity. Using a pipeline based on PDB structures I established that the occurrence of 3D motifs is not higher than expected from sampling random serines or threonines, however docking simulations for selected cases show that such interactions are possible. All of the current kinase target predictions based on kinase specificity motifs use statistical approaches to define the most frequent amino acids surrounding a phosphosite. These predictors can be still improved with new interactions data.

In Chapter 3 I presented an extensive and systematic analysis of phosphorylation sites, that uses the conservation in domains as a feature. In the analysed examples I show well known, functional phosphosites and suggest new mechanisms in domain families that show a conserved phospho-function. The overall effectiveness of the method has been additionally proven with a set of confirmed, functional human phosphosites. All of the predictions and alignments are available online, and the project has been published (Strumillo et al. [274]).

In the third project, described in Chapter 4, I include the elements from Chapter 3, but integrating additionally the ubiquitination and cancer missense mutations data. It also maps

the clinvar and mutations that are known to inhibit or activate proteins obtained from Uniprot database to the domain structure. The ubiquitination hotspots mapped to structures provide the same framework for analysis as the phosphorylation project, however the ubiquitination set is smaller than the phosphorylation data. The joint analysis of the three hotspot types allows for the further validation of the functional relevance of the PTM hotspot regions and suggests potential mechanisms of action of the cancer mutations.

The coverage of different PTMs in different species is very uneven, and both phosphorylation and ubiquitination analysis would benefit from greater number of data. Many other PTMs can be analysed in a similar hotspot fashion, however they still lack enough data across different domains and species. Hotspot analysis only addresses the PTMs conserved in domains, and not the PTMs amongst domains linkers or unstructured parts of proteins. Similar statistical analysis could be performed to identify PTM/mutation enrichment in n- and c-terminal regions for specific domain families or linker regions between specific domain combinations

The functional assignments of all the PTMs are the essential knowledge that is still missing, even for the most known PTM which is phosphorylation. As shown in the last two chapters, functional assignment analysis can be accomplished to some extent with the already existing data. However in Chapter 3 I was only able to evaluate a fraction of the available data, because of the domain focused approach. Addressing the functionality of ubiquitinations presents a set of its own problems. The set of known regulatory ubiquitinations is very small, and it consists mostly ubiquitinations tagging for degradation. The limitation of data problem has been highlighted in Chapter 4, where only some of the hotspots could have been assigned to a known regulatory role. Other challenges include establishing the meaning of the ubiquitination code. A complete dataset of E3s and its targets, with corresponding DUBs is highly desirable and could provide a great starting point into the molecular and functional context of ubiquitination. There are a lot of unanswered questions considering ubiquitin, and the answers could provide a better understanding of multiple signalling pathways, drug tolerance and immunity responses.

Future solutions for the field are envisioned with functional readouts for protein interactions, activity, localization, and stability coupled with phosphoproteomics. Especially interactions specific to phosphorylation sites will provide functional information and explore more signalling networks. By identifying interacting compounds of a protein, the protein function can become predictable. Large scale assays of peptide interaction, using phosphorylated and unphosphorylated peptides as control, has already proven efficient and successful in compounds identification (Schulze and Mann [249], Sharma et al. [257]). The functional

analysis can also be accomplished by mutating the PTM sites and phenotypes measurement (Nussinov et al. [209]). Another set of methods includes assays enabling affinity purification coupled with LC-MS to identify phosphorylation specific interactions in different cell types and perturbations (Lemeer et al. [165], Francavilla et al. [83]). Analysis of data coming from these experiments can be speed up with the creation of large peptide interaction libraries (Marx et al. [183]). It is also predicted, that similarly to proteomics, phosphoproteomics will shift from discovering new phosphosites into remeasurement mode in all model organisms (Aebersold [1], Mann et al. [178]).

Other approaches that integrate different 'omics' with PTMs data are in great need. Phosphoproteomics coupled with metabolomics links phosphorylation with metabolic enzymatic activity (Yugi et al. 2014). Similarly, transcriptomics coupled with phosphoproteomics could uncover phosphosites on transcription factors; Organelle-resolved phosphoproteomics (Krahmer et al. [153]) can identify the phosphosites that cause the alteration of localisation of proteins. Knowing all the associations between phosphosites and the protein stability could improve the functional predictions. A research exploring how phosphorylations influence protein stability has not been performed, but can be possible due to TPP - thermal proteome profiling. All of these research ideas could be extended to ubiquitination and other PTMs. Studies with the application of the CRISPR/Cas9 system (Bi et al. [22]), that allow for large-scale knockout studies provide interesting and promising prospect (Cong et al. [46], Mali et al. [177]). Another scenario for discovering complicated signalling networks is optogenetics. A kinase can be inhibited with a photocaged lysine. Such lysine only allows an ATP binding, and hence kinase activation, when exposed to the light, which uncages the lysine (Gautier et al. [88]). This universal approach can be used to examine many kinases and does not have the downsides of genetic manipulations. All the aforementioned research ideas require even better integration of MS technology, greater number of laboratories with the available advanced equipment, and more research on other organisms than human. Cross talk between PTMs requires protocols examining time-resolved quantitative changes of PTMs and protein levels, improved sensitivity and stoichiometry measurements.

References

[1] Aebersold, R. (2003). A mass spectrometric journey into protein and proteome research. *Journal of the American Society for Mass Spectrometry*, 14(7):685–695.

[2] Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A., and Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–1017.

[3] Altvater, M., Chang, Y., Melnik, A., Occhipinti, L., Schütz, S., Rothenbusch, U., Picotti, P., and Panse, V. G. (2012). Targeted proteomics reveals compositional dynamics of 60S pre-ribosomes after nuclear export. *Molecular systems biology*, 8:628.

[4] Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G., and Pandey, A. (2007). A curated compendium of phosphorylation motifs. *Nature biotechnology*, 25(3):285–286.

[5] Amanchy, R., Zhong, J., Molina, H., Chaerkady, R., Iwahori, A., Kalume, D. E., Grønborg, M., Joore, J., Cope, L., and Pandey, A. (2008). Identification of c-src tyrosine kinase substrates using mass spectrometry and peptide microarrays. *Journal of proteome research*, 7(9):3900–3910.

[6] Amano, M., Tsumura, Y., Taki, K., Harada, H., Mori, K., Nishioka, T., Kato, K., Suzuki, T., Nishioka, Y., Iwamatsu, A., and Kaibuchi, K. (2010). A proteomic approach for comprehensively screening substrates of protein kinases such as Rho-Kinase.

[7] An, C.-I., Ganio, E., and Hagiwara, N. (2013). Trip12, a HECT domain E3 ubiquitin ligase, targets sox6 for proteasomal degradation and affects fiber type-specific gene expression in muscle cells. *Skeletal muscle*, 3(1):11.

[8] Andersson, L. and Porath, J. (1986). Isolation of phosphoproteins by immobilized metal (fe3) affinity chromatography.

[9] Ballif, B. A., Villén, J., Beausoleil, S. A., Schwartz, D., and Gygi, S. P. (2004). Phosphoproteomic analysis of the developing mouse brain. *Molecular & cellular proteomics: MCP*, 3(11):1093–1101.

[10] Barford, D., Hu, S. H., and Johnson, L. N. (1991). Structural mechanism for glycogen phosphorylase control by phosphorylation and AMP. *Journal of molecular biology*, 218(1):233–260.

[11] Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villén, J., Li, J., Cohn, M. A., Cantley, L. C., and Gygi, S. P. (2004). Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(33):12130–12135.

[12] Beck, M., Claassen, M., and Aebersold, R. (2011). Comprehensive proteomics. *Current opinion in biotechnology*, 22(1):3–8.

[13] Beli, P., Lukashchuk, N., Wagner, S. A., Weinert, B. T., Olsen, J. V., Baskcomb, L., Mann, M., Jackson, S. P., and Choudhary, C. (2012). Proteomic investigations reveal a role for RNA processing factor THRAP3 in the DNA damage response. *Molecular cell*, 46(2):212–225.

[14] Belozerov, V. E., Y. Lin, Z., C. Gingras, A., McDermott, J. C., and Michael Siu, K. W. (2012). High-Resolution protein interaction map of the drosophila melanogaster p38 Mitogen-Activated protein kinases reveals limited functional redundancy.

[15] Beltrao, P., Albanèse, V., Kenner, L. R., Swaney, D. L., Burlingame, A., Villén, J., Lim, W. A., Fraser, J. S., Frydman, J., and Krogan, N. J. (2012). Systematic functional prioritization of protein posttranslational modifications. *Cell*, 150(2):413–425.

[16] Benayoun, B. A. and Veitia, R. A. (2009). A post-translational modification code for transcription factors: sorting through a sea of signals. *Trends in cell biology*, 19(5):189–197.

[17] Bennetzen, M. V., Larsen, D. H., Bunkenborg, J., Bartek, J., Lukas, J., and Andersen, J. S. (2010). Site-specific phosphorylation dynamics of the nuclear proteome during the DNA damage response. *Molecular & cellular proteomics: MCP*, 9(6):1314–1323.

[18] Bensimon, A., Schmidt, A., Ziv, Y., Elkon, R., Wang, S.-Y., Chen, D. J., Aebersold, R., and Shiloh, Y. (2010). ATM-dependent and -independent dynamics of the nuclear phosphoproteome after DNA damage. *Science signaling*, 3(151):rs3.

[19] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.

[20] Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J. C., Huang, J. H., Alexander, S., Du, J., Kau, T., Thomas, R. K., Shah, K., Soto, H., Perner, S., Prensner, J., Debiasi, R. M., Demichelis, F., Hatton, C., Rubin, M. A., Garraway, L. A., Nelson, S. F., Liau, L., Mischel, P. S., Cloughesy, T. F., Meyerson, M., Golub, T. A., Lander, E. S., Mellinghoff, I. K., and Sellers, W. R. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50):20007–20012.

[21] Betts, M. J., Wichmann, O., Utz, M., Andre, T., Petsalaki, E., Minguez, P., Parca, L., Roth, F. P., Gavin, A.-C., Bork, P., and Russell, R. B. (2017). Systematic identification of phosphorylation-mediated protein interaction switches. *PLoS computational biology*, 13(3):e1005462.

[22] Bi, P., Ramirez-Martinez, A., Li, H., Cannavino, J., McAnally, J. R., Shelton, J. M., Sánchez-Ortiz, E., Bassel-Duby, R., and Olson, E. N. (2017). Control of muscle formation by the fusogenic micropeptide myomixer. *Science*, 356(6335):323–327.

[23] Biemann, K., Cone, C., Webster, B. R., and Arsenault, G. P. (1966). Determination of the amino acid sequence in oligopeptides by computer interpretation of their High-Resolution mass spectra1. *Journal of the American Chemical Society*, 88(23):5598–5606.

[24] Biondi, R. M. and Nebreda, A. R. (2003). Signalling specificity of Ser/Thr protein kinases through docking-site-mediated interactions. *Biochemical Journal*, 372(Pt 1):1–13.

[25] Blagoev, B., Ong, S.-E., Kratchmarova, I., and Mann, M. (2004). Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nature biotechnology*, 22(9):1139–1145.

[26] Blom, N., Gammeltoft, S., and Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of molecular biology*, 294(5):1351–1362.

[27] Blume-Jensen, P. and Hunter, T. (2001). Oncogenic kinase signalling. *Nature*, 411(6835):355–365.

[28] Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S., and Heck, A. J. R. (2009). Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics.

[29] Bose, R., Holbert, M. A., Pickin, K. A., and Cole, P. A. (2006). Protein tyrosine kinase–substrate interactions. *Current opinion in structural biology*, 16(6):668–675.

[30] Bradley, D., Viéitez, C., Rajeeve, V., Cutillas, P. R., and Beltrao, P. (2018). Global analysis of specificity determinants in eukaryotic protein kinases.

[31] Brunet, M. A., Brunelle, M., Lucier, J.-F., Delcourt, V., Levesque, M., Grenier, F., Samandi, S., Leblanc, S., Aguilar, J.-D., Dufour, P., Jacques, J.-F., Fournier, I., Ouangraoua, A., Scott, M. S., Boisvert, F.-M., and Roucou, X. (2019). OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic acids research*, 47(D1):D403–D410.

[32] Buss, H., Dörrie, A., Schmitz, M. L., Frank, R., and others (2004). Phosphorylation of serine 468 by GSK-3β negatively regulates basal p65 NF-κB activity. *Journal of Biological*.

[33] Calpe, S., Wang, N., Romero, X., Berger, S. B., Lanyi, A., Engel, P., and Terhorst, C. (2008). The SLAM and SAP gene families control innate and adaptive immune responses. In *Advances in Immunology*, volume 97, pages 177–250. Academic Press.

[34] Cañas, B., López-Ferrer, D., Ramos-Fernández, A., Camafeita, E., and Calvo, E. (2006). Mass spectrometry technologies for proteomics. *Briefings in functional genomics*, 4(4):295–320.

[35] Casado, P., Wilkes, E. H., Miraki-Moud, F., Hadi, M. M., Rio-Machin, A., Rajeeve, V., Pike, R., Iqbal, S., Marfa, S., Lea, N., Best, S., Gribben, J., Fitzgibbon, J., and Cutillas, P. R. (2017). Proteomic and genomic integration identifies kinase and differentiation determinants of kinase inhibitor sensitivity in leukemia cells. *Leukemia*.

[36] Chang, F., Lemmon, C., Lietha, D., Eck, M., and Romer, L. (2011). Tyrosine phosphorylation of rac1: a role in regulation of cell spreading. *PloS one*, 6(12):e28587.

[37] Chen, X., Chan, W. L., Zhu, F.-Y., and Lo, C. (2014). Phosphoproteomic analysis of the non-seed vascular plant model selaginella moellendorffii. *Proteome science*, 12:16.

[38] Chou, M. F. and Schwartz, D. (2011a). Biological sequence motif discovery using motif-x. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, pages 13–15.

[39] Chou, M. F. and Schwartz, D. (2011b). Using the scan-x web site to predict protein post-translational modifications. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 13:Unit 13.16.

[40] Choudhary, C. and Mann, M. (2010). Decoding signalling networks by mass spectrometry-based proteomics. *Nature reviews. Molecular cell biology*, 11(6):427–439.

[41] Ciechanover, A., Finley, D., and Varshavsky, A. (1984). The ubiquitin-mediated proteolytic pathway and mechanisms of energy-dependent intracellular protein degradation. *Journal of cellular biochemistry*, 24(1):27–53.

[42] Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133.

[43] Clark, I. E., Dodson, M. W., Jiang, C., Cao, J. H., Huh, J. R., Seol, J. H., Yoo, S. J., Hay, B. A., and Guo, M. (2006). Drosophila pink1 is required for mitochondrial function and interacts genetically with parkin. *Nature*, 441(7097):1162–1166.

[44] Coba, M. P., Pocklington, A. J., Collins, M. O., Kopanitsa, M. V., Uren, R. T., Swamy, S., Croning, M. D. R., Choudhary, J. S., and Grant, S. G. N. (2009). Neurotransmitters drive combinatorial multistate postsynaptic density networks. *Science signaling*, 2(68):ra19.

[45] Cohen, P. and Knebel, A. (2006). KESTREL: a powerful method for identifying the physiological substrates of protein kinases. *Biochemical Journal*, 393(Pt 1):1–6.

[46] Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819–823.

[47] Conrads, T. P., Alving, K., Veenstra, T. D., Belov, M. E., Anderson, G. A., Anderson, D. J., Lipton, M. S., Paša-Tolić, L., Udseth, H. R., Chrisler, W. B., Thrall, B. D., and Smith, R. D. (2001). Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and15N-Metabolic labeling.

[48] Cowan-Jacob, S. W. (2006). Structural biology of protein tyrosine kinases. *Cellular and molecular life sciences: CMLS*, 63(22):2608–2625.

[49] Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics: MCP*, 13(9):2513–2526.

[50] Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367.

[51] Cox, J. and Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry*, 80:273–299.

[52] Creixell, P. and Linding, R. (2012). Cells, shared memory and breaking the PTM code. *Molecular systems biology*, 8:598.

[53] Creixell, P., Palmeri, A., Miller, C. J., Lou, H. J., Santini, C. C., Nielsen, M., Turk, B. E., and Linding, R. (2015a). Unmasking determinants of specificity in the human kinome. *Cell*, 163(1):187–201.

[54] Creixell, P., Schoof, E. M., Simpson, C. D., Longden, J., Miller, C. J., Lou, H. J., Perryman, L., Cox, T. R., Zivanovic, N., Palmeri, A., Wesolowska-Andersen, A., Helmer-Citterich, M., Ferkinghoff-Borg, J., Itamochi, H., Bodenmiller, B., Erler, J. T., Turk, B. E., and Linding, R. (2015b). Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell*, 163(1):202–217.

[55] Danielsson, A., Ost, A., Nystrom, F. H., and Strålfors, P. (2005). Attenuation of insulin-stimulated insulin receptor substrate-1 serine 307 phosphorylation in insulin resistance of type 2 diabetes. *The Journal of biological chemistry*, 280(41):34389–34392.

[56] Dar, H. H. and Chakraborti, P. K. (2010). Intermolecular phosphotransfer is crucial for efficient catalytic activity of nucleoside diphosphate kinase. *Biochemical Journal*, 430(3):539–549.

[57] Daub, H., Olsen, J. V., Bairlein, M., Gnad, F., Oppermann, F. S., Körner, R., Greff, Z., Kéri, G., Stemmann, O., and Mann, M. (2008). Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. *Molecular cell*, 31(3):438–448.

[58] Dayon, L. (2008). Hainard a. licker v. turck n. kuhn k. hochstrasser DF. burkhard PR. sanchez JC. relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. *Analytical chemistry*, 80:2921–2931.

[59] de Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fröhlich, F., Walther, T. C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217):1251–1254.

[60] DeFeo, D., Gonda, M. A., Young, H. A., Chang, E. H., Lowy, D. R., Scolnick, E. M., and Ellis, R. W. (1981). Analysis of two divergent rat genomic clones homologous to the transforming gene of harvey murine sarcoma virus. *Proceedings of the National Academy of Sciences of the United States of America*, 78(6):3328–3332.

[61] Delom, F. and Chevet, E. (2006). In vitro mapping of calnexin interaction with ribosomes. *Biochemical and biophysical research communications*, 341(1):39–44.

[62] Dente, L., Vetriani, C., Zucconi, A., Pelicci, G., Lanfrancone, L., Pelicci, P. G., and Cesareni, G. (1997). Modified phage peptide libraries as a tool to study specificity of phosphorylation and recognition of tyrosine containing peptides 1 1edited by j. karn.

[63] Deshaies, R. J. and Ferrell, Jr, J. E. (2001). Multisite phosphorylation and the countdown to S phase. *Cell*, 107(7):819–822.

[64] Díaz-Moreno, I., Hollingworth, D., Frenkiel, T. A., Kelly, G., Martin, S., Howell, S., García-Mayoral, M., Gherzi, R., Briata, P., and Ramos, A. (2009). Phosphorylation-mediated unfolding of a KH domain regulates KSRP localization via 14-3-3 binding. *Nature structural & molecular biology*, 16(3):238–246.

[65] Diella, F., Cameron, S., Gemünd, C., Linding, R., Via, A., Kuster, B., Sicheritz-Pontén, T., Blom, N., and Gibson, T. J. (2004). Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC bioinformatics*, 5(1):1–5.

[66] Dominguez, C., Boelens, R., and Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737.

[67] Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., and Mechtler, K. (2014). MS amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra.

[68] Dorovkov, M. V. and Ryazanov, A. G. (2004). Phosphorylation of annexin I by TRPM7 channel-kinase. *The Journal of biological chemistry*, 279(49):50643–50646.

[69] Duarte, M. L., Pena, D. A., Nunes Ferraz, F. A., Berti, D. A., Paschoal Sobreira, T. J., Costa-Junior, H. M., Abdel Baqui, M. M., Disatnik, M.-H., Xavier-Neto, J., Lopes de Oliveira, P. S., and Schechtman, D. (2014). Protein folding creates structure-based, non-contiguous consensus phosphorylation motifs recognized by kinases. *Science signaling*, 7(350):ra105.

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

[71] Edwards, A. M., Isserlin, R., Bader, G. D., Frye, S. V., Willson, T. M., and Yu, F. H. (2011). Too many roads not taken. *Nature*, 470(7333):163–165.

[72] Eidenmüller, J., Fath, T., Maas, T., Pool, M., Sontag, E., and Brandt, R. (2001). Phosphorylation-mimicking glutamate clusters in the proline-rich region are sufficient to simulate the functional deficiencies of hyperphosphorylated tau protein. *Biochemical Journal*, 357(Pt 3):759–767.

[73] Elia, A. E. H., Rellos, P., Haire, L. F., Chao, J. W., Ivins, F. J., Hoepker, K., Mohammad, D., Cantley, L. C., Smerdon, S. J., and Yaffe, M. B. (2003). The molecular basis for phosphodependent substrate targeting and regulation of plks by the polo-box domain. *Cell*, 115(1):83–95.

[74] Fabbro, E. D., Del Fabbro, E., Inui, A., and Strasser, F. (2012). Emerging treatments, current challenges, and future directions.

[75] Faza, M. B., Chang, Y., Occhipinti, L., Kemmler, S., and Panse, V. G. (2012). Role of Mex67-Mtr2 in the nuclear export of 40S pre-ribosomes. *PLoS genetics*, 8(8):e1002915.

[76] Feldman, R. M., Correll, C. C., Kaplan, K. B., and Deshaies, R. J. (1997). A complex of cdc4p, skp1p, and cdc53p/cullin catalyzes ubiquitination of the phosphorylated CDK inhibitor sic1p. *Cell*, 91(2):221–230.

[77] Fernández-de Cossio, J., Gonzalez, J., and Besada, V. (1995). A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. *Bioinformatics*, 11(4):427–434.

[78] Ferrell, Jr, J. E. and Ha, S. H. (2014). Ultrasensitivity part II: multisite phosphorylation, stoichiometric inhibitors, and positive feedback. *Trends in biochemical sciences*, 39(11):556–569.

[79] Ficarro, S. B., McCleland, M. L., Stukenberg, P. T., Burke, D. J., Ross, M. M., Shabanowitz, J., Hunt, D. F., and White, F. M. (2002). Phosphoproteome analysis by mass spectrometry and its application to saccharomyces cerevisiae. *Nature biotechnology*, 20(3):301–305.

[80] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). The pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1):D279–85.

[81] Fischer, E. H. and Krebs, E. G. (1955). Conversion of phosphorylase b to phosphorylase a in muscle extracts. *The Journal of biological chemistry*, 216(1):121–132.

[82] Fort, K. L., Dyachenko, A., Potel, C. M., Corradini, E., Marino, F., Barendregt, A., Makarov, A. A., Scheltema, R. A., and Heck, A. J. R. (2016). Implementation of ultraviolet photodissociation on a benchtop Q exactive mass spectrometer and its application to phosphoproteomics.

[83] Francavilla, C., Rigbolt, K. T. G., Emdal, K. B., Carraro, G., Vernet, E., Bekker-Jensen, D. B., Streicher, W., Wikström, M., Sundström, M., Bellusci, S., Cavallaro, U., Blagoev, B., and Olsen, J. V. (2013). Functional proteomics defines the molecular switch underlying FGF receptor trafficking and cellular outputs. *Molecular cell*, 51(6):707–722.

[84] Franck, W. L., Gokce, E., Randall, S. M., Oh, Y., Eyre, A., Muddiman, D. C., and Dean, R. A. (2015). Phosphoproteome analysis links protein phosphorylation to cellular remodeling and metabolic adaptation during magnaporthe oryzae appressorium development. *Journal of proteome research*, 14(6):2408–2424.

[85] Frese, C. K., Maarten Altelaar, A. F., van den Toorn, H., Nolting, D., Griep-Raming, J., Heck, A. J. R., and Mohammed, S. (2012). Toward full peptide sequence coverage by dual fragmentation combining Electron-Transfer and Higher-Energy collision dissociation tandem mass spectrometry.

[86] Friedman, E. (1995). The role of ras GTPase activating protein in human tumorigenesis. *Pathobiology: journal of immunopathology, molecular and cellular biology*, 63(6):348–350.

[87] Fukunaga, R. and Hunter, T. (1997). MNK1, a new MAP kinase-activated protein kinase, isolated by a novel expression screening method for identifying protein kinase substrates. *The EMBO journal*.

[88] Gautier, A., Deiters, A., and Chin, J. W. (2011). Light-activated kinases enable temporal dissection of signaling networks in living cells. *Journal of the American Chemical Society*, 133(7):2124–2127.

[89] Goldknopf, I. L., Taylor, C. W., Baum, R. M., Yeoman, L. C., Olson, M. O., Prestayko, A. W., and Busch, H. (1975). Isolation and characterization of protein a24, a" histone-like" non-histone chromosomal protein. *The Journal of biological chemistry*, 250(18):7182–7187.

[90] Goldsmith, E. J., Akella, R., Min, X., Zhou, T., and Humphreys, J. M. (2007). Substrate and docking interactions in serine/threonine protein kinases. *Chemical reviews*, 107(11):5065–5081.

[91] Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P., Faergeman, N. J., Mann, M., and Jensen, O. N. (2005). Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Molecular & cellular proteomics: MCP*, 4(3):310–327.

[92] Grundke-Iqbal, I., Iqbal, K., Tung, Y. C., Quinlan, M., Wisniewski, H. M., and Binder, L. I. (1986). Abnormal phosphorylation of the microtubule-associated protein tau (tau) in alzheimer cytoskeletal pathology. *Proceedings of the National Academy of Sciences of the United States of America*, 83(13):4913–4917.

[93] Gu, B. and Zhu, W.-G. (2012). Surf the post-translational modification network of p53 regulation. *International journal of biological sciences*, 8(5):672.

[94] Habelhah, H., Shah, K., Huang, L., Burlingame, A. L., Shokat, K. M., and Ronai, Z. (2001). Identification of new JNK substrate using ATP pocket mutant JNK and a corresponding ATP analogue. *The Journal of biological chemistry*, 276(21):18090–18095.

[95] Hamill, S., Lou, H. J., Turk, B. E., and Boggon, T. J. (2016). Structural basis for noncanonical substrate recognition of Cofilin/ADF proteins by LIM kinases. *Molecular cell*, 62(3):397–408.

[96] Hanks, S. K. and Hunter, T. (1995). Protein kinases 6. the eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 9(8):576–596.

[97] Hattori, T., Kishino, T., Stephen, S., Eberspaecher, H., Maki, S., Takigawa, M., de Crombrugghe, B., and Yasuda, H. (2013). E6-AP/UBE3A protein acts as a ubiquitin ligase toward SOX9 protein. *The Journal of biological chemistry*, 288(49):35138–35148.

[98] Hedstrom, L. (2009). IMP dehydrogenase: structure, mechanism, and inhibition. *Chemical reviews*, 109(7):2903–2928.

[99] Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., and Flicek, P. (2016). Ensembl comparative genomics resources. *Database: the journal of biological databases and curation*, 2016.

[100] Hirata, F., Thibodeau, L. M., and Hirata, A. (2010). Ubiquitination and SUMOylation of annexin A1 and helicase activity. *Biochimica et biophysica acta*, 1800(9):899–905.

[101] Hoffman, N. J., Parker, B. L., Chaudhuri, R., Fisher-Wellman, K. H., Kleinert, M., Humphrey, S. J., Yang, P., Holliday, M., Trefely, S., Fazakerley, D. J., Stöckli, J., Burchfield, J. G., Jensen, T. E., Jothi, R., Kiens, B., Wojtaszewski, J. F. P., Richter, E. A., and James, D. E. (2015). Global phosphoproteomic analysis of human skeletal muscle reveals a network of Exercise-Regulated kinases and AMPK substrates. *Cell metabolism*, 22(5):922–935.

[102] Holland, P. M. and Cooper, J. A. (1999). Protein modification: docking sites for kinases. *Current biology: CB*, 9(9):R329–31.

[103] Horn, H., Schoof, E. M., Kim, J., Robin, X., Miller, M. L., Diella, F., Palma, A., Cesareni, G., Jensen, L. J., and Linding, R. (2014). KinomeXplorer: an integrated platform for kinome biology studies. *Nature methods*, 11(6):603–604.

[104] Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2011). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research*, page gkr1122.

[105] Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research*, 40(Database issue):D261–70.

[106] Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015a). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research*, 43(Database issue):D512–20.

[107] Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015b). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research*, 43(Database issue):D512–20.

[108] Hu, J., Yu, H., Kornev, A. P., Zhao, J., Filbert, E. L., Taylor, S. S., and Shaw, A. S. (2011). Mutation that blocks ATP binding creates a pseudokinase stabilizing the scaffolding function of kinase suppressor of ras, CRAF and BRAF. *Proceedings of the National Academy of Sciences of the United States of America*, 108(15):6067–6072.

[109] Hu, Y., Sopko, R., Chung, V., Studer, R. A., Landry, S. D., Liu, D., Rabinow, L., Gnad, F., Beltrao, P., and Perrimon, N. (2018). iProteinDB: an integrative database of drosophila post-translational modifications.

[110] Huang, C. Y. and Ferrell, Jr, J. E. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19):10078–10083.

[111] Huang, H.-D., Lee, T.-Y., Tzeng, S.-W., and Horng, J.-T. (2005). KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic acids research*, 33(Web Server issue):W226–9.

[112] Huang, S.-Y., Tsai, M.-L., Chen, G.-Y., Wu, C.-J., and Chen, S.-H. (2007a). A systematic MS-based approach for identifying in vitro substrates of PKA and PKG in rat uteri. *Journal of proteome research*, 6(7):2674–2684.

[113] Huang, S.-Y., Tsai, M.-L., Chen, G.-Y., Wu, C.-J., and Chen, S.-H. (2007b). A systematic MS-Based approach for IdentifyinginvitroSubstrates of PKA and PKG in rat uteri.

[114] Hubbard, S. R. and Miller, W. T. (2007). Receptor tyrosine kinases: mechanisms of activation and signaling. *Current opinion in cell biology*, 19(2):117–123.

[115] Humphrey, S. J., Yang, G., Yang, P., Fazakerley, D. J., Stöckli, J., Yang, J. Y., and James, D. E. (2013). Dynamic adipocyte phosphoproteome reveals that akt directly regulates mTORC2. *Cell metabolism*, 17(6):1009–1020.

[116] Hunter, T. (1995). Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell*, 80(2):225–236.

[117] Hunter, T. (2007). The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Molecular cell*, 28(5):730–738.

[118] Hurley, J. H., Dean, A. M., Thorsness, P. E., Koshland, Jr, D. E., and Stroud, R. M. (1990). Regulation of isocitrate dehydrogenase by phosphorylation involves no long-range conformational change in the free enzyme. *The Journal of biological chemistry*, 265(7):3599–3602.

[119] Hutti, J. E., Jarrell, E. T., Chang, J. D., Abbott, D. W., Storz, P., Toker, A., Cantley, L. C., and Turk, B. E. (2004). A rapid method for determining protein kinase phosphorylation specificity. *Nature methods*, 1(1):27–29.

[120] Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic acids research*, 32(3):1037–1049.

[121] Iesmantavicius, V., Weinert, B. T., and Choudhary, C. (2014). Convergence of ubiquitylation and phosphorylation signaling in rapamycin-treated yeast cells. *Molecular & cellular proteomics: MCP*, 13(8):1979–1992.

[122] Iliuk, A. B., Martin, V. A., Alicie, B. M., Geahlen, R. L., and Tao, W. A. (2010). In-depth analyses of kinase-dependent tyrosine phosphoproteomes based on metal ion-functionalized soluble nanopolymers. *Molecular & cellular proteomics: MCP*, 9(10):2162–2172.

[123] Imamura, H., Sugiyama, N., Wakabayashi, M., and Ishihama, Y. (2014). Large-scale identification of phosphorylation sites for profiling protein kinase selectivity. *Journal of proteome research*, 13(7):3410–3419.

[124] Imamura, H., Wagih, O., Niinae, T., Sugiyama, N., Beltrao, P., and Ishihama, Y. (2017). Identifications of putative PKA substrates with quantitative phosphoproteomics and Primary-Sequence-Based scoring. *Journal of proteome research*, 16(4):1825–1830.

[125] Inazuka, F., Sugiyama, N., Tomita, M., Abe, T., Shioi, G., and Esumi, H. (2012). Muscle-specific knock-out of NUAK family SNF1-like kinase 1 (NUAK1) prevents high fat diet-induced glucose intolerance. *The Journal of biological chemistry*, 287(20):16379–16389.

[126] Ingley, E. and Hemmings, B. A. (2000). PKB/Akt interacts with inosine-5' monophosphate dehydrogenase through its pleckstrin homology domain. *FEBS letters*, 478(3):253–259.

[127] Invergo, B. M., Brochet, M., Yu, L., Choudhary, J., Beltrao, P., and Billker, O. (2017). Sub-minute phosphoregulation of cell cycle systems during plasmodium gamete formation. *Cell reports*, 21(7):2017–2029.

[128] Jeong, J. S., Jiang, L., Albino, E., Marrero, J., Rho, H. S., Hu, J., Hu, S., Vera, C., Bayron-Poueymiroy, D., Rivera-Pacheco, Z. A., Ramos, L., Torres-Castro, C., Qian, J., Bonaventura, J., Boeke, J. D., Yap, W. Y., Pino, I., Eichinger, D. J., Zhu, H., and Blackshaw, S. (2012). Rapid identification of monospecific monoclonal antibodies using a human proteome microarray.

[129] Johnson, J. R., Santos, S. D., Johnson, T., Pieper, U., Strumillo, M., Wagih, O., Sali, A., Krogan, N. J., and Beltrao, P. (2015). Prediction of functionally important Phospho-Regulatory events in xenopus laevis oocytes. *PLoS computational biology*, 11(8):e1004362.

[130] Johnson, L. N., Noble, M. E., and Owen, D. J. (1996). Active and inactive protein kinases: structural basis for regulation. *Cell*, 85(2):149–158.

[131] Johnson, S. A. and Hunter, T. (2005). Kinomics: methods for deciphering the kinome. *Nature methods*, 2(1):17–25.

[132] Jones, P. M. and George, A. M. (2004). The ABC transporter structure and mechanism: perspectives on recent research. *Cellular and molecular life sciences: CMLS*, 61(6):682–699.

[133] Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., Lander, E. S., and Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences of the United States of America*, 112(40):E5486–95.

[134] Kannan, N. and Neuwald, A. F. (2004). Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2α. *Protein science: a publication of the Protein Society*, 13(8):2059–2077.

[135] Kannan, N., Taylor, S. S., Zhai, Y., Craig Venter, J., and Manning, G. (2007). Structural and functional diversity of the microbial kinome.

[136] Kanshin, E., Giguère, S., Jing, C., Tyers, M., and Thibault, P. (2017). Machine learning of global phosphoproteomic profiles enables discrimination of direct versus indirect kinase substrates. *Molecular & cellular proteomics: MCP*, 16(5):786–798.

[137] Karnoub, A. E. and Weinberg, R. A. (2008). Ras oncogenes: split personalities. *Nature reviews. Molecular cell biology*, 9(7):517–531.

[138] Kato, M., Wynn, R. M., Chuang, J. L., Tso, S.-C., Machius, M., Li, J., and Chuang, D. T. (2008). Structural basis for inactivation of the human pyruvate dehydrogenase complex by phosphorylation: role of disordered phosphorylation loops. *Structure*, 16(12):1849–1859.

[139] Katoh, K., Rozewicki, J., and Yamada, K. D. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics*.

[140] Kemp, B. E. and Pearson, R. B. (1990). Protein kinase recognition sequence motifs. *Trends in biochemical sciences*, 15(9):342–346.

[141] Kettenbach, A. N., Schweppe, D. K., Faherty, B. K., Pechenick, D., Pletnev, A. A., and Gerber, S. A. (2011). Quantitative phosphoproteomics identifies substrates and functional modules of aurora and polo-like kinase activities in mitotic cells. *Science signaling*, 4(179):rs5.

[142] Khmelinskii, A., Meurer, M., Duishoev, N., Delhomme, N., and Knop, M. (2011). Seamless gene tagging by endonuclease-driven homologous recombination. *PloS one*, 6(8):e23794.

[143] Kim, J.-Y., Welsh, E. A., Oguz, U., Fang, B., Bai, Y., Kinose, F., Bronk, C., Remsing Rix, L. L., Beg, A. A., Rix, U., Eschrich, S. A., Koomen, J. M., and Haura, E. B. (2013). Dissection of TBK1 signaling via phosphoproteomics in lung cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, 110(30):12414–12419.

[144] Knebel, A., Morrice, N., and Cohen, P. (2001). A novel method to identify protein kinase substrates: eEF2 kinase is phosphorylated and inhibited by SAPK4/p38δ. *The EMBO journal*.

[145] Knighton, D. R., Zheng, J. H., Ten Eyck, L. F., Ashford, V. A., Xuong, N. H., Taylor, S. S., and Sowadski, J. M. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, 253(5018):407–414.

[146] Koenig, M. and Grabe, N. (2004). Highly specific prediction of phosphorylation sites in proteins. *Bioinformatics*, 20(18):3620–3627.

[147] Kölling, R. and Hollenberg, C. P. (1994). The ABC-transporter ste6 accumulates in the plasma membrane in a ubiquitinated form in endocytosis mutants. *The EMBO journal*, 13(14):3261–3271.

[148] Komander, D. and Rape, M. (2012). The ubiquitin code. *Annual review of biochemistry*, 81:203–229.

[149] Korotchkina, L. G. and Patel, M. S. (2001). Probing the mechanism of inactivation of human pyruvate dehydrogenase by phosphorylation of three sites. *The Journal of biological chemistry*, 276(8):5731–5738.

[150] Kosako, H. and Nagano, K. (2011). Quantitative phosphoproteomics strategies for understanding protein kinase-mediated signal transduction pathways. *Expert review of proteomics*, 8(1):81–94.

[151] Koscielny, G., Yaikhom, G., Iyer, V., Meehan, T. F., Morgan, H., Atienza-Herrero, J., Blake, A., Chen, C.-K., Easty, R., Di Fenza, A., Fiegel, T., Grifiths, M., Horne, A., Karp, N. A., Kurbatova, N., Mason, J. C., Matthews, P., Oakley, D. J., Qazi, A., Regnart, J., Retha, A., Santos, L. A., Sneddon, D. J., Warren, J., Westerberg, H., Wilson, R. J.,

Melvin, D. G., Smedley, D., Brown, S. D. M., Flicek, P., Skarnes, W. C., Mallon, A.-M., and Parkinson, H. (2014). The international mouse phenotyping consortium web portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic acids research*, 42(Database issue):D802–9.

[152] Kotrba, P., Inui, M., and Yukawa, H. (2001). The ptsi gene encoding enzyme I of the phosphotransferase system of corynebacterium glutamicum.

[153] Krahmer, N., Najafi, B., Schueder, F., Quagliarini, F., Steger, M., Seitz, S., Kasper, R., Salinas, F., Cox, J., Uhlenhaut, N. H., Walther, T. C., Jungmann, R., Zeigerer, A., Borner, G. H. H., and Mann, M. (2018). Organellar proteomics and Phospho-Proteomics reveal subcellular reorganization in Diet-Induced hepatic steatosis. *Developmental cell*, 47(2):205–221.e7.

[154] Kurosaki, T. (2002). Regulation of b-cell signal transduction by adaptor proteins. *Nature reviews. Immunology*, 2(5):354–363.

[155] Kweon, H. K. and Håkansson, K. (2006). Selective zirconium dioxide-based enrichment of phosphorylated peptides for mass spectrometric analysis. *Analytical chemistry*, 78(6):1743–1749.

[156] la Cour, T., Kiemer, L., Mølgaard, A., Gupta, R., Skriver, K., and Brunak, S. (2004). Analysis and prediction of leucine-rich nuclear export signals. *Protein engineering, design & selection: PEDS*, 17(6):527–536.

[157] Lachaise, F., Martin, G., Drougard, C., Perl, A., Vuillaume, M., Wegnez, M., Sarasin, A., and Daya-Grosjean, L. (2001). Relationship between posttranslational modification of transaldolase and catalase deficiency in UV-sensitive repair-deficient xeroderma pigmentosum fibroblasts and SV40-transformed human cells. *Free radical biology & medicine*, 30(12):1365–1373.

[158] Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., Kattman, B. L., and Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1):D1062–D1067.

[159] Landry, C. R., Levy, E. D., and Michnick, S. W. (2009). Weak functional constraints on phosphoproteomes. *Trends in genetics: TIG*, 25(5):193–197.

[160] Laporte, D. C., Stueland, C. S., and Ikeda, T. P. (1989). Isocitrate dehydrogenase kinase/phosphatase. *Biochimie*, 71(9-10):1051–1057.

[161] Lappalainen, I., Thusberg, J., Shen, B., and Vihinen, M. (2008). Genome wide analysis of pathogenic SH2 domain mutations. *Proteins*, 72(2):779–792.

[162] Larsen, M. R., Thingholm, T. E., Jensen, O. N., Roepstorff, P., and Jørgensen, T. J. D. (2005). Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns.

[163] Leberer, E., Thomas, D. Y., and Whiteway, M. (1997). Pheromone signalling and polarized morphogenesis in yeast. *Current opinion in genetics & development*, 7(1):59–66.

[164] Lee, E. E., Ma, J., Sacharidou, A., Mi, W., Salato, V. K., Nguyen, N., Jiang, Y., Pascual, J. M., North, P. E., Shaul, P. W., Mettlen, M., and Wang, R. C. (2015). A protein kinase C phosphorylation motif in GLUT1 affects glucose transport and is mutated in GLUT1 deficiency syndrome. *Molecular cell*, 58(5):845–853.

[165] Lemeer, S., Bluwstein, A., Wu, Z., Leberfinger, J., Müller, K., Kramer, K., and Kuster, B. (2012). Phosphotyrosine mediated protein interactions of the discoidin domain receptor 1. *Journal of proteomics*, 75(12):3465–3477.

[166] Lemeer, S. and Heck, A. J. R. (2009). The phosphoproteomics data explosion.

[167] Lesaicherre, M. L., Uttamchandani, M., Chen, G. Y. J., and Yao, S. Q. (2002). Antibody-based fluorescence detection of kinase activity on a peptide array. *Bioorganic & medicinal chemistry letters*, 12(16):2085–2088.

[168] Li, T., Li, F., and Zhang, X. (2008). Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins: Structure, Function, and Bioinformatics*, 70(2):404–414.

[169] Linding, R., Jensen, L. J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M. B., and Pawson, T. (2008). NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic acids research*, 36(Database issue):D695–9.

[170] Lonard, D. M. and O'malley, B. W. (2007). Nuclear receptor coregulators: judges, juries, and executioners of cellular regulation. *Molecular cell*, 27(5):691–700.

[171] Lundby, A., Secher, A., Lage, K., Nordsborg, N. B., Dmytriyev, A., Lundby, C., and Olsen, J. V. (2012). Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues.

[172] Lv, D.-W., Subburaj, S., Cao, M., Yan, X., Li, X., Appels, R., Sun, D.-F., Ma, W., and Yan, Y.-M. (2014). Proteome and phosphoproteome characterization reveals new response and defense mechanisms of brachypodium distachyon leaves under salt stress. *Molecular & cellular proteomics: MCP*, 13(2):632–652.

[173] Lv, X.-B., Wu, W., Tang, X., Wu, Y., Zhu, Y., Liu, Y., Cui, X., Chu, J., Hu, P., Li, J., Guo, Q., Cai, Z., Wu, J., Hu, K., and Ouyang, N. (2015). Regulation of SOX10 stability via ubiquitination-mediated degradation by Fbxw7α modulates melanoma cell migration. *Oncotarget*, 6(34):36370–36382.

[174] Ma, Y.-M., Tao, R.-Y., Liu, Q., Li, J., Tian, J.-Y., Zhang, X.-L., Xiao, Z.-Y., and Ye, F. (2011). PTP1B inhibitor improves both insulin resistance and lipid abnormalities in vivo and in vitro. *Molecular and cellular biochemistry*, 357(1-2):65–72.

[175] Mah, A. S., Elia, A. E. H., Devgan, G., Ptacek, J., Schutkowski, M., Snyder, M., Yaffe, M. B., and Deshaies, R. J. (2005). Substrate specificity analysis of protein kinase complex Dbf2-Mob1 by peptide library and proteome array screening. *BMC biochemistry*, 6:22.

[176] Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., and Church, G. M. (2013a). RNA-guided human genome engineering via cas9. *Science*, 339(6121):823–826.

[177] Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., and Church, G. M. (2013b). RNA-guided human genome engineering via cas9. *Science*, 339(6121):823–826.

[178] Mann, M., Kulak, N. A., Nagaraj, N., and Cox, J. (2013). The coming age of complete, accurate, and ubiquitous proteomes. *Molecular cell*, 49(4):583–590.

[179] Mann, M., Ong, S. E., Grønborg, M., Steen, H., Jensen, O. N., and Pandey, A. (2002). Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends in biotechnology*, 20(6):261–268.

[180] Manning, B. D. and Cantley, L. C. (2002). Hitting the target: emerging technologies in the search for kinase substrates. *Science's STKE: signal transduction knowledge environment*, 2002(162):e49.

[181] Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934.

[182] Markevich, N. I., Hoek, J. B., and Kholodenko, B. N. (2004). Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *The Journal of cell biology*, 164(3):353–359.

[183] Marx, H., Lemeer, S., Schliep, J. E., Matheron, L., Mohammed, S., Cox, J., Mann, M., Heck, A. J. R., and Kuster, B. (2013). A large synthetic peptide and phosphopeptide reference library for mass spectrometry–based proteomics.

[184] Matic, K., Eninger, T., Bardoni, B., Davidovic, L., and Macek, B. (2014). Quantitative phosphoproteomics of murine Fmr1-KO cell lines provides new insights into FMRP-Dependent signal transduction mechanisms.

[185] Matsuoka, S., Ballif, B. A., Smogorzewska, A., McDonald, 3rd, E. R., Hurov, K. E., Luo, J., Bakalarski, C. E., Zhao, Z., Solimini, N., Lerenthal, Y., Shiloh, Y., Gygi, S. P., and Elledge, S. J. (2007). ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science*, 316(5828):1160–1166.

[186] Mattsson, P. T., Lappalainen, I., Bäckesjö, C. M., and others (2000). Six x-linked agammaglobulinemia-causing missense mutations in the src homology 2 domain of bruton's tyrosine kinase: phosphotyrosine-binding and circular .... *The Journal of*.

[187] McClatchy, D. B., Dong, M.-Q., Wu, C. C., Venable, J. D., and Yates, J. R. (2007). 15N metabolic labeling of mammalian tissue with slow protein turnover.

[188] McDonald, B. M., Wydro, M. M., Lightowlers, R. N., and Lakey, J. H. (2009). Probing the orientation of yeast VDAC1 in vivo. *FEBS letters*, 583(4):739–742.

[189] Meissner, F., Scheltema, R. A., Mollenkopf, H.-J., and Mann, M. (2013). Direct proteomic quantification of the secretome of activated immune cells. *Science*, 340(6131):475–478.

[190] Merid, S. K., Goranskaya, D., and Alexeyenko, A. (2014). Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. *BMC bioinformatics*, 15:308.

[191] Meurer, A., Smith, C. P., Paprocki, M., Čertík, O., Kirpichev, S. B., Rocklin, M., Kumar, A., Ivanov, S., Moore, J. K., Singh, S., Rathnayake, T., Vig, S., Granger, B. E., Muller, R. P., Bonazzi, F., Gupta, H., Vats, S., Johansson, F., Pedregosa, F., Curry, M. J., Terrel, A. R., Roučka, Š., Saboo, A., Fernando, I., Kulal, S., Cimrman, R., and Scopatz, A. (2017). SymPy: symbolic computing in python. *PeerJ Computer Science*, 3:e103.

[192] Meyerovitch, J., Backer, J. M., and Kahn, C. R. (1989). Hepatic phosphotyrosine phosphatase activity and its alterations in diabetic rats. *The Journal of clinical investigation*, 84(3):976–983.

[193] Mijn, J. C. v. d., van der Mijn, J. C., Labots, M., Piersma, S. R., Pham, T. V., Knol, J. C., Broxterman, H. J., Verheul, H. M., and Jiménez, C. R. (2015). Evaluation of different phospho-tyrosine antibodies for label-free phosphoproteomics.

[194] Miller, M. L., Jensen, L. J., Diella, F., Jørgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S. A., Bordeaux, J., Sicheritz-Ponten, T., Olhovsky, M., Pasculescu, A., Alexander, J., Knapp, S., Blom, N., Bork, P., Li, S., Cesareni, G., Pawson, T., Turk, B. E., Yaffe, M. B., Brunak, S., and Linding, R. (2008). Linear motif atlas for phosphorylation-dependent signaling. *Science signaling*, 1(35):ra2.

[195] Mocan, I., Georgescauld, F., Gonin, P., Thoraval, D., Cervoni, L., Giartosio, A., Dabernat-Arnaud, S., Crouzet, M., Lacombe, M.-L., and Lascu, I. (2007). Protein phosphorylation corrects the folding defect of the neuroblastoma (S120G) mutant of human nucleoside diphosphate kinase A/Nm23-H1. *Biochemical Journal*, 403(1):149–156.

[196] Moebitz, H. and Fabbro, D. (2012). Conformational bias: a key concept for protein kinase inhibition. *Eur Pharm Rev*, 17:41–51.

[197] Mok, J., Kim, P. M., Lam, H. Y. K., Piccirillo, S., Zhou, X., Jeschke, G. R., Sheridan, D. L., Parker, S. A., Desai, V., Jwa, M., Cameroni, E., Niu, H., Good, M., Remenyi, A., Ma, J.-L. N., Sheu, Y.-J., Sassi, H. E., Sopko, R., Chan, C. S. M., De Virgilio, C., Hollingsworth, N. M., Lim, W. A., Stern, D. F., Stillman, B., Andrews, B. J., Gerstein, M. B., Snyder, M., and Turk, B. E. (2010). Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Science signaling*, 3(109):ra12.

[198] Moorhead, G. B. G., Trinkle-Mulcahy, L., and Ulke-Lemée, A. (2007). Emerging roles of nuclear protein phosphatases. *Nature reviews. Molecular cell biology*, 8(3):234–244.

[199] Mosca, R., Céol, A., and Aloy, P. (2012). Interactome3D: adding structural details to protein networks. *Nature methods*, 10(1):47–53.

[200] Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic acids research*, 42(Database issue):D374–9.

[201] Mott, H. R. and Owen, D. (2015). Structures of ras superfamily effector complexes: What have we learnt in two decades? *Critical reviews in biochemistry and molecular biology*, 50(2):85–133.

[202] Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology*, 7:548.

[203] Nakagawa, H., Toyoda, Y., Wakabayashi-Nakao, K., Tamaki, H., Osumi, M., and Ishikawa, T. (2011). Ubiquitin-mediated proteasomal degradation of ABC transporters: a new aspect of genetic polymorphisms and clinical impacts. *Journal of pharmaceutical sciences*, 100(9):3602–3619.

[204] Nasu, J., Murakami, K., Miyagawa, S., Yamashita, R., Ichimura, T., Wakita, T., Hotta, H., Miyamura, T., Suzuki, T., Satoh, T., and Shoji, I. (2010). E6AP ubiquitin ligase mediates ubiquitin-dependent degradation of peroxiredoxin 1. *Journal of cellular biochemistry*, 111(3):676–685.

[205] Needham, E. J., Parker, B. L., Burykin, T., James, D. E., and Humphrey, S. J. (2019). Illuminating the dark phosphoproteome. *Science signaling*, 12(565).

[206] Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814.

[207] Nishi, H., Hashimoto, K., and Panchenko, A. R. (2011). Phosphorylation in protein-protein binding: effect on stability and function. *Structure*, 19(12):1807–1815.

[208] Nolen, B. J., Littlefield, R. S., and Pollard, T. D. (2004). Crystal structures of actin-related protein 2/3 complex with bound ATP or ADP. *Proceedings of the National Academy of Sciences of the United States of America*, 101(44):15627–15632.

[209] Nussinov, R., Tsai, C.-J., and Jang, H. (2019). Protein ensembles link genotype to phenotype. *PLoS computational biology*, 15(6):e1006648.

[210] Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic acids research*, 31(13):3635–3641.

[211] Ochoa, D., Jarnuczak, A. F., Gehre, M., Soucheray, M., and others (2019). The functional landscape of the human phosphoproteome. *BioRxiv*.

[212] Olsen, J. V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M. L., Jensen, L. J., Gnad, F., Cox, J., Jensen, T. S., Nigg, E. A., Brunak, S., and Mann, M. (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Science signaling*, 3(104):ra3.

[213] Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.

[214] Oruganty, K., Talevich, E. E., Neuwald, A. F., and Kannan, N. (2016). Identification and classification of small molecule kinases: insights into substrate recognition and specificity. *BMC evolutionary biology*, 16:7.

[215] Paik, Y.-K., Jeong, S.-K., Omenn, G. S., Uhlen, M., Hanash, S., Cho, S. Y., Lee, H.-J., Na, K., Choi, E.-Y., Yan, F., Zhang, F., Zhang, Y., Snyder, M., Cheng, Y., Chen, R., Marko-Varga, G., Deutsch, E. W., Kim, H., Kwon, J.-Y., Aebersold, R., Bairoch, A., Taylor, A. D., Kim, K. Y., Lee, E.-Y., Hochstrasser, D., Legrain, P., and Hancock, W. S. (2012). The Chromosome-Centric human proteome project for cataloging proteins encoded in the genome. *Nature biotechnology*, 30(3):221–223.

[216] Palma, S. D., Di Palma, S., Zoumaro-Djayoon, A., Peng, M., Post, H., Preisinger, C., Munoz, J., and Heck, A. J. R. (2013). Finding the same needles in the haystack? a comparison of phosphotyrosine peptides enriched by immuno-affinity precipitation and metal-based affinity chromatography.

[217] Paradis, S. and Ruvkun, G. (1998). Caenorhabditis elegans Akt/PKB transduces insulin receptor-like signals from AGE-1 PI3 kinase to the DAF-16 transcription factor. *Genes & development*, 12(16):2488–2498.

[218] Patel, M. S., Nemeria, N. S., Furey, W., and Jordan, F. (2014). The pyruvate dehydrogenase complexes: structure-based function and regulation. *The Journal of biological chemistry*, 289(24):16615–16623.

[219] Pawson, T. and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science*, 300(5618):445–452.

[220] Pearson, R. B. and Kemp, B. E. (1991). [3] protein kinase phosphorylation site sequences and consensus specificity motifs: Tabulations. In *Methods in Enzymology*, volume Volume 200, pages 62–81. Academic Press.

[221] Peña, C., Schütz, S., Fischer, U., Chang, Y., and Panse, V. G. (2016). Prefabrication of a ribosomal protein subcomplex essential for eukaryotic ribosome formation. *eLife*, 5.

[222] Pieper, U., Webb, B. M., Dong, G. Q., Schneidman-Duhovny, D., Fan, H., Kim, S. J., Khuri, N., Spill, Y. G., Weinkam, P., Hammel, M., Tainer, J. A., Nilges, M., and Sali, A. (2014). ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic acids research*, 42(Database issue):D336–46.

[223] Pillay, T. S. (2009). A fisherman's tale: Phage display as a discovery tool. *Discovery medicine*, 4(23):315–318.

[224] Pinna, L. A. and Ruzzene, M. (1996a). How do protein kinases recognize their substrates? *Biochimica et biophysica acta*, 1314(3):191–225.

[225] Pinna, L. A. and Ruzzene, M. (1996b). How do protein kinases recognize their substrates? *Biochimica et biophysica acta*, 1314(3):191–225.

[226] Plewczynski, D., Jaroszewski, L., Godzik, A., Kloczkowski, A., and Rychlewski, L. (2005). Molecular modeling of phosphorylation sites in proteins using a database of local structure segments. *Journal of molecular modeling*, 11(6):431–438.

[227] Posewitz, M. C. and Tempst, P. (1999). Immobilized gallium(III) affinity chromatography of phosphopeptides. *Analytical chemistry*, 71(14):2883–2892.

[228] Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., and Others (2009). Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772.

[229] Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(Database issue):D61–5.

[230] Pylayeva-Gupta, Y., Grabocka, E., and Bar-Sagi, D. (2011). RAS oncogenes: weaving a tumorigenic web. *Nature reviews. Cancer*, 11(11):761–774.

[231] Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature protocols*, 2(8):1896–1906.

[232] Rauniyar, N., Lavalleé-Adam, M., Yates, J. R., and others (2017). Quantitative analysis of the proteome response to the histone deacetylase inhibitor (HDACi) vorinostat in Niemann-Pick type C1 disease. *& Cellular Proteomics*.

[233] Ravichandran, L. V., Chen, H., Li, Y., and Quon, M. J. (2001). Phosphorylation of PTP1B at ser(50) by akt impairs its ability to dephosphorylate the insulin receptor. *Molecular endocrinology*, 15(10):1768–1780.

[234] Rhoads, T. W., Prasad, A., Kwiecien, N. W., Merrill, A. E., Zawack, K., Westphall, M. S., Schroeder, F. C., Kimble, J., and Coon, J. J. (2015). NeuCode labeling in nematodes: Proteomic and phosphoproteomic impact of ascaroside treatment in caenorhabditis elegans. *Molecular & cellular proteomics: MCP*, 14(11):2922–2935.

[235] Rigbolt, K. T. G., Prokhorova, T. A., Akimov, V., Henningsen, J., Johansen, P. T., Kratchmarova, I., Kassem, M., Mann, M., Olsen, J. V., and Blagoev, B. (2011). System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. *Science signaling*, 4(164):rs3.

[236] Rikova, K., Guo, A., Zeng, Q., Possemato, A., Yu, J., Haack, H., Nardone, J., Lee, K., Reeves, C., Li, Y., Hu, Y., Tan, Z., Stokes, M., Sullivan, L., Mitchell, J., Wetzel, R., Macneill, J., Ren, J. M., Yuan, J., Bakalarski, C. E., Villen, J., Kornhauser, J. M., Smith, B., Li, D., Zhou, X., Gygi, S. P., Gu, T.-L., Polakiewicz, R. D., Rush, J., and Comb, M. J. (2007). Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell*, 131(6):1190–1203.

[237] Ringrose, J. H., van den Toorn, H. W. P., Eitel, M., Post, H., Neerincx, P., Schierwater, B., Altelaar, A. F. M., and Heck, A. J. R. (2013). Deep proteome profiling of trichoplax adhaerens reveals remarkable features at the origin of metazoan multicellularity. *Nature communications*, 4:1408.

[238] Robinson, M. R., Taliaferro, J. M., Dalby, K. N., and Brodbelt, J. S. (2016). 193 nm ultraviolet photodissociation mass spectrometry for phosphopeptide characterization in the positive and negative ion modes. *Journal of proteome research*, 15(8):2739–2748.

[239] Robinson, V. L., Buckler, D. R., and Stock, A. M. (2000). A tale of two components: a novel kinase and a regulatory switch. *Nature structural biology*, 7(8):626–633.

[240] Rogerson, D. T., Sachdeva, A., Wang, K., Haq, T., Kazlauskaite, A., Hancock, S. M., Huguenin-Dezot, N., Muqit, M. M. K., Fry, A. M., Bayliss, R., and Chin, J. W. (2015). Efficient genetic encoding of phosphoserine and its nonhydrolyzable analog. *Nature chemical biology*, 11(7):496–503.

[241] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004). Multiplexed protein quantitation insaccharomyces cerevisiaeusing amine-reactive isobaric tagging reagents.

[242] Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., Liang, X., Nahnsen, S., Nilse, L., Pfeuffer, J., Rosenberger, G., Rurik, M., Schmitt, U., Veit, J., Walzer, M., Wojnar, D., Wolski, W. E., Schilling, O., Choudhary, J. S., Malmström, L., Aebersold, R., Reinert, K., and Kohlbacher, O. (2016). OpenMS: a flexible open-source software platform for mass spectrometry data analysis.

[243] Sacco, F., Perfetto, L., Castagnoli, L., and Cesareni, G. (2012). The human phosphatase interactome: An intricate family portrait. *FEBS letters*, 586(17):2732–2739.

[244] Saunders, N. F. W., Brinkworth, R. I., Huber, T., Kemp, B. E., and Kobe, B. (2008). Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC bioinformatics*, 9(1):1–11.

[245] Savitski, M. M., Lemeer, S., Boesche, M., Lang, M., Mathieson, T., Bantscheff, M., and Kuster, B. (2011). Confident phosphorylation site localization using the mascot delta score.

[246] Schilling, B., Rardin, M. J., MacLean, B. X., Zawadzka, A. M., Frewen, B. E., Cusack, M. P., Sorensen, D. J., Bereman, M. S., Jing, E., Wu, C. C., Verdin, E., Ronald Kahn, C., MacCoss, M. J., and Gibson, B. W. (2012). Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline.

[247] Schlüter, H., Apweiler, R., Holzhütter, H.-G., and Jungblut, P. R. (2009). Finding one's way in proteomics: a protein species nomenclature. *Chemistry Central journal*, 3:11.

[248] Schmid-Burgk, J. L., Chauhan, D., Schmidt, T., Ebert, T. S., Reinhardt, J., Endl, E., and Hornung, V. (2016). A genome-wide CRISPR (clustered regularly interspaced short palindromic repeats) screen identifies NEK7 as an essential component of NLRP3 inflammasome activation.

[249] Schulze, W. X. and Mann, M. (2004). A novel proteomic screen for peptide-protein interactions. *The Journal of biological chemistry*, 279(11):10756–10764.

[250] Scott, J. D. and Pawson, T. (2009). Cell signaling in space and time: where proteins come together and when they're apart. *Science*, 326(5957):1220–1224.

[251] Sekiguchi, T., Kurihara, Y., and Fukumura, J. (2007). Phosphorylation of threonine 204 of DEAD-box RNA helicase DDX3 by cyclin b/cdc2 in vitro. *Biochemical and biophysical research communications*, 356(3):668–673.

[252] Seringhaus, M., Paccanaro, A., Borneman, A., Snyder, M., and Gerstein, M. (2006). Predicting essential genes in fungal genomes. *Genome research*, 16(9):1126–1135.

[253] Sha, D., Chin, L.-S., and Li, L. (2010). Phosphorylation of parkin by parkinson disease-linked kinase PINK1 activates parkin E3 ligase function and NF-kappaB signaling. *Human molecular genetics*, 19(2):352–363.

[254] Shah, K., Liu, Y., Deirmengian, C., and Shokat, K. M. (1997). Engineering unnatural nucleotide specificity for rous sarcoma virus tyrosine kinase to uniquely label its direct substrates. *Proceedings of the National Academy of Sciences of the United States of America*, 94(8):3565–3570.

[255] Shankar, S., Pitchiaya, S., Malik, R., Kothari, V., Hosono, Y., Yocum, A. K., Gundlapalli, H., White, Y., Firestone, A., Cao, X., Dhanasekaran, S. M., Stuckey, J. A., Bollag, G., Shannon, K., Walter, N. G., Kumar-Sinha, C., and Chinnaiyan, A. M. (2016). KRAS engages AGO2 to enhance cellular transformation. *Cell reports*, 14(6):1448–1461.

[256] Sharma, K., D'Souza, R. C. J., Tyanova, S., Schaab, C., Wiśniewski, J. R., Cox, J., and Mann, M. (2014). Ultradeep human phosphoproteome reveals a distinct regulatory nature of tyr and Ser/Thr-based signaling. *Cell reports*, 8(5):1583–1594.

[257] Sharma, K., Weber, C., Bairlein, M., Greff, Z., Kéri, G., Cox, J., Olsen, J. V., and Daub, H. (2009). Proteomics strategy for quantitative protein interaction profiling in cell extracts. *Nature methods*, 6(10):741–744.

[258] Shepherd, P. R., Withers, D. J., and Siddle, K. (1998). Phosphoinositide 3-kinase: the key switch mechanism in insulin signalling. *Biochemical Journal*, 333 (Pt 3):471–490.

[259] Shin, D., Na, W., Lee, J.-H., Kim, G., Baek, J., Park, S. H., Choi, C. Y., and Lee, S. (2017). Site-specific monoubiquitination downregulates rab5 by disrupting effector binding and guanine nucleotide conversion. *eLife*, 6.

[260] Shinde, S. R. and Maddika, S. (2016). PTEN modulates EGFR late endocytic trafficking and degradation by dephosphorylating rab7. *Nature communications*, 7:10689.

[261] Shirakata, Y., Ishii, K., Yagita, H., Okumura, K., Taniguchi, M., and Takemori, T. (1999). Distinct subcellular localization and substrate specificity of extracellular signal-regulated kinase in B cells upon stimulation with IgM and CD40. *Journal of immunology*, 163(12):6589–6597.

[262] Sickmann, A. and Meyer, H. E. (2001). Phosphoamino acid analysis. *Proteomics*, 1(2):200–206.

[263] Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*, 40(Web Server issue):W452–7.

[264] Sims, 3rd, R. J. and Reinberg, D. (2008). Is there a code embedded in proteins that is based on post-translational modifications? *Nature reviews. Molecular cell biology*, 9:815.

[265] Skowyra, D., Craig, K. L., Tyers, M., Elledge, S. J., and Harper, J. W. (1997). F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell*, 91(2):209–219.

[266] Smith, L. M., The Consortium for Top Down Proteomics, and Kelleher, N. L. (2013). Proteoform: a single term describing protein complexity.

[267] Song, C., Ye, M., Liu, Z., Cheng, H., Jiang, X., Han, G., Songyang, Z., Tan, Y., Wang, H., Ren, J., Xue, Y., and Zou, H. (2012). Systematic analysis of protein phosphorylation networks from phosphoproteomic data. *Molecular & cellular proteomics: MCP*, 11(10):1070–1083.

[268] Songyang, Z., Carraway, 3rd, K. L., Eck, M. J., Harrison, S. C., Feldman, R. A., Mohammadi, M., Schlessinger, J., Hubbard, S. R., Smith, D. P., and Eng, C. (1995). Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. *Nature*, 373(6514):536–539.

[269] Songyang, Z., Lu, K. P., Kwon, Y. T., Tsai, L. H., and others (1996). A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase .... *and cellular biology*.

[270] Šoštarić, N., O'Reilly, F. J., Giansanti, P., Heck, A. J. R., Gavin, A.-C., and van Noort, V. (2018). Effects of acetylation and phosphorylation on subunit interactions in three large eukaryotic complexes. *Molecular & cellular proteomics: MCP*, 17(12):2387–2401.

[271] Staudinger, J., Zhou, J., Burgess, R., Elledge, S. J., and Olson, E. N. (1995). PICK1: a perinuclear binding protein and substrate for protein kinase C isolated by the yeast two-hybrid system. *The Journal of cell biology*, 128(3):263–271.

[272] Steger, M., Diez, F., Dhekne, H. S., Lis, P., Nirujogi, R. S., Karayel, O., Tonelli, F., Martinez, T. N., Lorentzen, E., Pfeffer, S. R., Alessi, D. R., and Mann, M. (2017). Systematic proteomic analysis of LRRK2-mediated rab GTPase phosphorylation establishes a connection to ciliogenesis. *eLife*, 6.

[273] Strickfaden, S. C., Winters, M. J., Ben-Ari, G., Lamson, R. E., Tyers, M., and Pryciak, P. M. (2007). A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. *Cell*, 128(3):519–531.

[274] Strumillo, M. J., Oplová, M., Viéitez, C., Ochoa, D., Shahraz, M., Busby, B. P., Sopko, R., Studer, R. A., Perrimon, N., Panse, V. G., and Beltrao, P. (2019). Conserved phosphorylation hotspots in eukaryotic protein domain families. *Nature communications*, 10(1):1977.

[275] Studer, R. A., Rodriguez-Mias, R. A., Haas, K. M., Hsu, J. I., Viéitez, C., Solé, C., Swaney, D. L., Stanford, L. B., Liachko, I., Böttcher, R., Dunham, M. J., de Nadal, E., Posas, F., Beltrao, P., and Villén, J. (2016a). Evolution of protein phosphorylation across 18 fungal species. *Science*, 354(6309):229–232.

[276] Studer, R. A., Rodriguez-Mias, R. A., Haas, K. M., Hsu, J. I., Viéitez, C., Solé, C., Swaney, D. L., Stanford, L. B., Liachko, I., Böttcher, R., Dunham, M. J., de Nadal, E., Posas, F., Beltrao, P., and Villén, J. (2016b). Evolution of protein phosphorylation across 18 fungal species. *Science*, 354(6309):229–232.

[277] Su, M.-G. and Lee, T.-Y. (2013). Incorporating substrate sequence motifs and spatial amino acid composition to identify kinase-specific phosphorylation sites on protein three-dimensional structures. *BMC bioinformatics*, 14 Suppl 16:S2.

[278] Taylor, J. A. and Johnson, R. S. (1997). Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry: RCM*, 11(9):1067–1075.

[279] Taylor, S. S. and Kornev, A. P. (2011). Protein kinases: evolution of dynamic regulatory proteins. *Trends in biochemical sciences*, 36(2):65–77.

[280] Thomason, P. and Kay, R. (2000). Eukaryotic signal transduction via histidine-aspartate phosphorelay. *Journal of cell science*, 113 (Pt 18):3141–3150.

[281] Tien, A.-C., Lin, M.-H., Su, L.-J., Hong, Y.-R., Cheng, T.-S., Lee, Y.-C. G., Lin, W.-J., Still, I. H., and Huang, C.-Y. F. (2004). Identification of the substrates and interaction proteins of aurora kinases from a Protein-Protein interaction model.

[282] Tomasello, M. F., Guarino, F., Reina, S., Messina, A., and De Pinto, V. (2013). The voltage-dependent anion selective channel 1 (VDAC1) topography in the mitochondrial outer membrane as detected in intact cell. *PloS one*, 8(12):e81522.

[283] Torres, M. P., Dewhurst, H., and Sundararaman, N. (2016). Proteome-wide structural analysis of PTM hotspots reveals regulatory elements predicted to impact biological function and disease. *Molecular & cellular proteomics: MCP*, 15(11):3513–3528.

[284] Toulmay, A. and Schneiter, R. (2006). A two-step method for the introduction of single or multiple defined point mutations into the genome of saccharomyces cerevisiae. *Yeast*, 23(11):825–831.

[285] Tran, J. C., Zamdborg, L., Ahlf, D. R., Lee, J. E., Catherman, A. D., Durbin, K. R., Tipton, J. D., Vellaichamy, A., Kellie, J. F., Li, M., Wu, C., Sweet, S. M. M., Early, B. P., Siuti, N., LeDuc, R. D., Compton, P. D., Thomas, P. M., and Kelleher, N. L. (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*, 480(7376):254–258.

[286] Treeck, M., Sanders, J. L., Elias, J. E., and Boothroyd, J. C. (2011). The phosphoproteomes of plasmodium falciparum and toxoplasma gondii reveal unusual adaptations within and beyond the parasites' boundaries. *Cell host & microbe*, 10(4):410–419.

[287] Troiani, S., Uggeri, M., Moll, J., Isacchi, A., Kalisz, H. M., Rusconi, L., and Valsasina, B. (2005). Searching for biomarkers of Aurora-A kinase activity: identification of in vitro substrates through a modified KESTREL approach. *Journal of proteome research*, 4(4):1296–1303.

[288] Trost, B. and Kusalik, A. (2011). Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics*, 27(21):2927–2935.

[289] Ubersax, J. A. and Ferrell, Jr, J. E. (2007). Mechanisms of specificity in protein phosphorylation. *Nature reviews. Molecular cell biology*, 8(7):530–541.

[290] UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515.

[291] UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic acids research*, 46(5):2699.

[292] Vadlamudi, R. K., Li, F., Adam, L., Nguyen, D., Ohta, Y., Stossel, T. P., and Kumar, R. (2002). Filamin is essential in actin cytoskeletal assembly mediated by p21-activated kinase 1. *Nature cell biology*, 4(9):681–690.

[293] Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.-A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H.-J., Albar, J. P., Martinez-Bartolomé, S., Apweiler, R., Omenn, G. S., Martens, L., Jones, A. R., and Hermjakob, H. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology*, 32(3):223–226.

[294] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, Jr, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339(6127):1546–1558.

[295] Wagih, O., Sugiyama, N., Ishihama, Y., and Beltrao, P. (2016). Uncovering Phosphorylation-Based specificities through functional interaction networks. *Molecular & cellular proteomics: MCP*, 15(1):236–245.

[296] Walsh, C. T., Garneau-Tsodikova, S., and Gatto, Jr, G. J. (2005). Protein posttranslational modifications: the chemistry of proteome diversifications. *Angewandte Chemie*, 44(45):7342–7372.

[297] Wang, H., Gau, B., Slade, W. O., Juergens, M., Li, P., and Hicks, L. M. (2014). The global phosphoproteome of chlamydomonas reinhardtii reveals complex organellar phosphorylation in the flagella and thylakoid membrane. *Molecular & cellular proteomics: MCP*, 13(9):2337–2353.

[298] Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology*, 337(3):635–645.

[299] Weisberg, E., Manley, P., Mestan, J., Cowan-Jacob, S., Ray, A., and Griffin, J. D. (2006). AMN107 (nilotinib): a novel and selective inhibitor of BCR-ABL. *British journal of cancer*, 94(12):1765–1769.

[300] Witze, E. S., Old, W. M., Resing, K. A., and Ahn, N. G. (2007). Mapping protein post-translational modifications with mass spectrometry. *Nature methods*, 4(10):798–806.

[301] Wolschin, F., Wienkoop, S., and Weckwerth, W. (2005). Enrichment of phosphorylated proteins and peptides from complex mixtures using metal oxide/hydroxide affinity chromatography (MOAC). *Proteomics*, 5(17):4389–4397.

[302] Wu, P., Nielsen, T. E., and Clausen, M. H. (2016). Small-molecule kinase inhibitors: an analysis of FDA-approved drugs. *Drug discovery today*, 21(1):5–10.

[303] Wu, R., Haas, W., Dephoure, N., Huttlin, E. L., Zhai, B., Sowa, M. E., and Gygi, S. P. (2011). A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nature methods*, 8(8):677–683.

[304] Xiao, Q., Miao, B., Bi, J., Wang, Z., and Li, Y. (2016). Prioritizing functional phosphorylation sites based on multiple feature integration. *Scientific reports*, 6:24735.

[305] Xu, D., Wang, H., and You, G. (2016). Posttranslational regulation of organic anion transporters by ubiquitination: Known and novel. *Medicinal research reviews*, 36(5):964–979.

[306] Yadav, B., Pemovska, T., Szwajda, A., Kulesskiy, E., Kontro, M., Karjalainen, R., Majumder, M. M., Malani, D., Murumägi, A., Knowles, J., Porkka, K., Heckman, C., Kallioniemi, O., Wennerberg, K., and Aittokallio, T. (2014). Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Scientific reports*, 4:5193.

[307] Yan, H., Jahanshahi, M., Horvath, E. A., Liu, H.-Y., and Pfleger, C. M. (2010). Rabex-5 ubiquitin ligase activity restricts ras signaling to establish pathway homeostasis in drosophila. *Current biology: CB*, 20(15):1378–1382.

[308] Yang, G., Murashige, D. S., Humphrey, S. J., and James, D. E. (2015). A positive feedback loop between akt and mTORC2 via SIN1 phosphorylation.

[309] Yang, P., Humphrey, S. J., James, D. E., Yang, Y. H., and Jothi, R. (2016). Positive-unlabeled ensemble learning for kinase substrate prediction from dynamic phosphoproteomics data. *Bioinformatics*, 32(2):252–259.

[310] Yang, W.-L., Wu, C.-Y., Wu, J., and Lin, H.-K. (2010). Regulation of akt signaling activation by ubiquitination. *Cell cycle*, 9(3):487–497.

[311] Yang, X., Hubbard, E. J., and Carlson, M. (1992). A protein kinase substrate identified by the two-hybrid system. *Science*, 257(5070):680–682.

[312] Yao, Q., Ge, H., Wu, S., Zhang, N., Chen, W., Xu, C., Gao, J., Thelen, J. J., and Xu, D. (2014). P³DB 3.0: From plant phosphorylation sites to protein networks. *Nucleic acids research*, 42(Database issue):D1206–13.

[313] Yao, X., Freas, A., Ramirez, J., Demirev, P. A., and Fenselau, C. (2001). Proteolytic18O labeling for comparative proteomics: Model studies with two serotypes of adenovirus.

[314] Yao, Z., Yaeger, R., Rodrik-Outmezguine, V. S., Tao, A., Torres, N. M., Chang, M. T., Drosten, M., Zhao, H., Cecchi, F., Hembrough, T., Michels, J., Baumert, H., Miles, L., Campbell, N. M., de Stanchina, E., Solit, D. B., Barbacid, M., Taylor, B. S., and Rosen, N. (2017). Tumours with class 3 BRAF mutants are sensitive to the inhibition of activated RAS. *Nature*, 548(7666):234–238.

[315] Youn, A. and Simon, R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, 27(2):175–181.

[316] Yuan, S., Fu, Y., Wang, X., Shi, H., Huang, Y., Song, X., Li, L., Song, N., and Luo, Y. (2008). Voltage-dependent anion channel 1 is involved in endostatin-induced endothelial cell apoptosis. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 22(8):2809–2820.

[317] Zang, J. B., Nosyreva, E. D., Spencer, C. M., Volk, L. J., Musunuru, K., Zhong, R., Stone, E. F., Yuva-Paylor, L. A., Huber, K. M., Paylor, R., Darnell, J. C., and Darnell, R. B. (2009). A mouse model of the human fragile X syndrome I304N mutation. *PLoS genetics*, 5(12):e1000758.

[318] Zanivan, S., Meves, A., Behrendt, K., Schoof, E. M., Neilson, L. J., Cox, J., Tang, H. R., Kalna, G., van Ree, J. H., van Deursen, J. M., Trempus, C. S., Machesky, L. M., Linding, R., Wickström, S. A., Fässler, R., and Mann, M. (2013). In vivo SILAC-based proteomics reveals phosphoproteome changes during mouse skin carcinogenesis. *Cell reports*, 3(2):552–566.

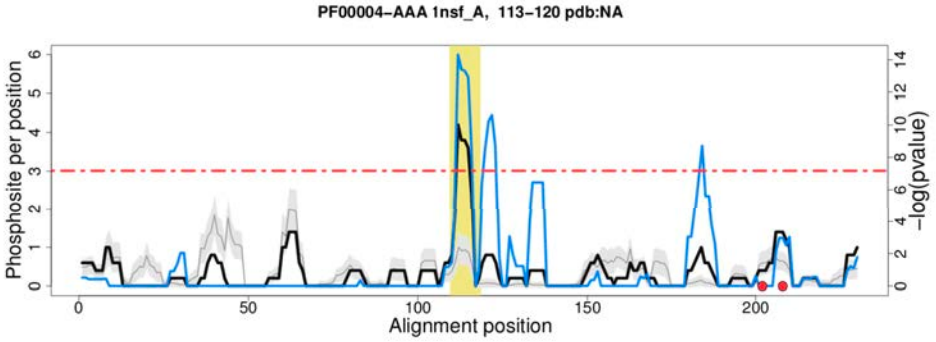
[319] Zeke, A., Lukács, M., Lim, W. A., and Reményi, A. (2009). Scaffolds: interaction platforms for cellular signalling circuits. *Trends in cell biology*, 19(8):364–374.

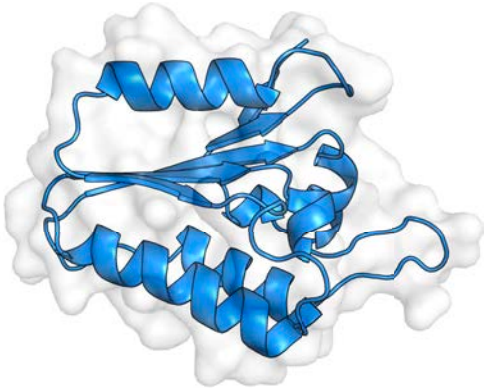
[320] Zhang, M. S., Brunner, S. F., Huguenin-Dezot, N., Liang, A. D., Schmied, W. H., Rogerson, D. T., and Chin, J. W. (2017). Biosynthesis and genetic encoding of phosphothreonine through parallel selection and deep sequencing. *Nature methods*, 14(7):729–736.

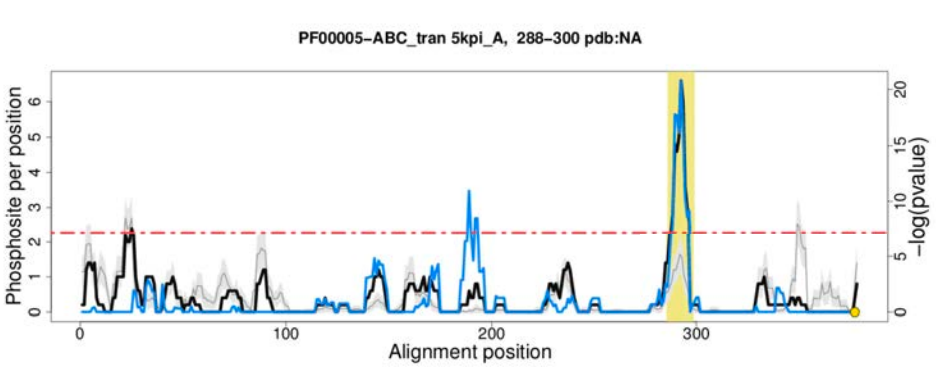
[321] Zhou, H., Ye, M., Dong, J., Corradini, E., Cristobal, A., Heck, A. J. R., Zou, H., and Mohammed, S. (2013). Robust phosphoproteome enrichment using monodisperse microsphere–based immobilized titanium (IV) ion affinity chromatography. *Nature protocols*, 8:461.

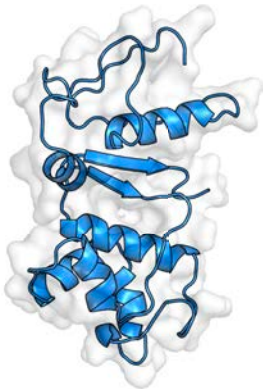
Appendix A

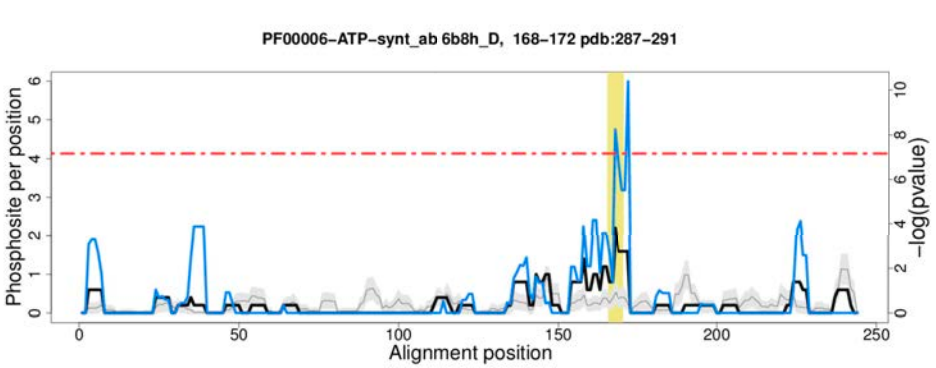
Remaining Phosphorylation hotspots

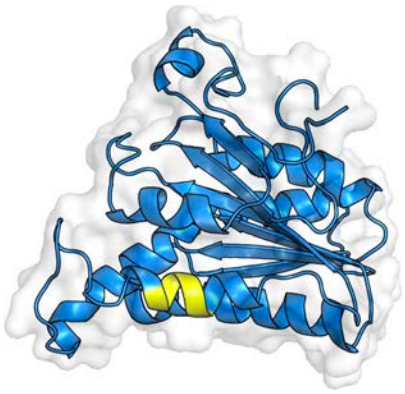

PF00004-AAA 1nsf_A, 113-120 pdb:NA
Phosphosite per position
-log(pvalue)
Alignment position

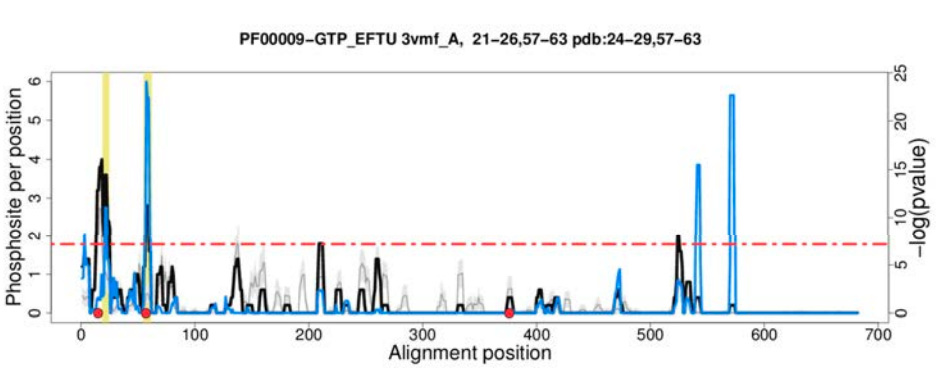


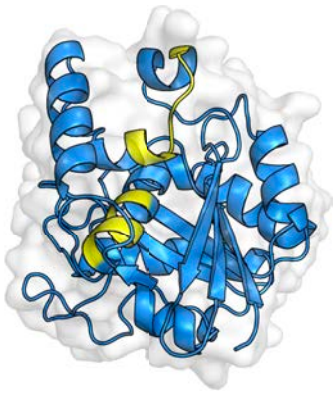

PF00005-ABC_tran 5kpi_A, 288-300 pdb:NA
Phosphosite per position
-log(pvalue)
Alignment position

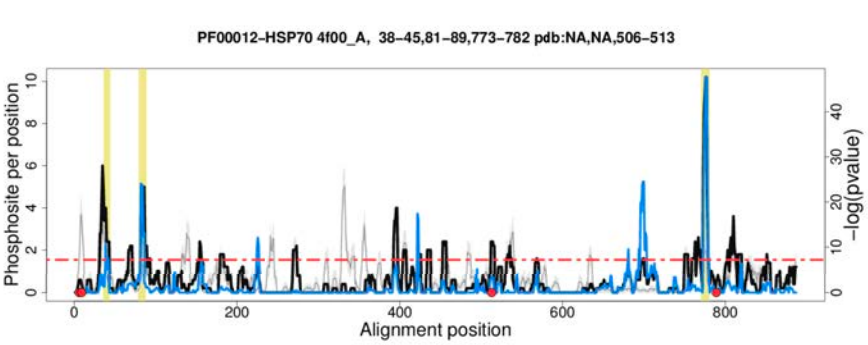


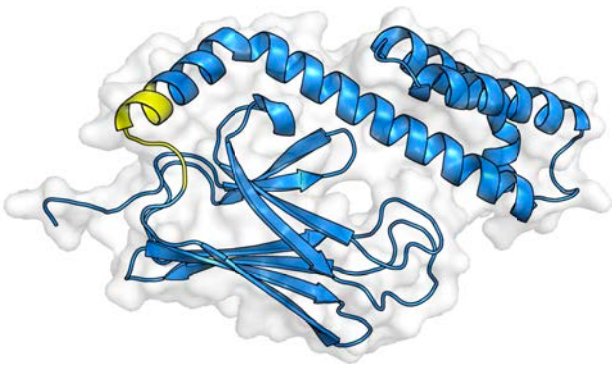

PF00006-ATP-synt_ab 6b8h_D, 168-172 pdb:287-291
Phosphosite per position
-log(pvalue)
Alignment position

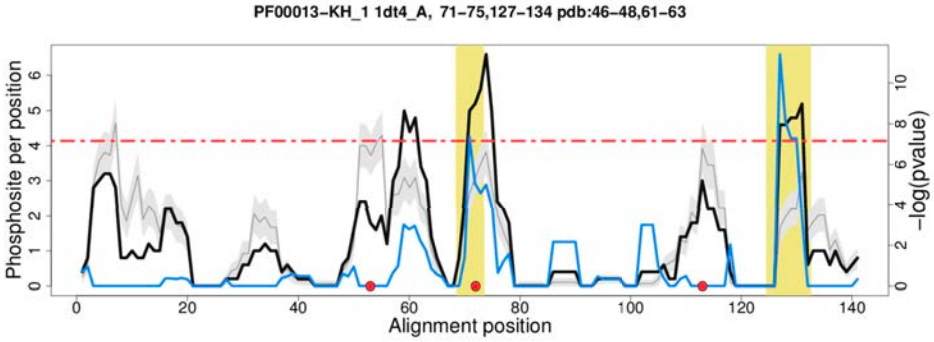


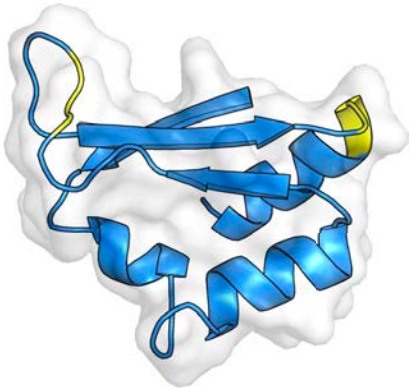

PF00009-GTP_EFTU 3vmf_A, 21-26,57-63 pdb:24-29,57-63
Phosphosite per position
-log(pvalue)
Alignment position

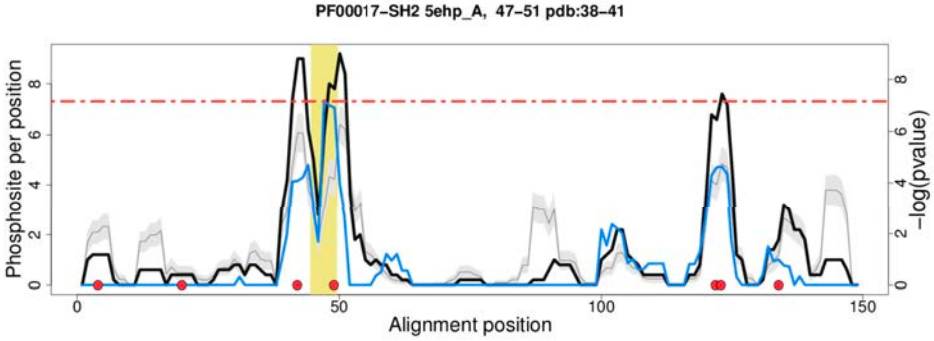


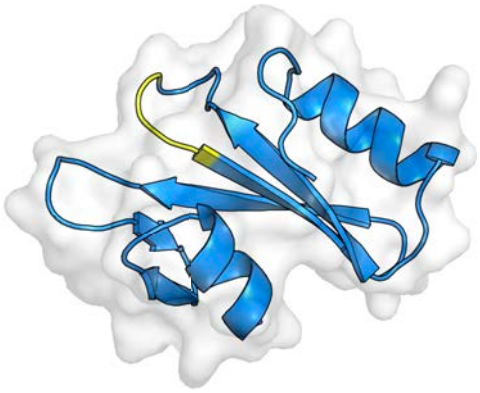

PF00012-HSP70 4f00_A, 38-45,81-89,773-782 pdb:NA,NA,506-513
Phosphosite per position
-log(pvalue)
Alignment position

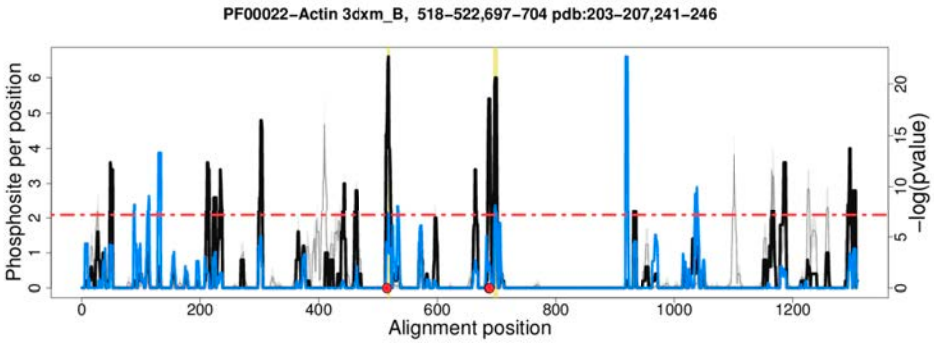


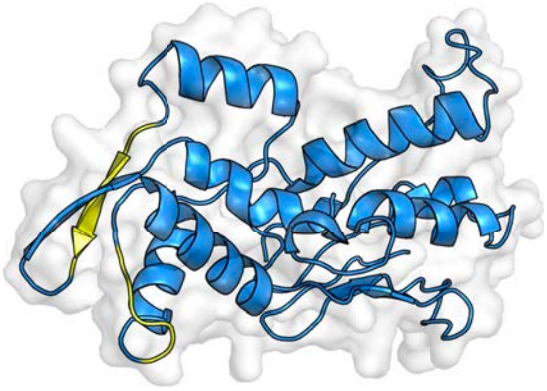

PF00013–KH_1 1dt4_A, 71–75,127–134 pdb:46–48,61–63
Phosphosite per position
–log(pvalue)
Alignment position

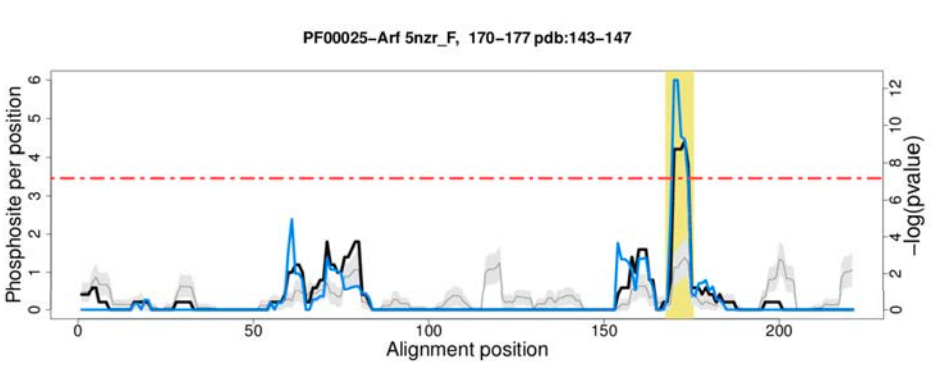


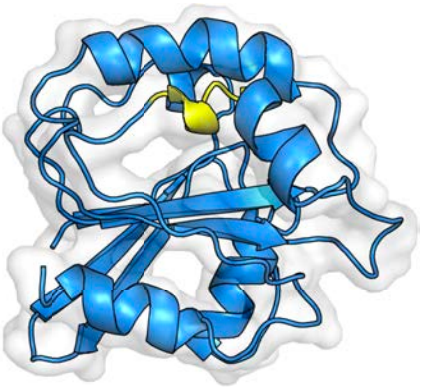

PF00017–SH2 5ehp_A, 47–51 pdb:38–41
Phosphosite per position
–log(pvalue)
Alignment position

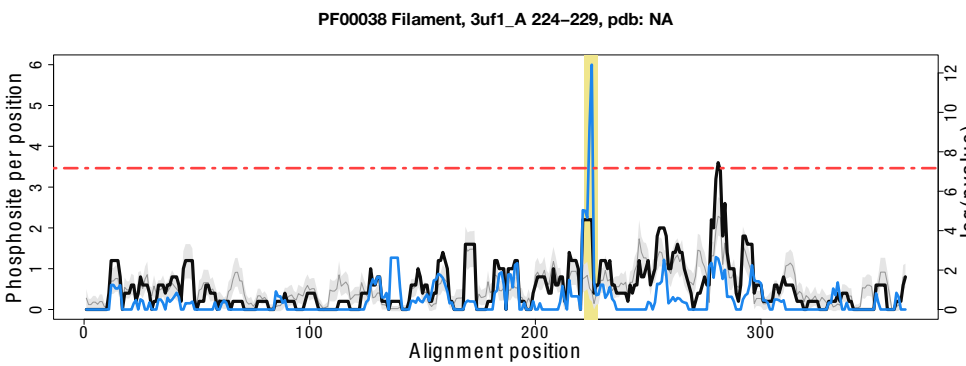



PF00022–Actin 3dxm_B, 518–522,697–704 pdb:203–207,241–246
Phosphosite per position
–log(pvalue)
Alignment position

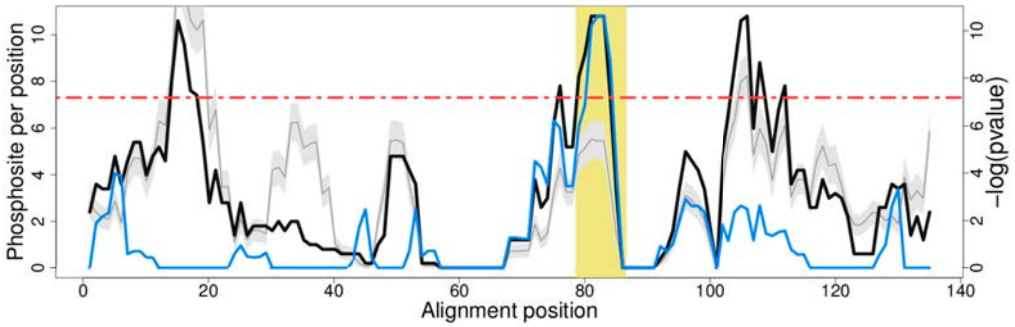


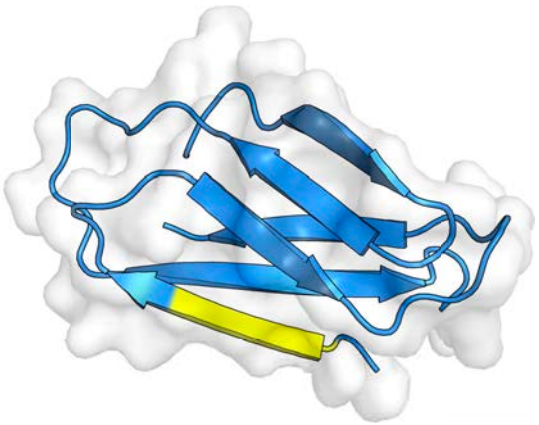

PF00025–Arf 5nzr_F, 170–177 pdb:143–147
Phosphosite per position
–log(pvalue)
Alignment position



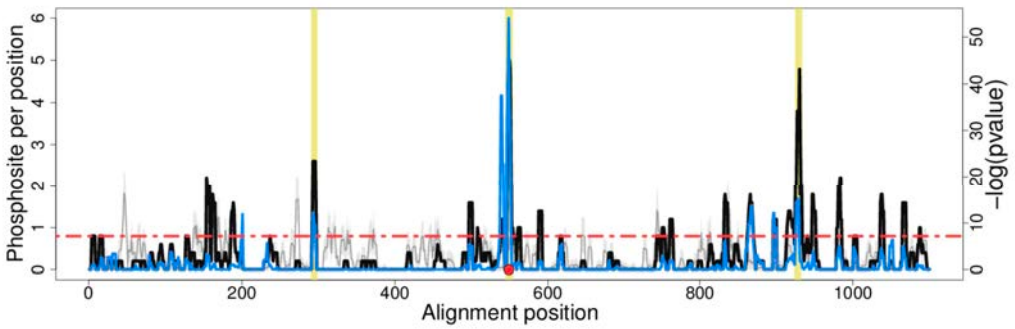

PF00038 Filament, 3uf1_A 224–229, pdb: NA
Phosphosite per position
–log(pvalue)
Alignment position

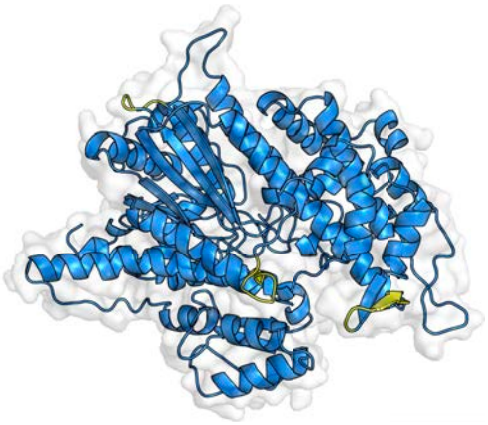
PF00041-fn3 5e95_B, 81-88 pdb:46-49



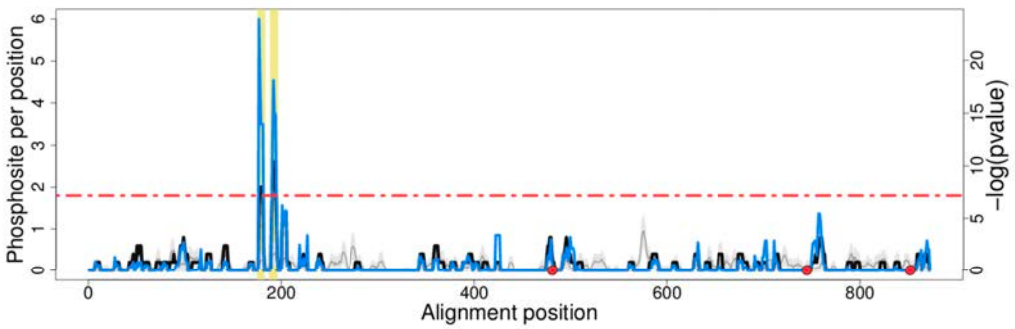


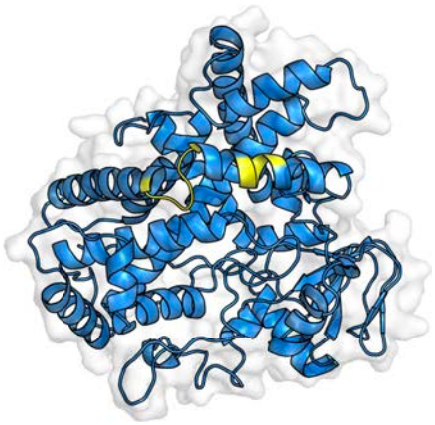
PF00063-Myosin_head 3mkd_A, 293-300,547-556,926-934 pdb:249-252,401-408,622-630



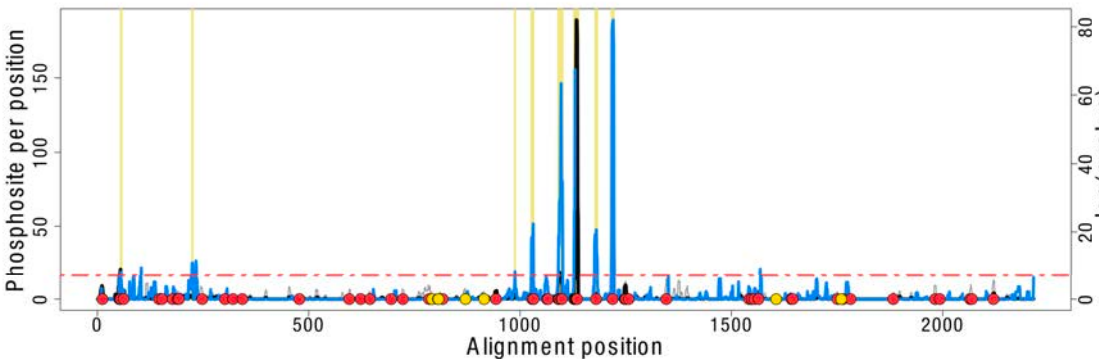


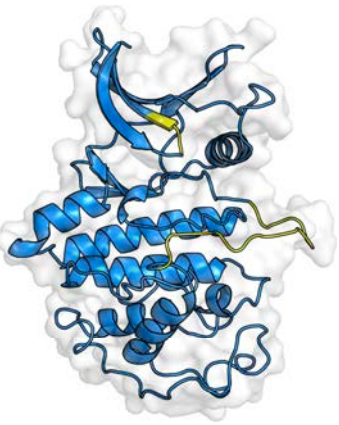
PF00067-p450 4rui_A, 177-184,190-197 pdb:130-133,138-143



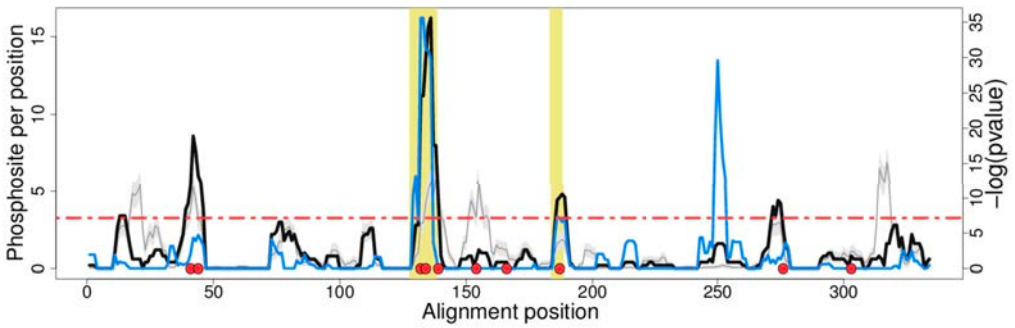


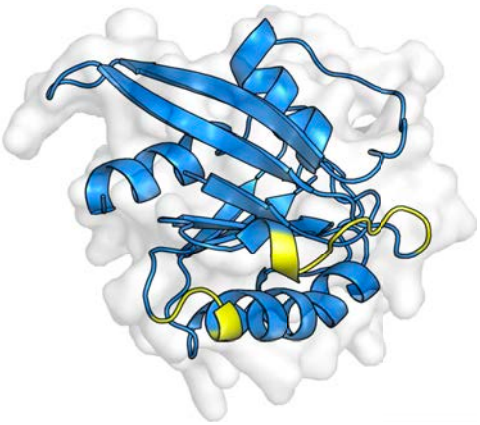
PF00069-Pkinase 5uq2_A, 55-60,223-228,988-992,1027-1035,1090-1104,1128-1141,1178-1186,1216-1225 pdb:15-17,NA,NA,NA,151-155,156-162,163-163,164-166



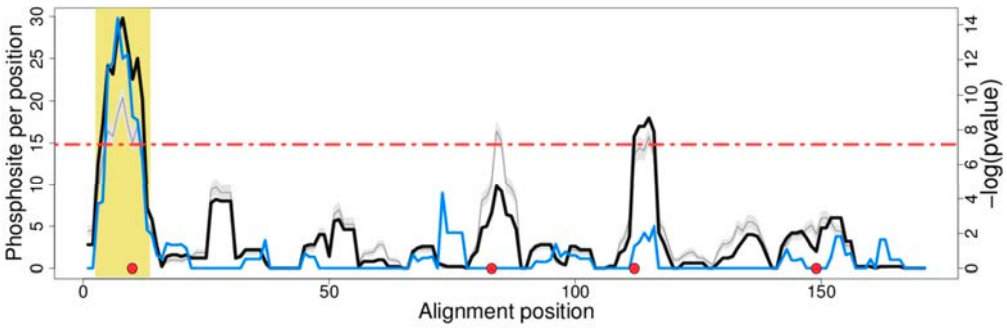


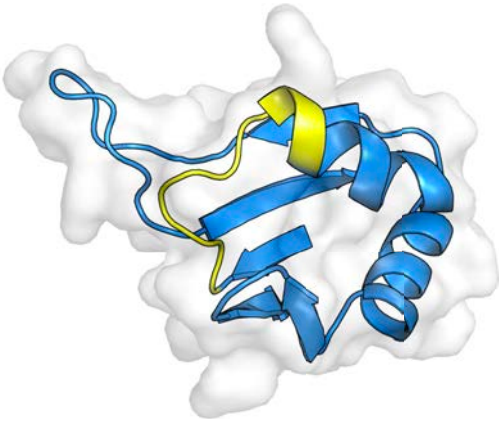
PF00071-Ras 5dha_A, 130-140,186-190 pdb:71-78,110-113



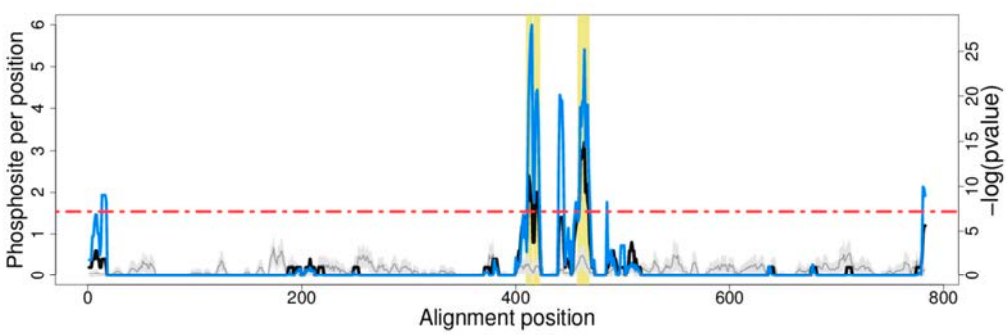


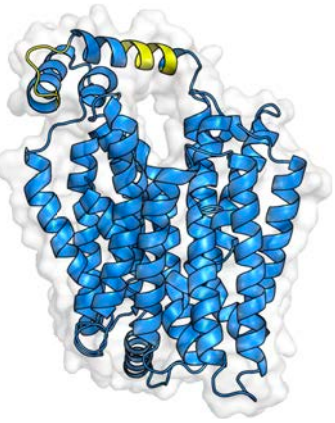
PF00076-RRM_1 1b7f_A, 5-15 pdb:131-140



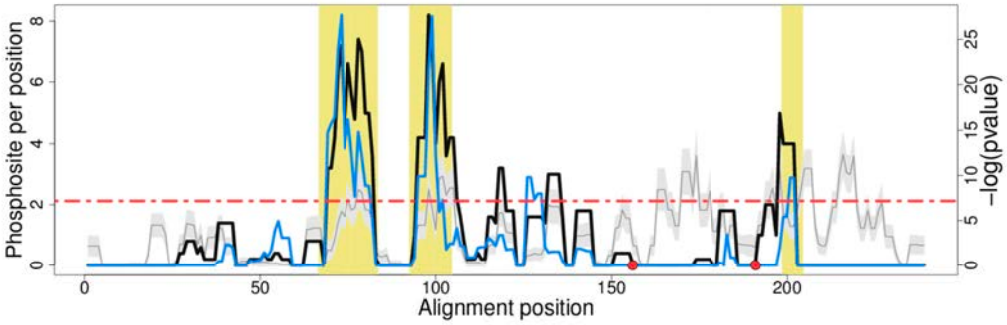


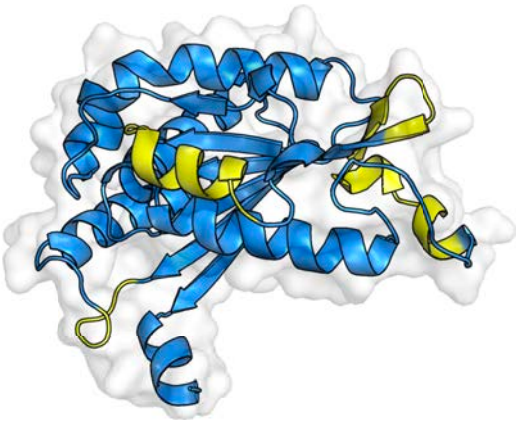
PF00083-Sugar_tr 5eqg_A, 412-417,419-424,460-470 pdb:232-235,236-237,245-251



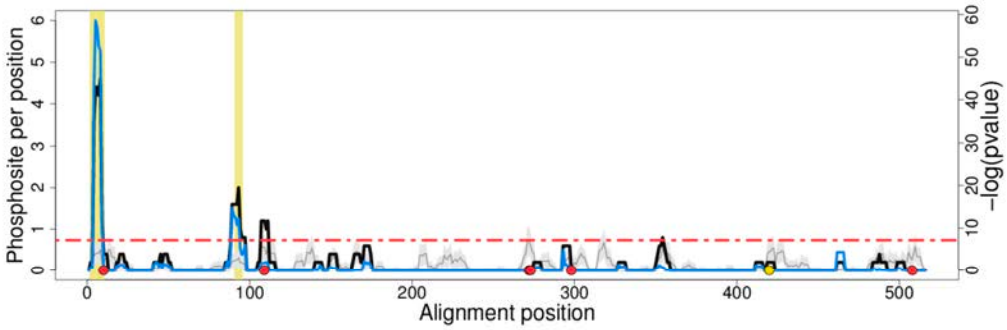


PF00091-Tubulin 5xi5_B, 69-85,95-106,201-206 pdb:42-58,68-79,171-175

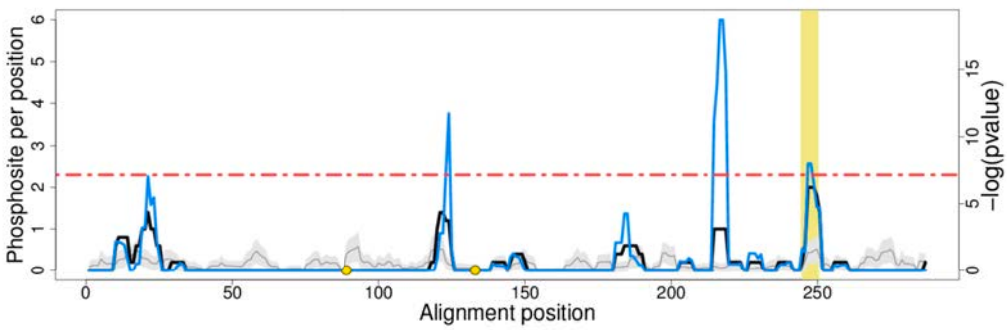


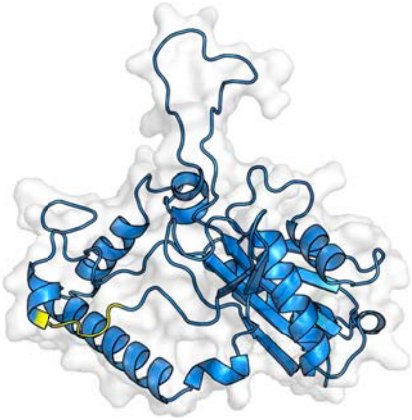


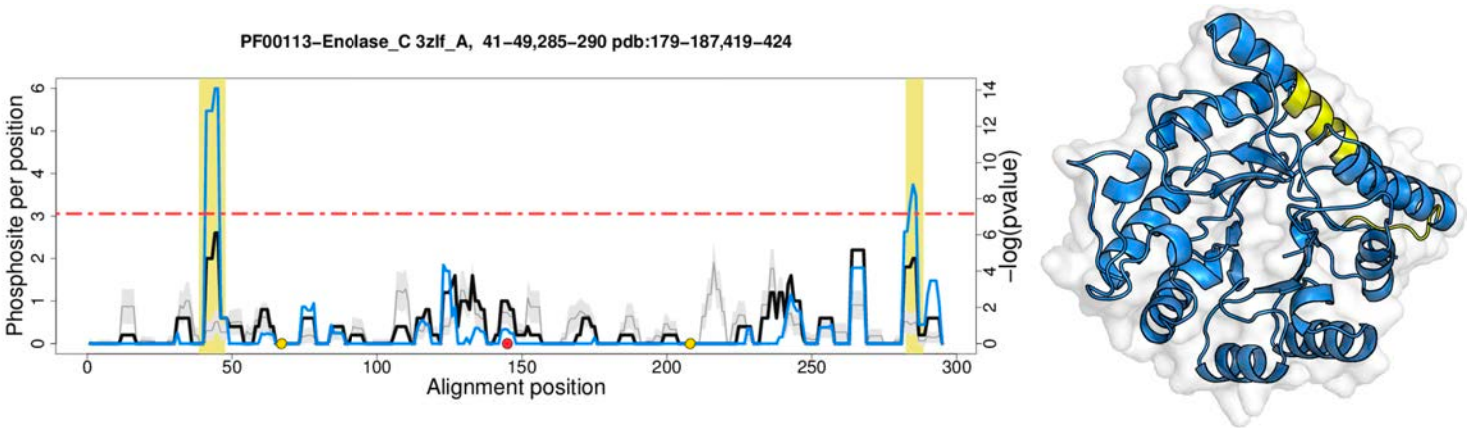
PF00102-Y_phosphatase 2fjn_A, 4-12,93-97 pdb:543-550,562-563

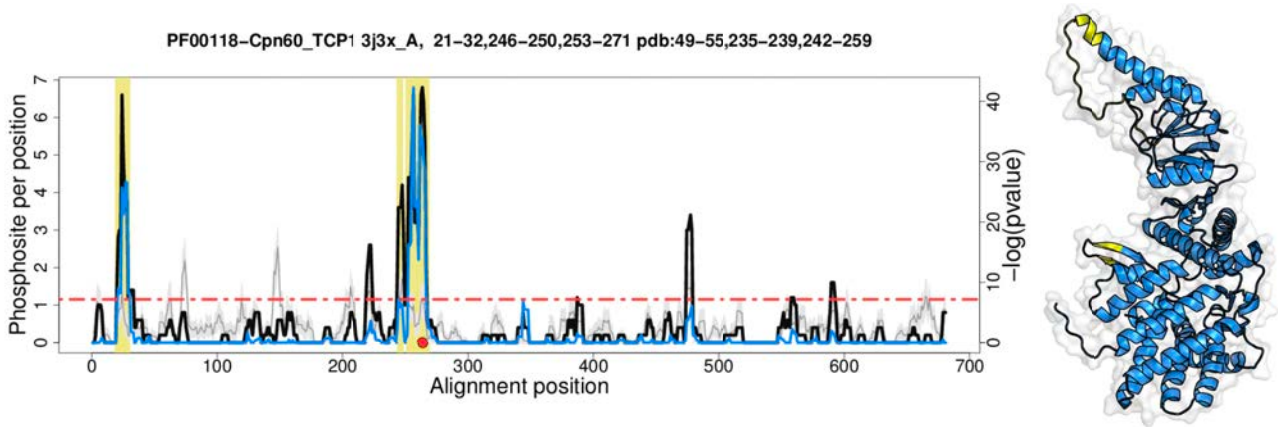


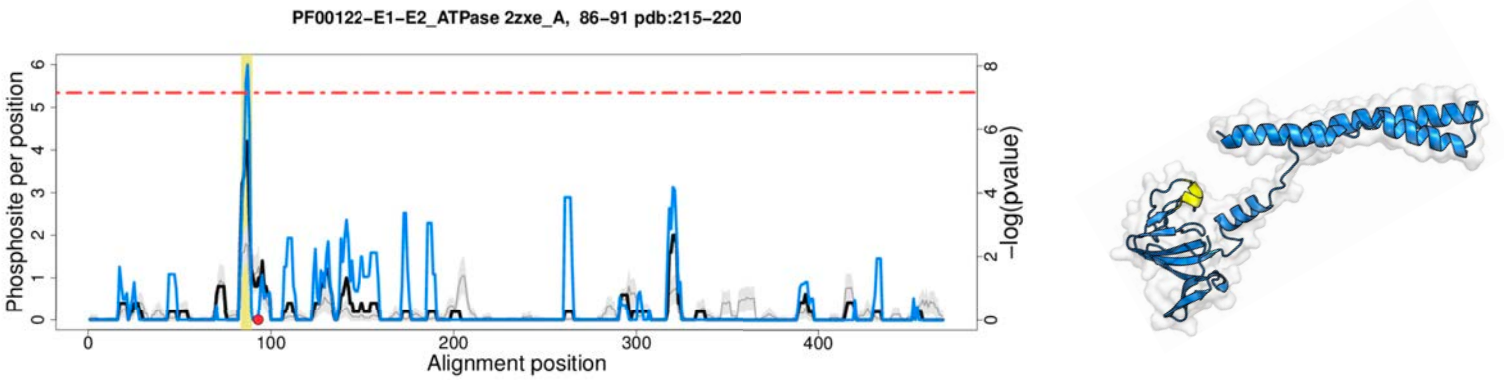
PF00108-Thiolase_N 4n44_A, 247-252 pdb:220-225

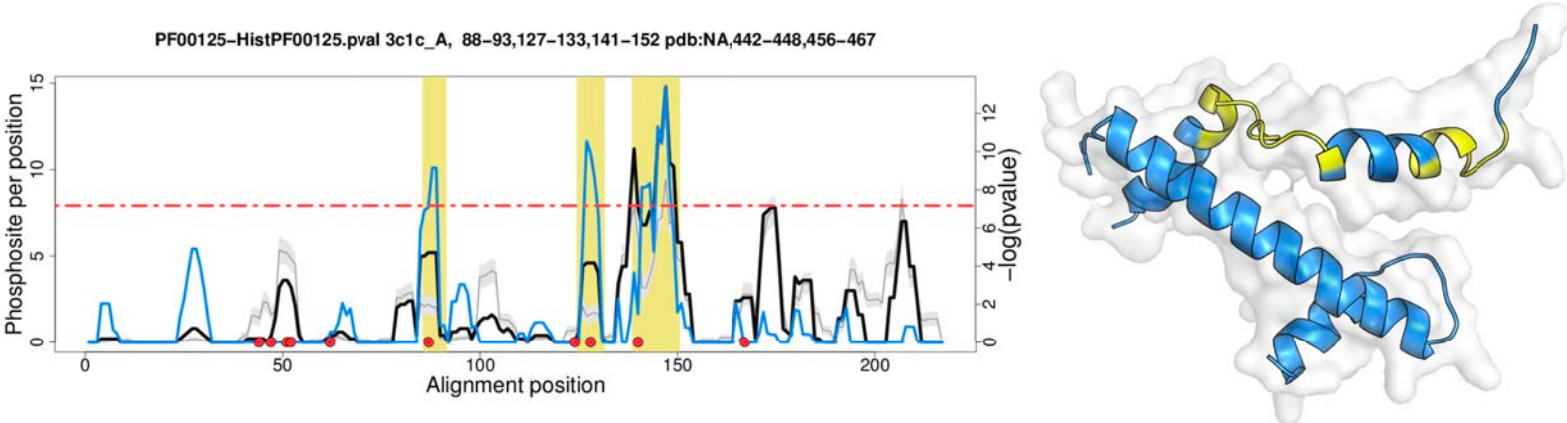


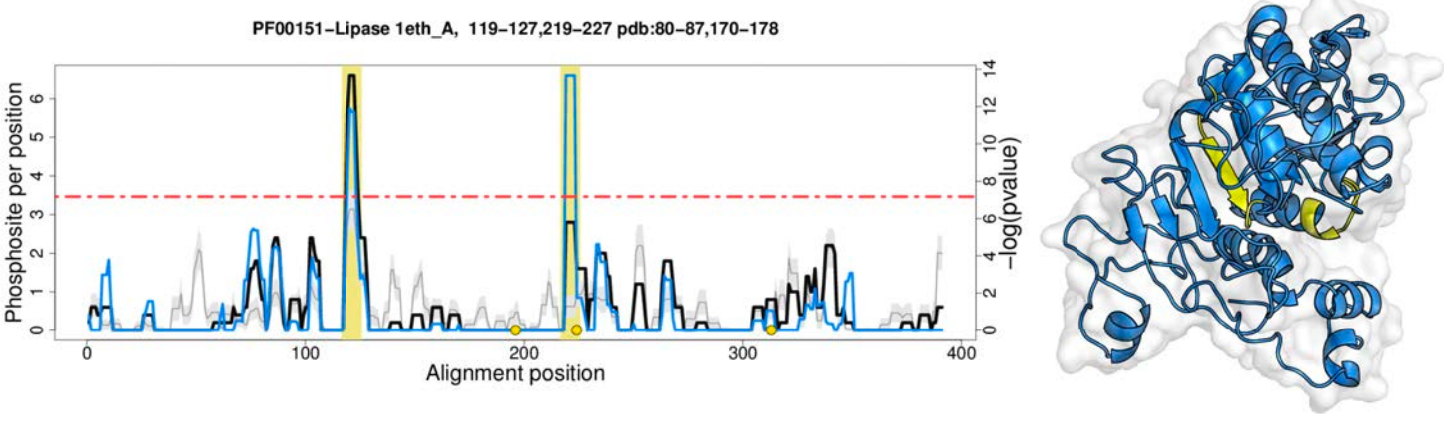


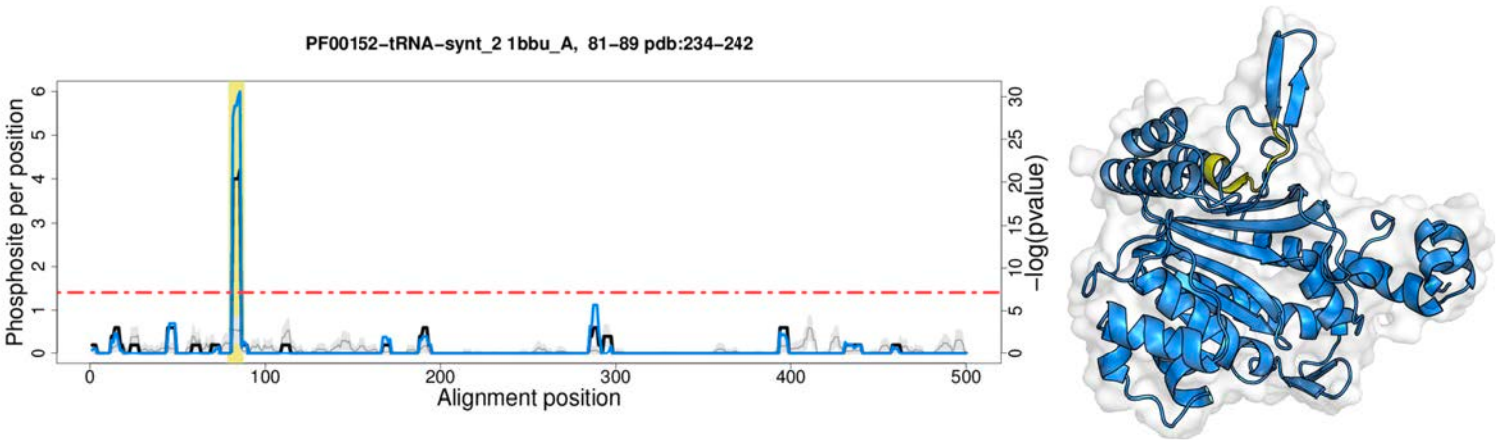

PF00113–Enolase_C 3zlf_A, 41–49,285–290 pdb:179–187,419–424
Phosphosite per position
−log(pvalue)
Alignment position

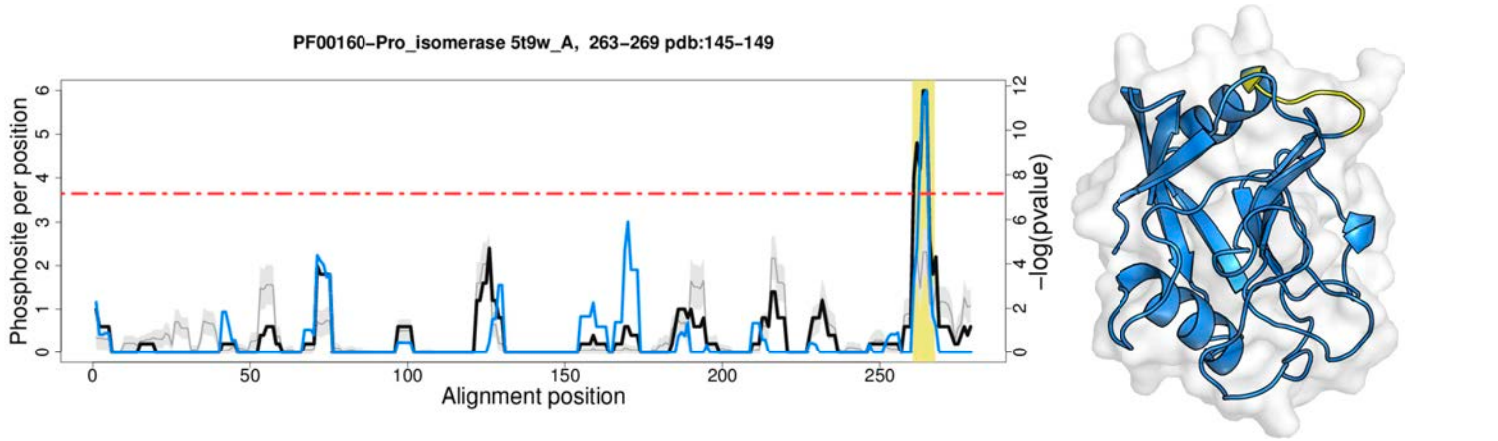

PF00118–Cpn60_TCP1 3j3x_A, 21–32,246–250,253–271 pdb:49–55,235–239,242–259
Phosphosite per position
−log(pvalue)
Alignment position

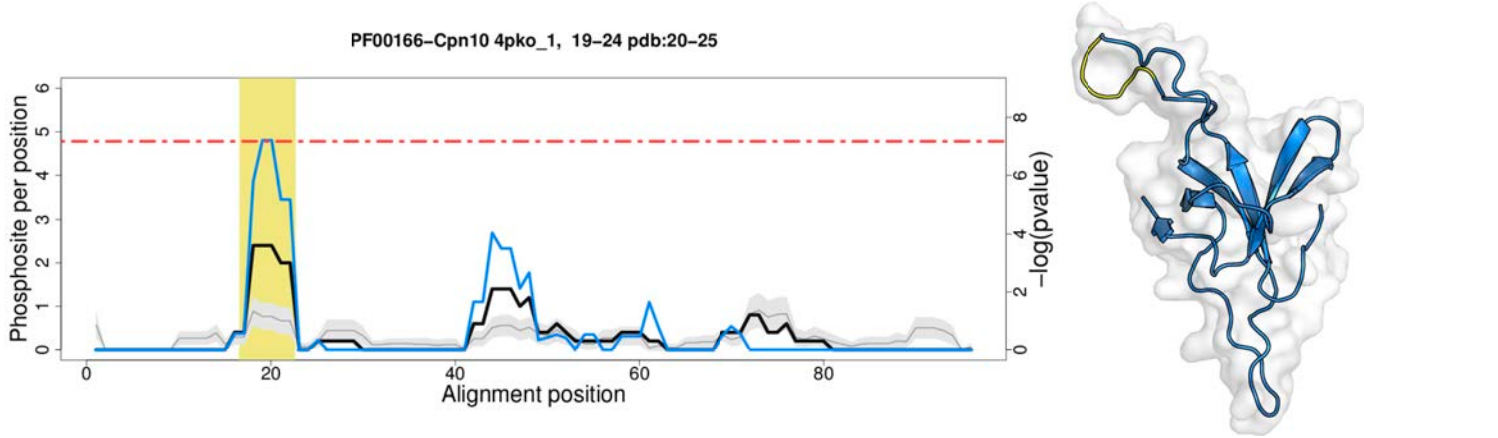

PF00122–E1–E2_ATPase 2zxe_A, 86–91 pdb:215–220
Phosphosite per position
−log(pvalue)
Alignment position

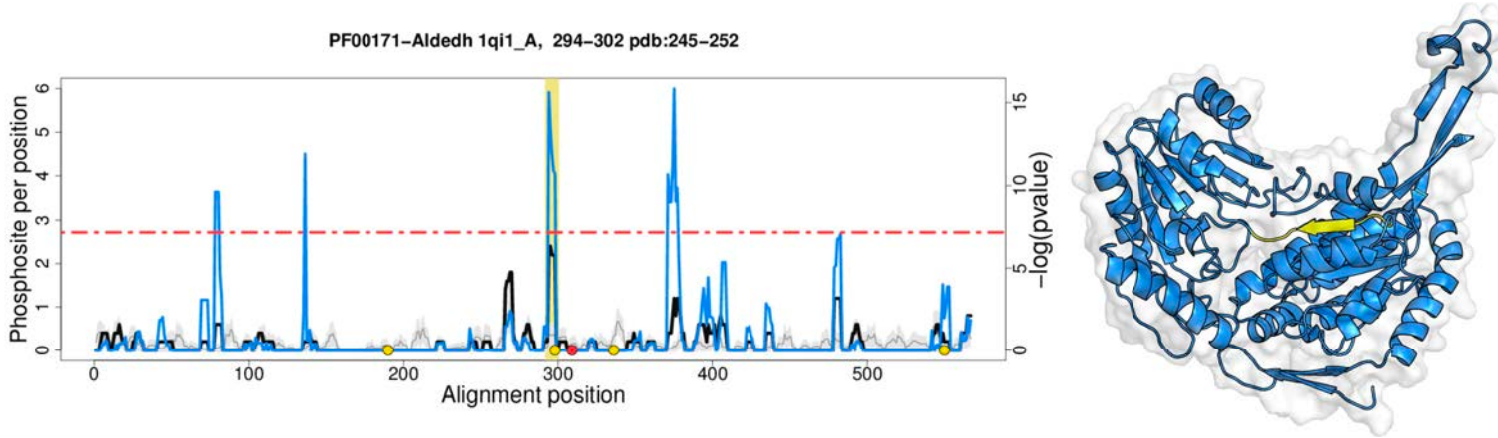

PF00125–HistPF00125.pval 3c1c_A, 88–93,127–133,141–152 pdb:NA,442–448,456–467
Phosphosite per position
−log(pvalue)
Alignment position

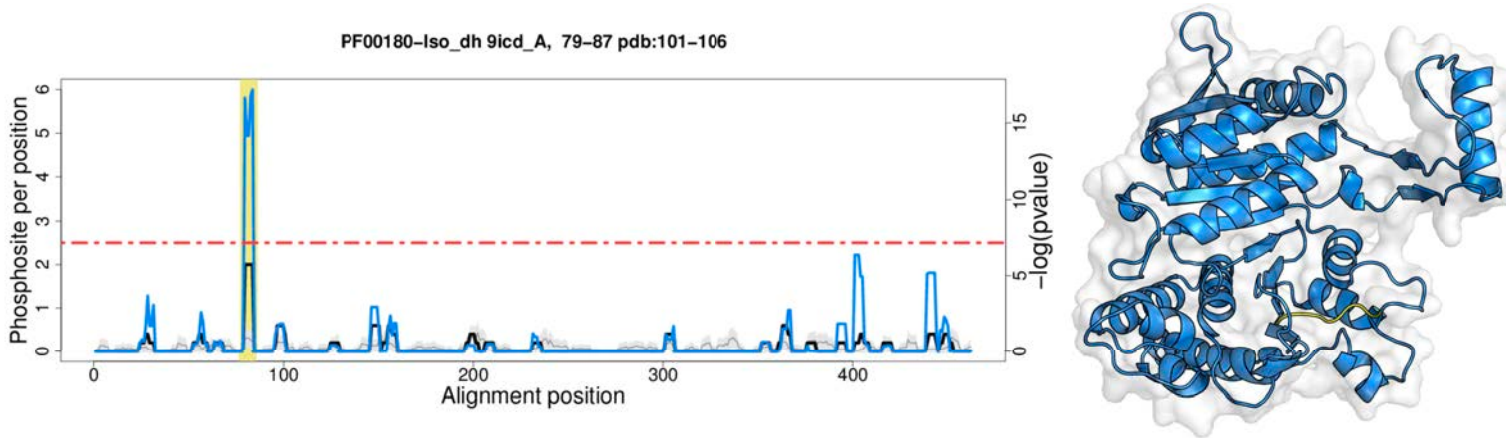

PF00151–Lipase 1eth_A, 119–127,219–227 pdb:80–87,170–178
Phosphosite per position
−log(pvalue)
Alignment position

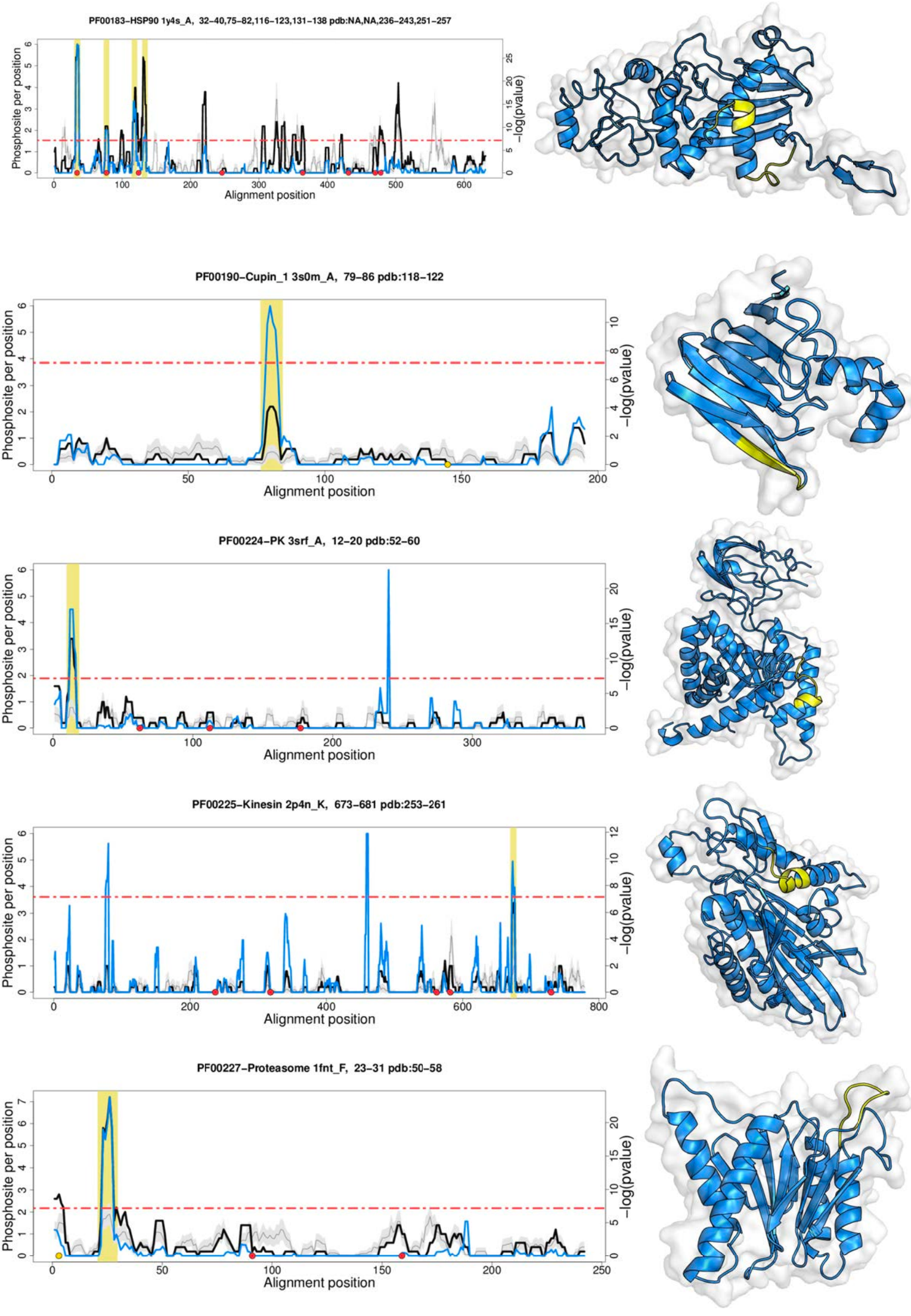

PF00152–tRNA–synt_2 1bbu_A, 81–89 pdb:234–242
Phosphosite per position
–log(pvalue)
Alignment position

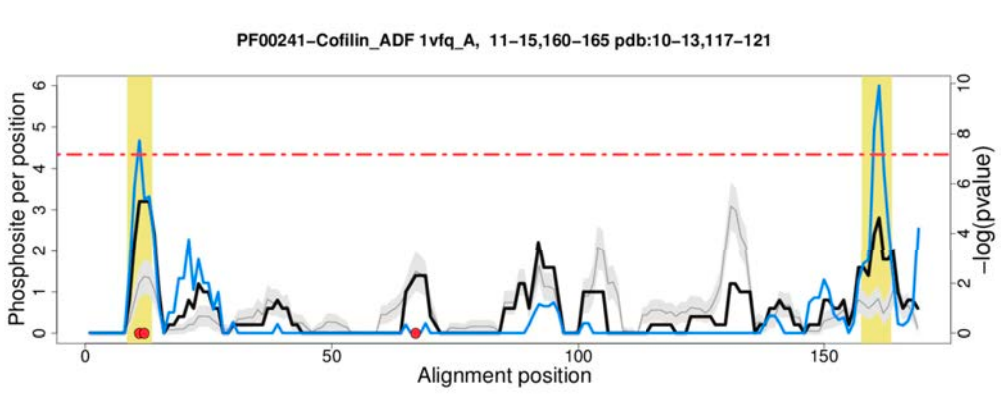

PF00160–Pro_isomerase 5t9w_A, 263–269 pdb:145–149
Phosphosite per position
–log(pvalue)
Alignment position

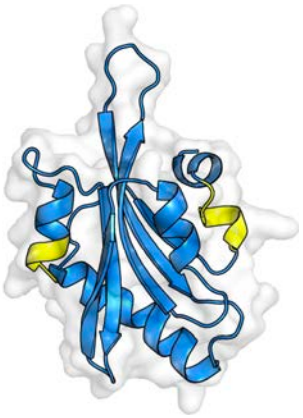

PF00166–Cpn10 4pko_1, 19–24 pdb:20–25
Phosphosite per position
–log(pvalue)
Alignment position

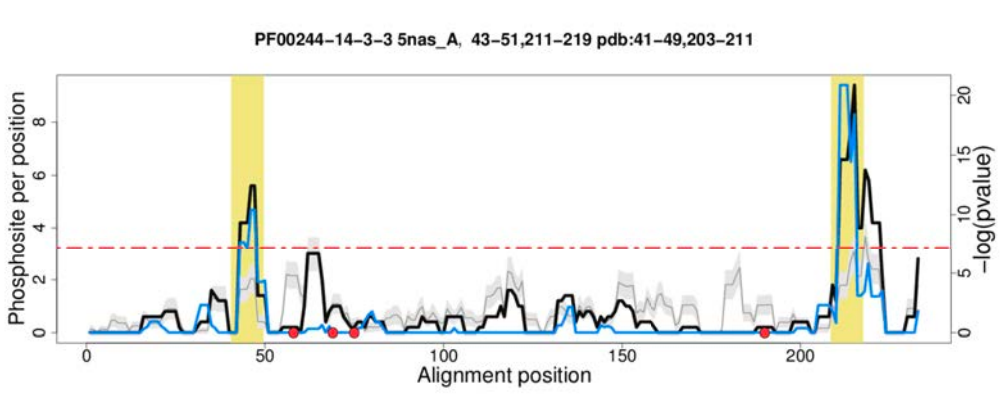

PF00171–Aldedh 1qi1_A, 294–302 pdb:245–252
Phosphosite per position
–log(pvalue)
Alignment position

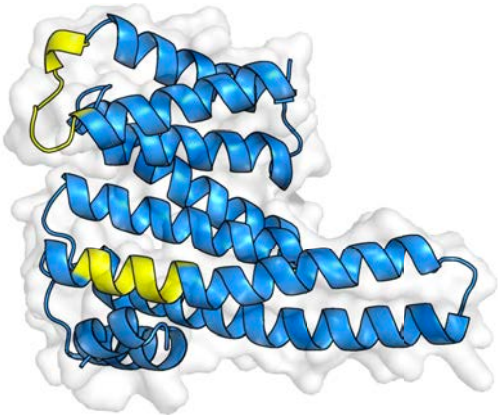

PF00180–Iso_dh 9icd_A, 79–87 pdb:101–106
Phosphosite per position
–log(pvalue)
Alignment position

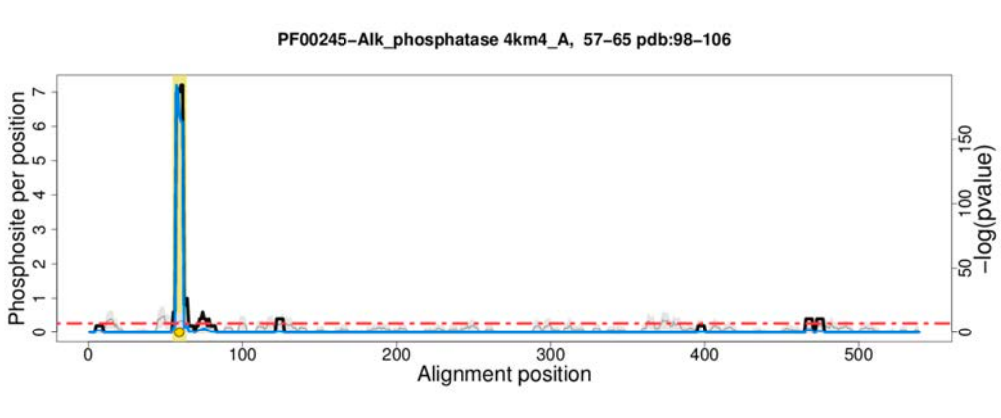

PF00183–HSP90 1y4s_A, 32–40,75–82,116–123,131–138 pdb:NA,NA,236–243,251–257
Phosphosite per position
−log(pvalue)
Alignment position
PF00190–Cupin_1 3s0m_A, 79–86 pdb:118–122
Phosphosite per position
−log(pvalue)
Alignment position
PF00224–PK 3srf_A, 12–20 pdb:52–60
Phosphosite per position
−log(pvalue)
Alignment position
PF00225–Kinesin 2p4n_K, 673–681 pdb:253–261
Phosphosite per position
−log(pvalue)
Alignment position
PF00227–Proteasome 1fnt_F, 23–31 pdb:50–58
Phosphosite per position
−log(pvalue)
Alignment position

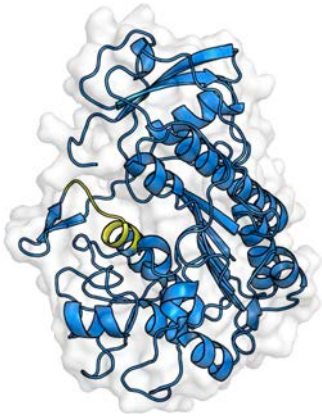

PF00241–Cofilin_ADF 1vfq_A, 11–15,160–165 pdb:10–13,117–121
Phosphosite per position
−log(pvalue)
Alignment position

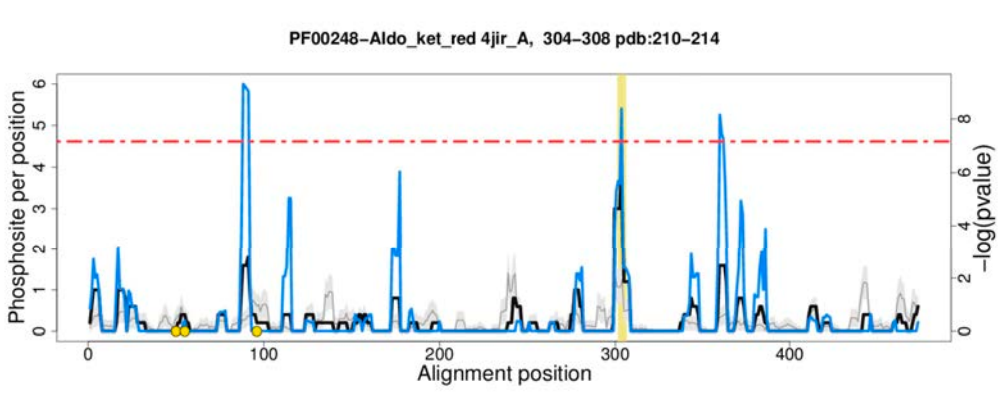


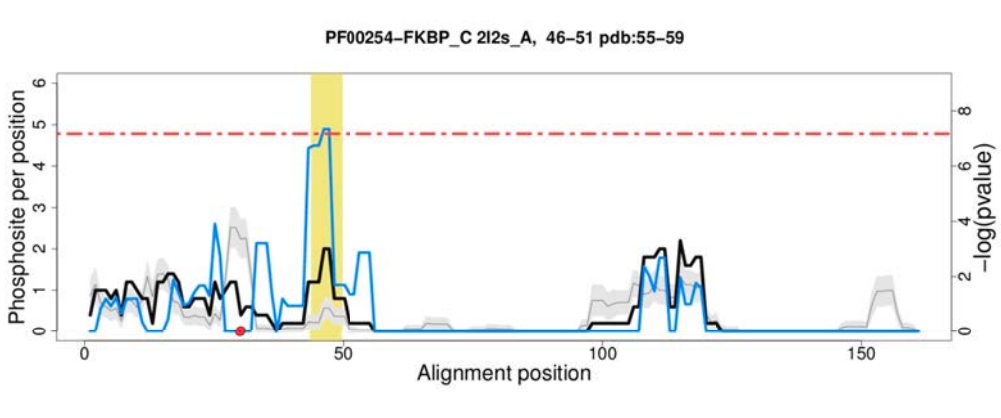

PF00244–14–3–3 5nas_A, 43–51,211–219 pdb:41–49,203–211
Phosphosite per position
−log(pvalue)
Alignment position

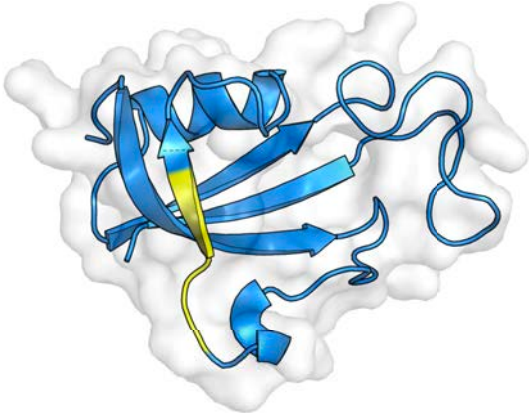


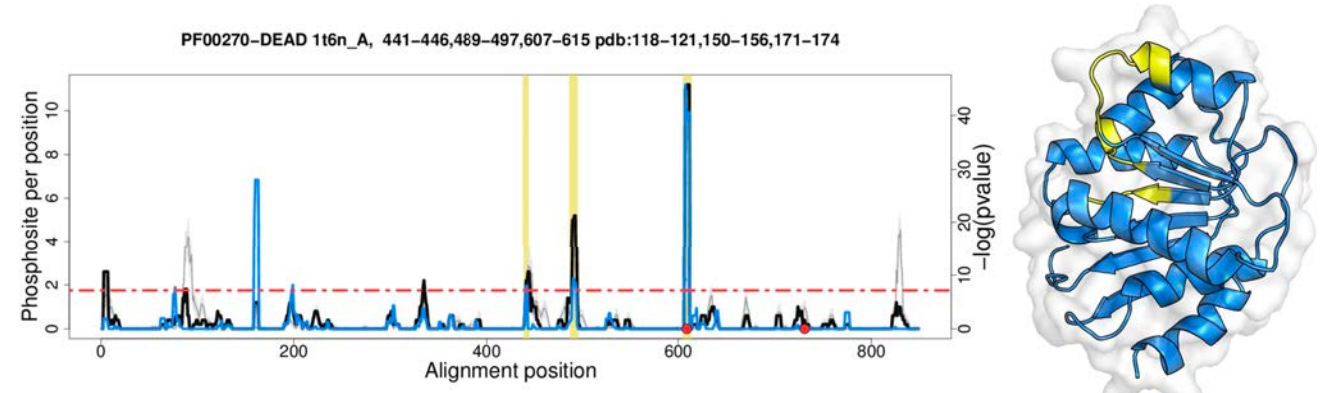

PF00245–Alk_phosphatase 4km4_A, 57–65 pdb:98–106
Phosphosite per position
−log(pvalue)
Alignment position

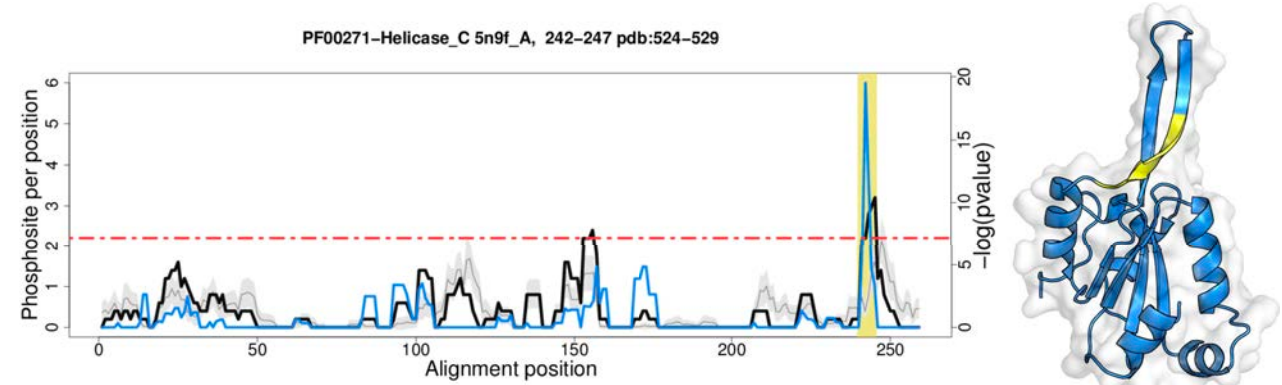


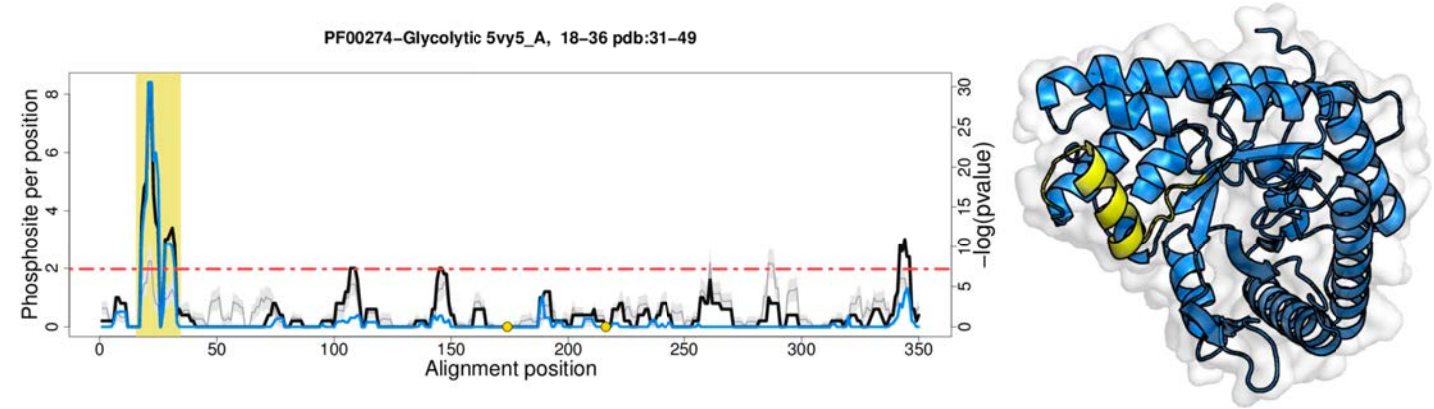

PF00248–Aldo_ket_red 4jir_A, 304–308 pdb:210–214
Phosphosite per position
−log(pvalue)
Alignment position

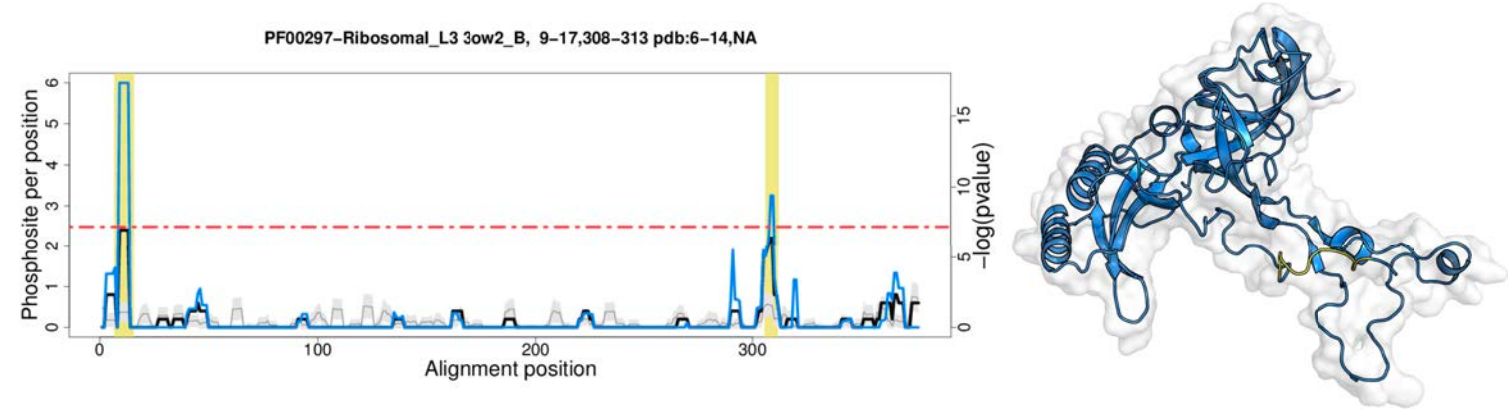

PF00254–FKBP_C 2l2s_A, 46–51 pdb:55–59
Phosphosite per position
−log(pvalue)
Alignment position

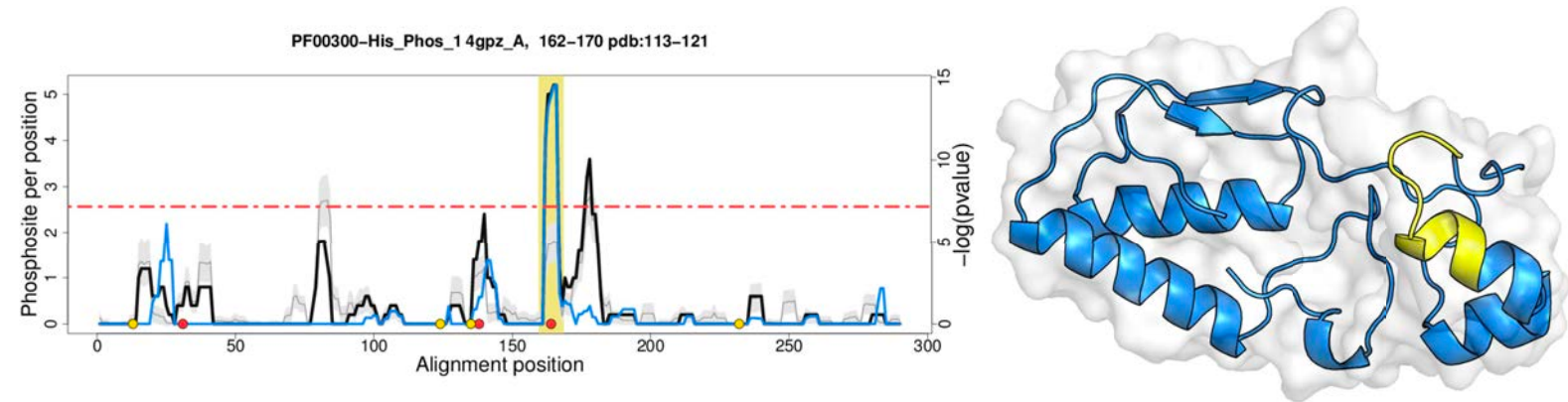


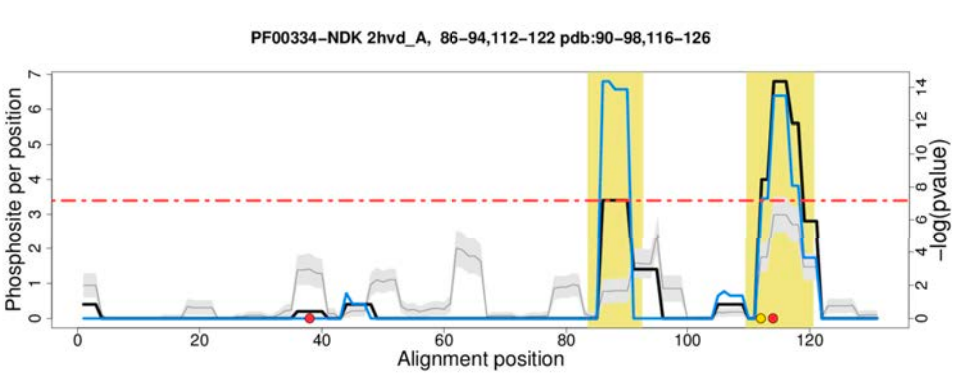

PF00270–DEAD 1t6n_A, 441–446,489–497,607–615 pdb:118–121,150–156,171–174
Phosphosite per position
–log(pvalue)
Alignment position

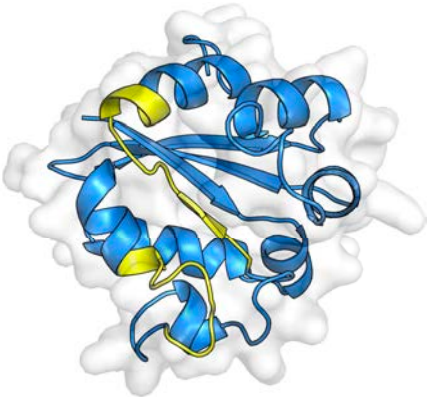

PF00271–Helicase_C 5n9f_A, 242–247 pdb:524–529
Phosphosite per position
–log(pvalue)
Alignment position

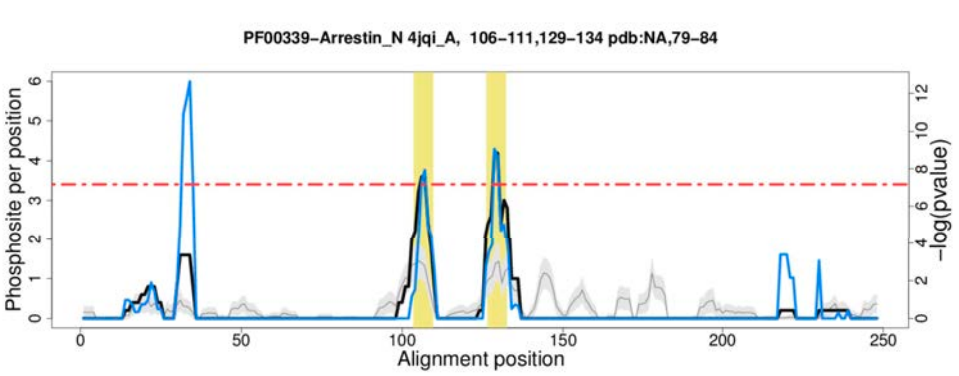

PF00274–Glycolytic 5vy5_A, 18–36 pdb:31–49
Phosphosite per position
–log(pvalue)
Alignment position

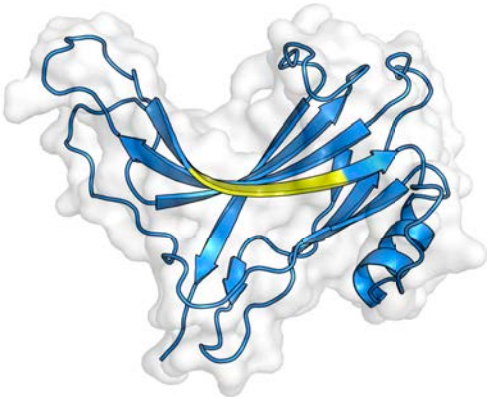

PF00297–Ribosomal_L3 3ow2_B, 9–17,308–313 pdb:6–14,NA
Phosphosite per position
–log(pvalue)
Alignment position

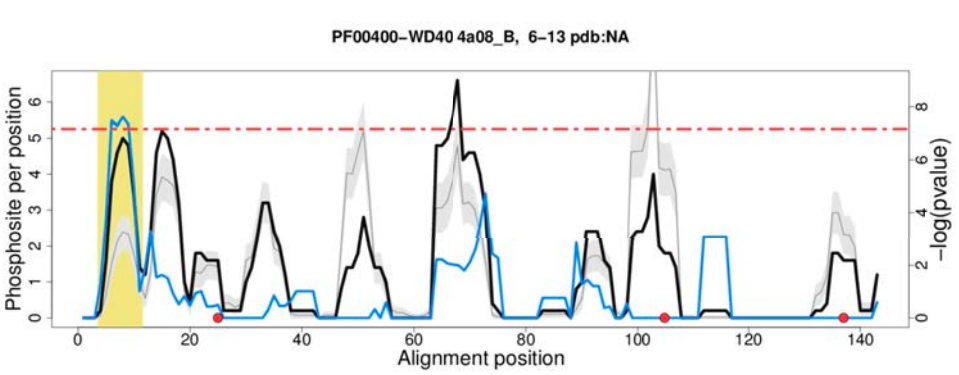

PF00300–His_Phos_1 4gpz_A, 162–170 pdb:113–121
Phosphosite per position
–log(pvalue)
Alignment position

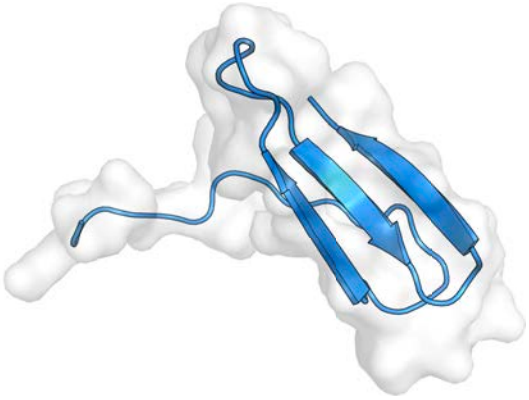

PF00334–NDK 2hvd_A, 86–94,112–122 pdb:90–98,116–126
Phosphosite per position
−log(pvalue)
Alignment position

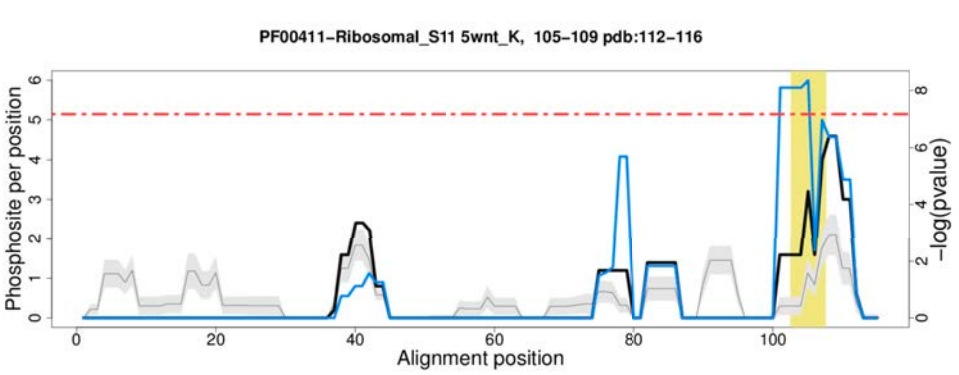


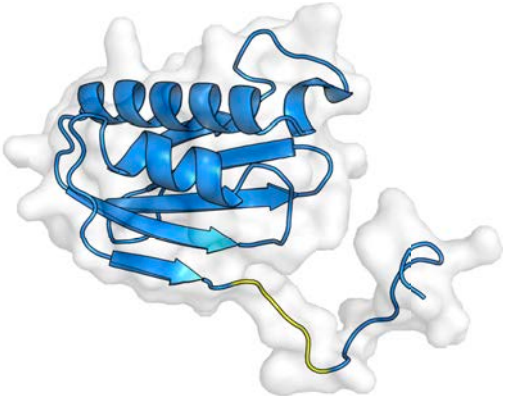

PF00339–Arrestin_N 4jqi_A, 106–111,129–134 pdb:NA,79–84
Phosphosite per position
−log(pvalue)
Alignment position

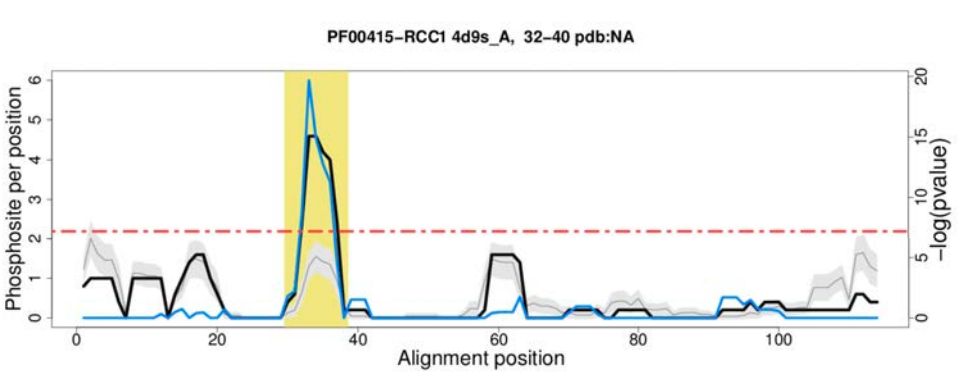


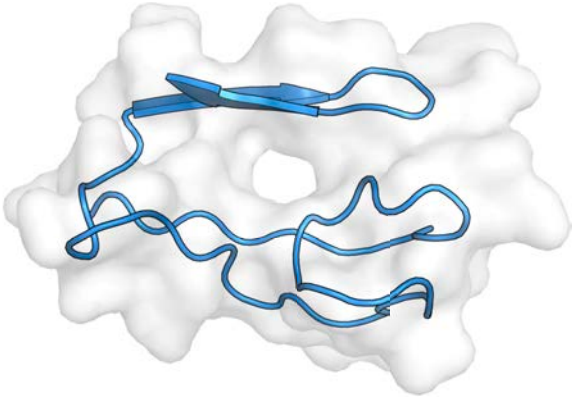

PF00400–WD40 4a08_B, 6–13 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position

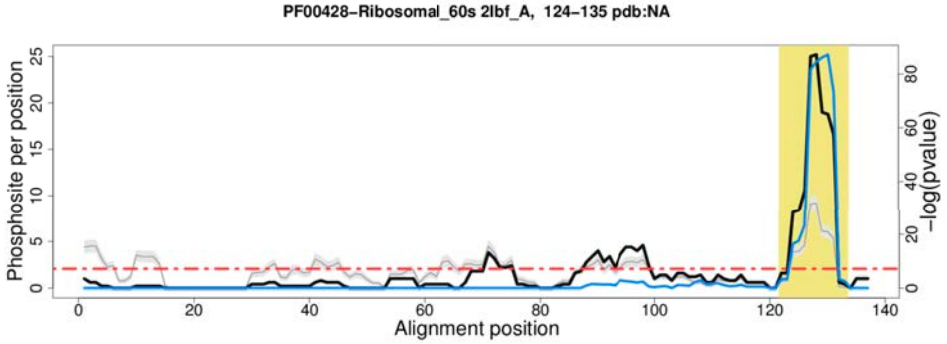


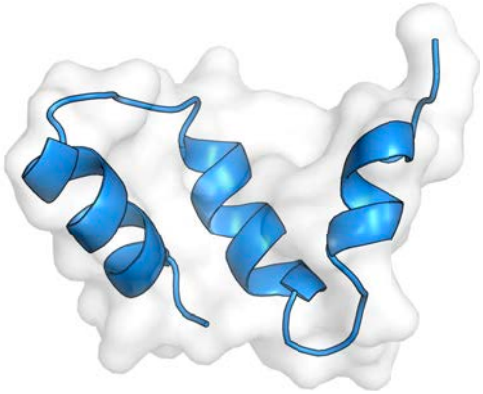

PF00411–Ribosomal_S11 5wnt_K, 105–109 pdb:112–116
Phosphosite per position
−log(pvalue)
Alignment position

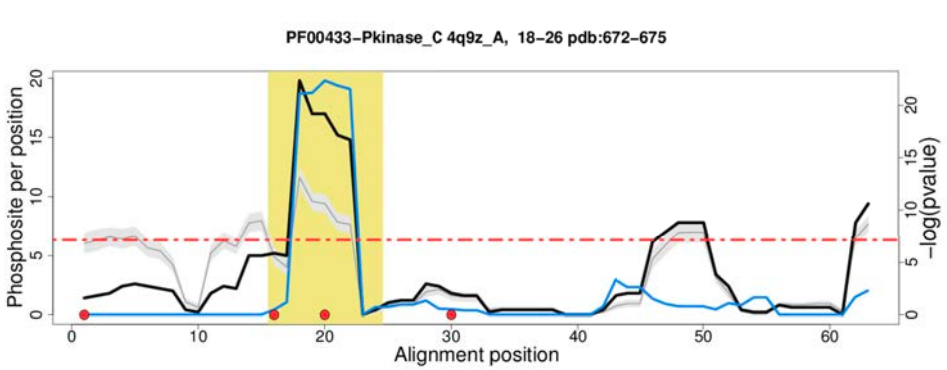


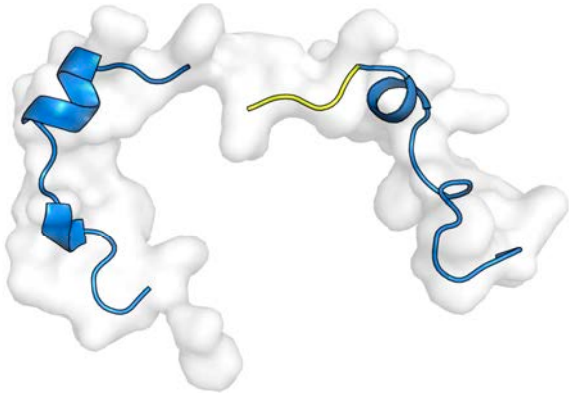

PF00415–RCC1 4d9s_A, 32–40 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position

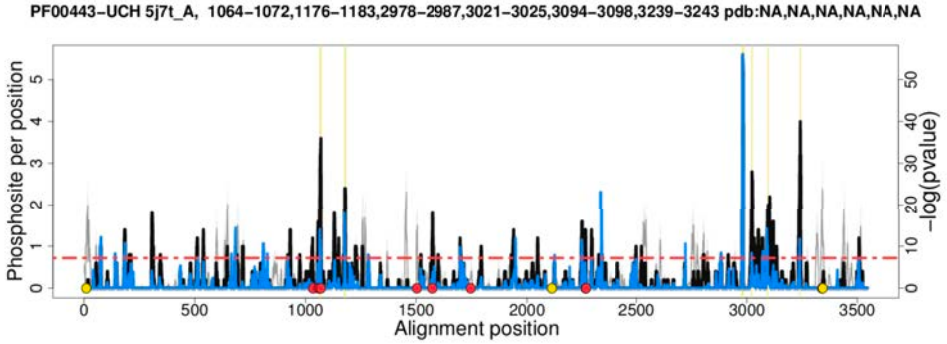


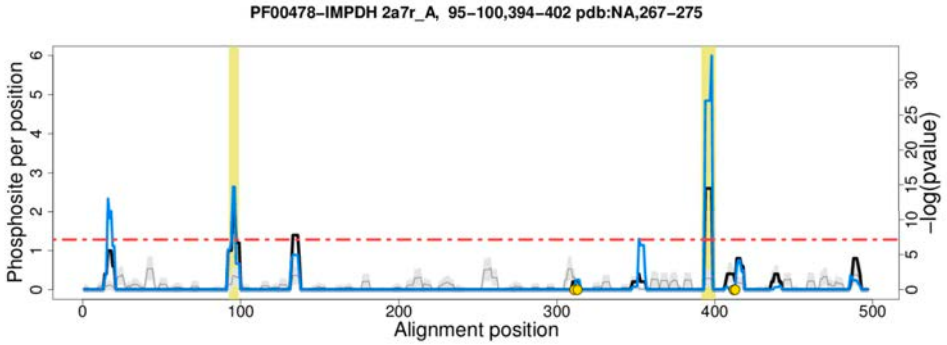

PF00428-Ribosomal_60s 2lbf_A, 124-135 pdb:NA
Phosphosite per position
-log(pvalue)
Alignment position

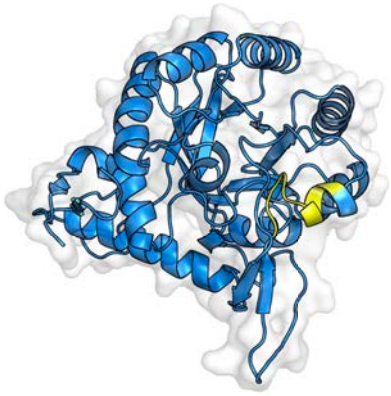


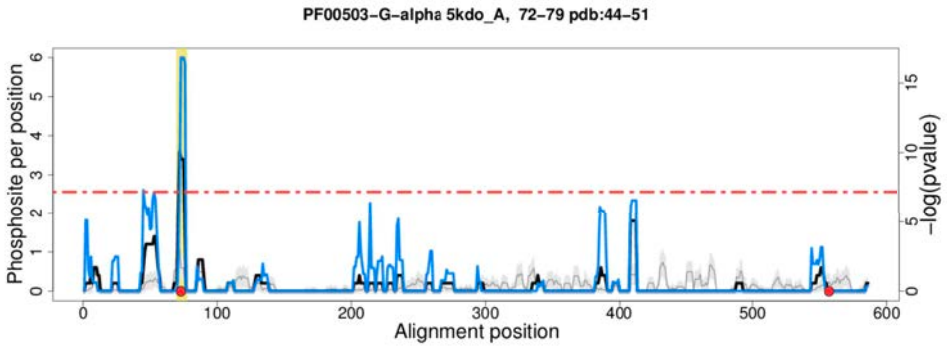

PF00433-Pkinase_C 4q9z_A, 18-26 pdb:672-675
Phosphosite per position
-log(pvalue)
Alignment position

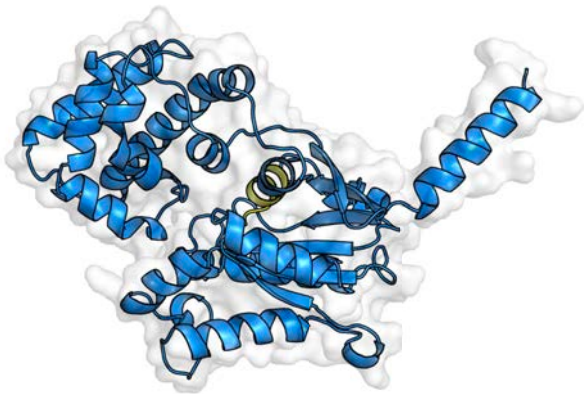


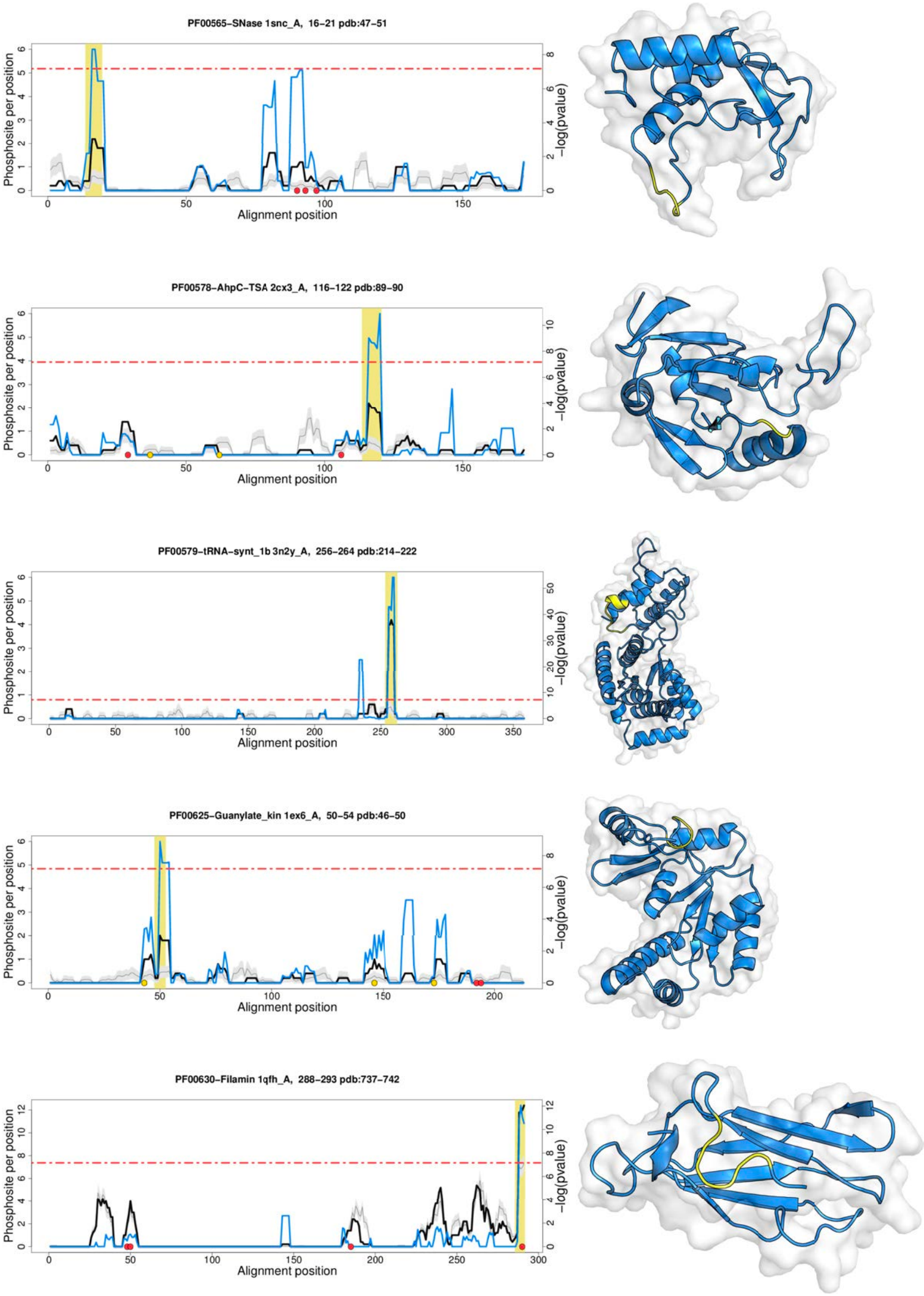

PF00443-UCH 5j7t_A, 1064-1072,1176-1183,2978-2987,3021-3025,3094-3098,3239-3243 pdb:NA,NA,NA,NA,NA,NA
Phosphosite per position
-log(pvalue)
Alignment position

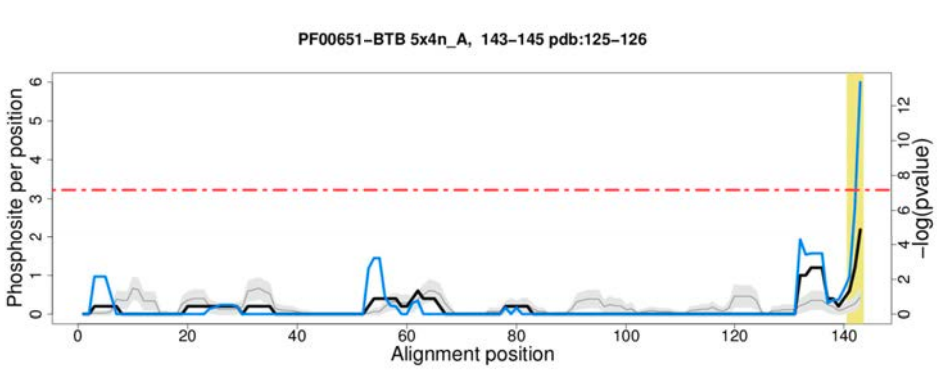

PF00478-IMPDH 2a7r_A, 95-100,394-402 pdb:NA,267-275
Phosphosite per position
-log(pvalue)
Alignment position

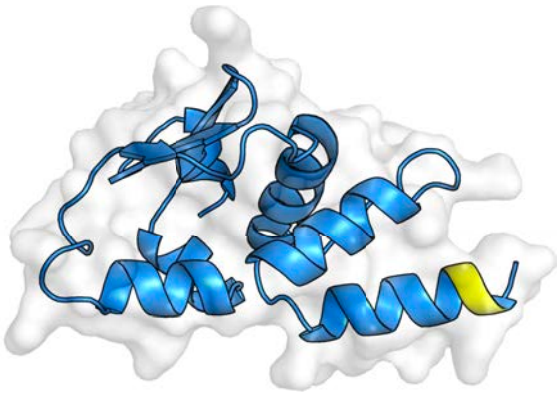


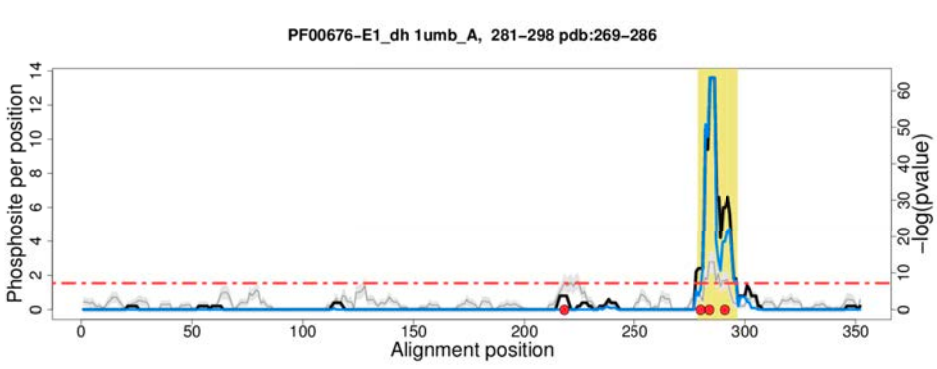

PF00503-G-alpha 5kdo_A, 72-79 pdb:44-51
Phosphosite per position
-log(pvalue)
Alignment position

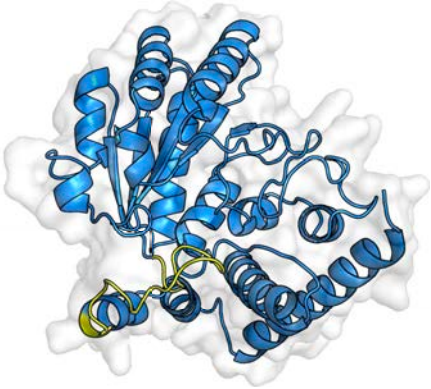


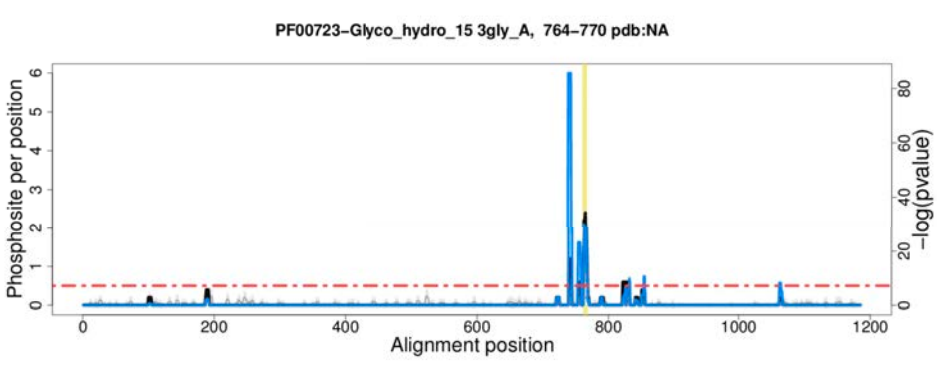

PF00565-SNase 1snc_A, 16-21 pdb:47-51
Phosphosite per position
-log(pvalue)
Alignment position
PF00578-AhpC-TSA 2cx3_A, 116-122 pdb:89-90
Phosphosite per position
-log(pvalue)
Alignment position
PF00579-tRNA-synt_1b 3n2y_A, 256-264 pdb:214-222
Phosphosite per position
-log(pvalue)
Alignment position
PF00625-Guanylate_kin 1ex6_A, 50-54 pdb:46-50
Phosphosite per position
-log(pvalue)
Alignment position
PF00630-Filamin 1qfh_A, 288-293 pdb:737-742
Phosphosite per position
-log(pvalue)
Alignment position

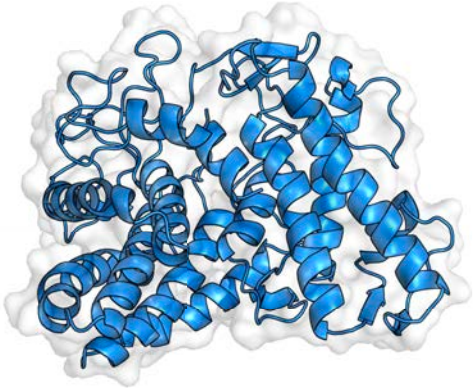

PF00651-BTB 5x4n_A, 143-145 pdb:125-126
Phosphosite per position
-log(pvalue)
Alignment position

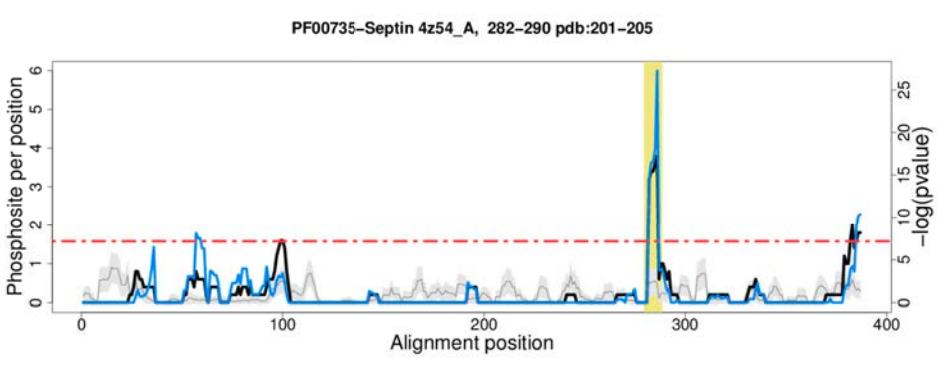


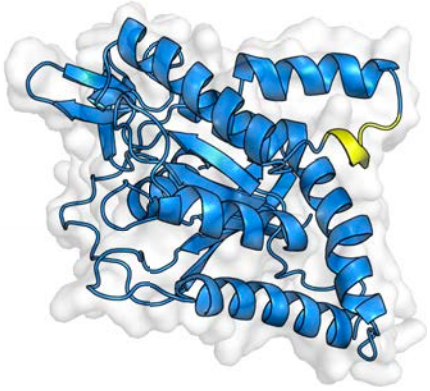

PF00676-E1_dh 1umb_A, 281-298 pdb:269-286
Phosphosite per position
-log(pvalue)
Alignment position

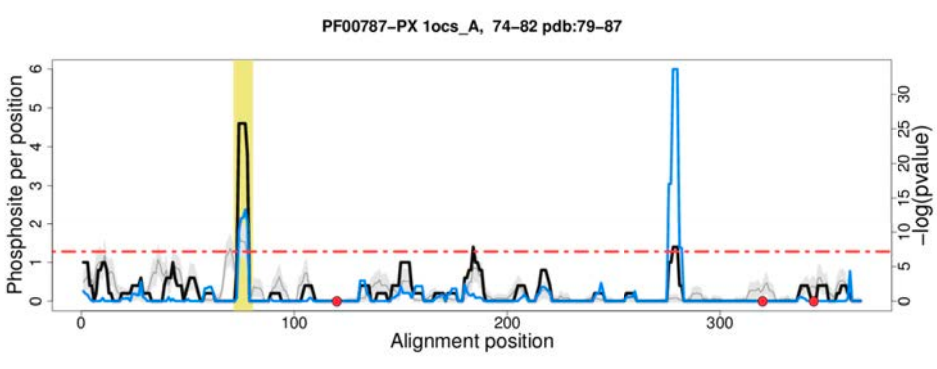


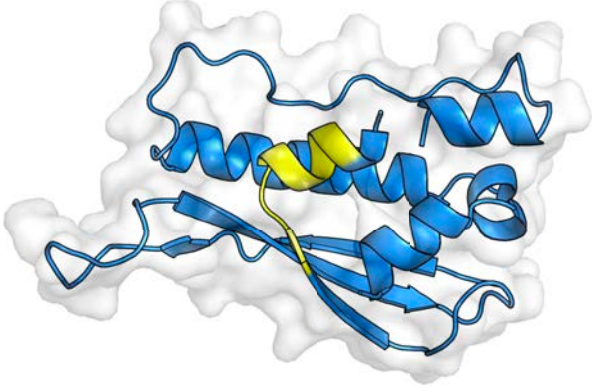

PF00723-Glyco_hydro_15 3gly_A, 764-770 pdb:NA
Phosphosite per position
-log(pvalue)
Alignment position

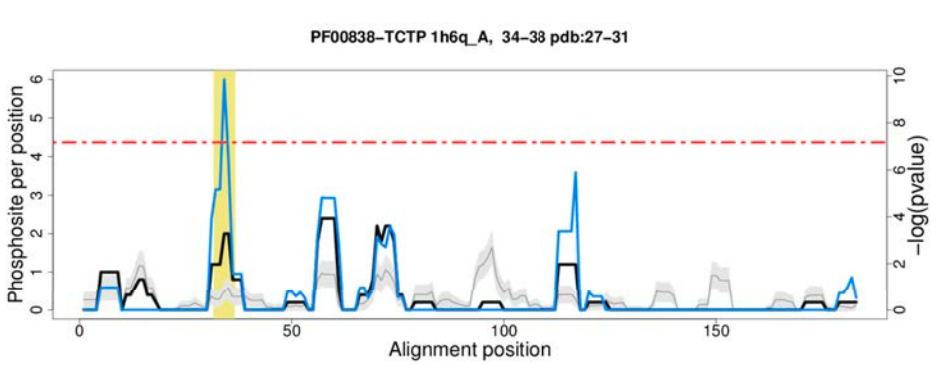


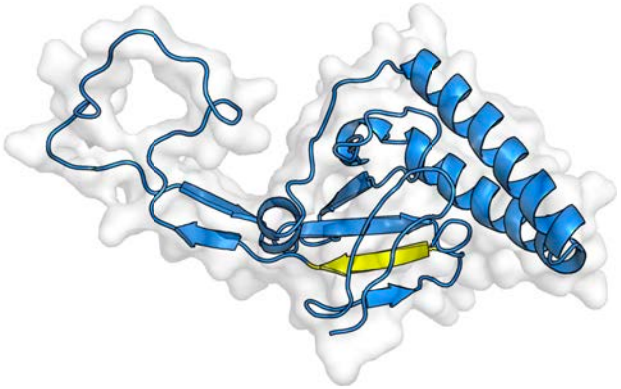

PF00735-Septin 4z54_A, 282-290 pdb:201-205
Phosphosite per position
-log(pvalue)
Alignment position

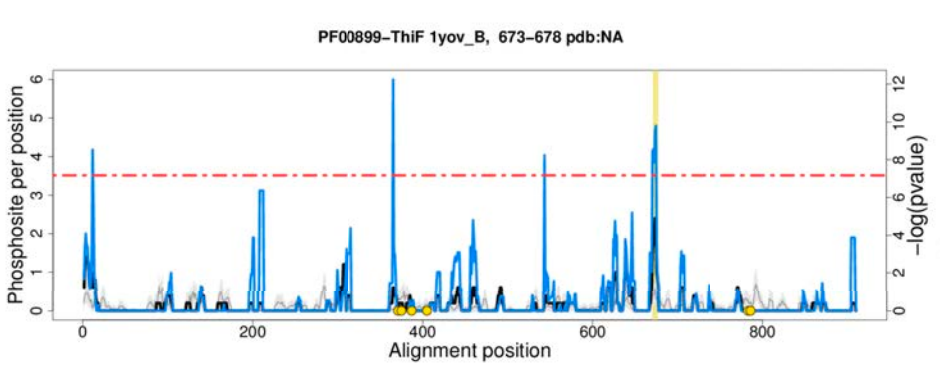


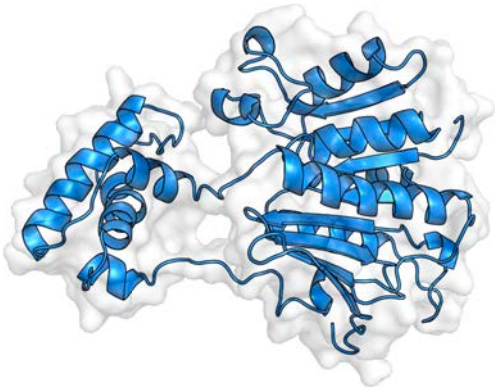

PF00787-PX 1ocs_A, 74-82 pdb:79-87
Phosphosite per position
-log(pvalue)
Alignment position

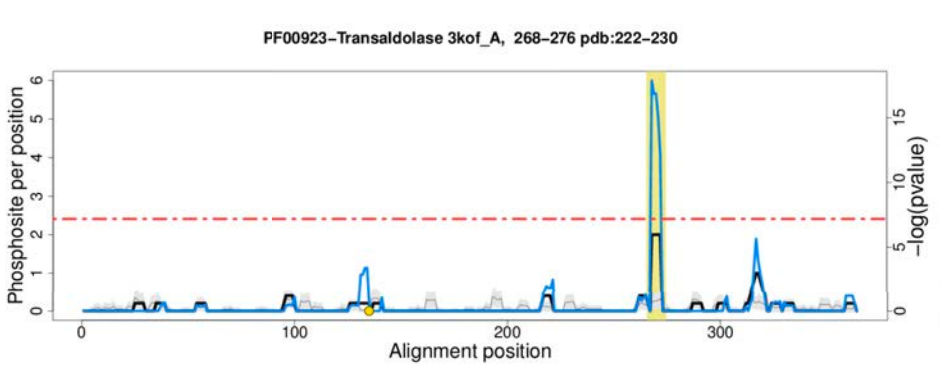


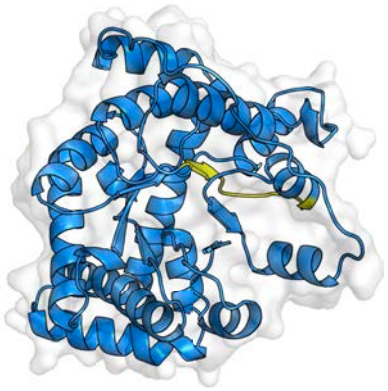

PF00838–TCTP 1h6q_A, 34–38 pdb:27–31
Phosphosite per position
−log(pvalue)
Alignment position

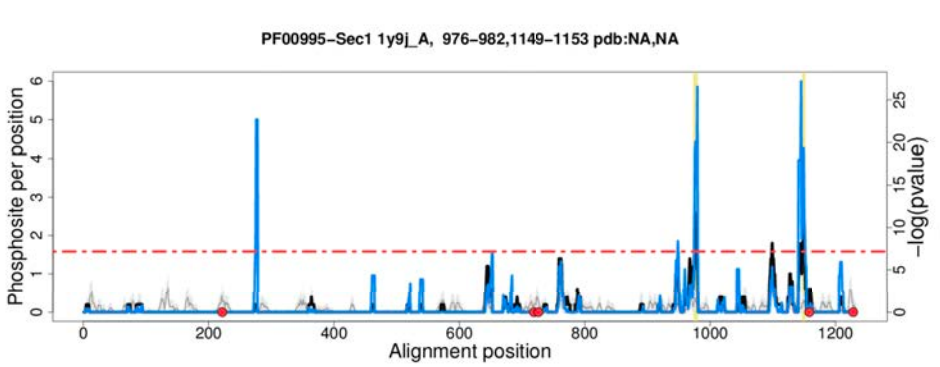


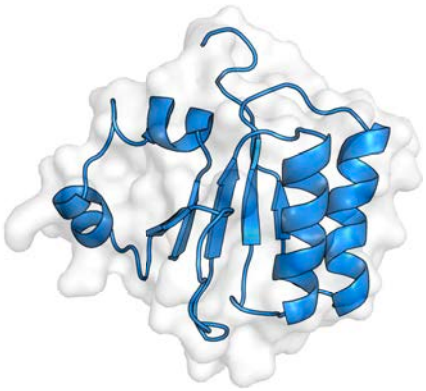

PF00899–ThiF 1yov_B, 673–678 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position

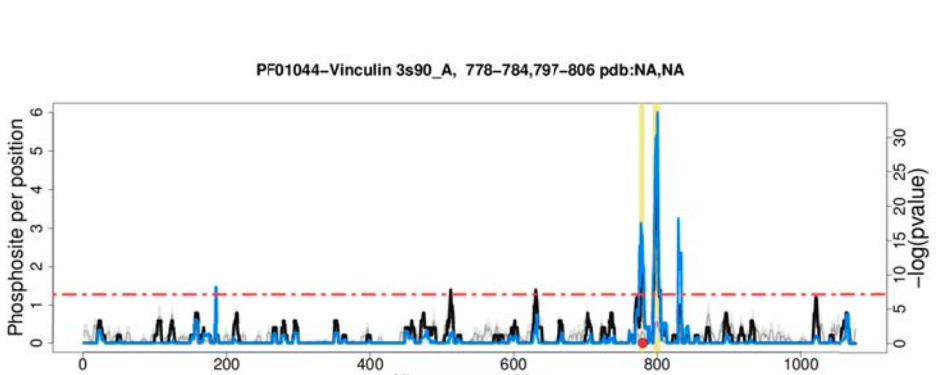


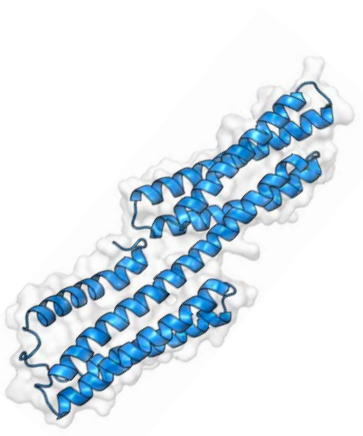

PF00923–Transaldolase 3kof_A, 268–276 pdb:222–230
Phosphosite per position
−log(pvalue)
Alignment position

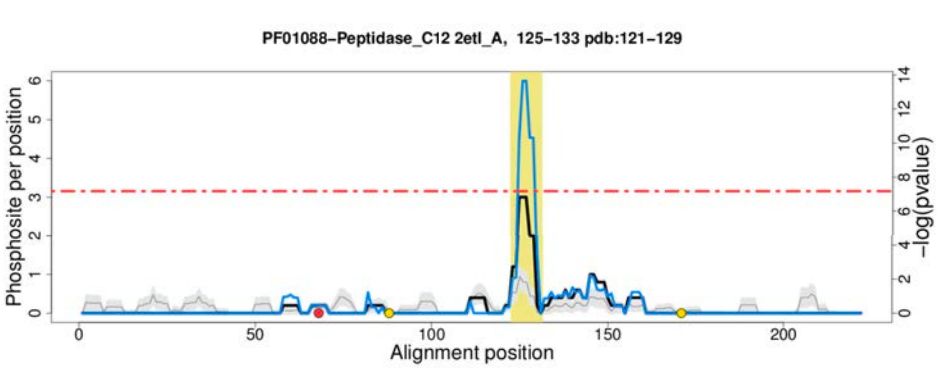


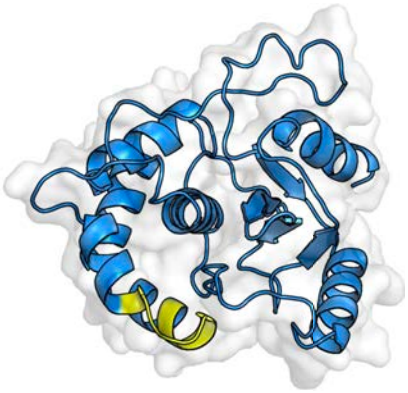

PF00995–Sec1 1y9j_A, 976–982,1149–1153 pdb:NA,NA
Phosphosite per position
−log(pvalue)
Alignment position

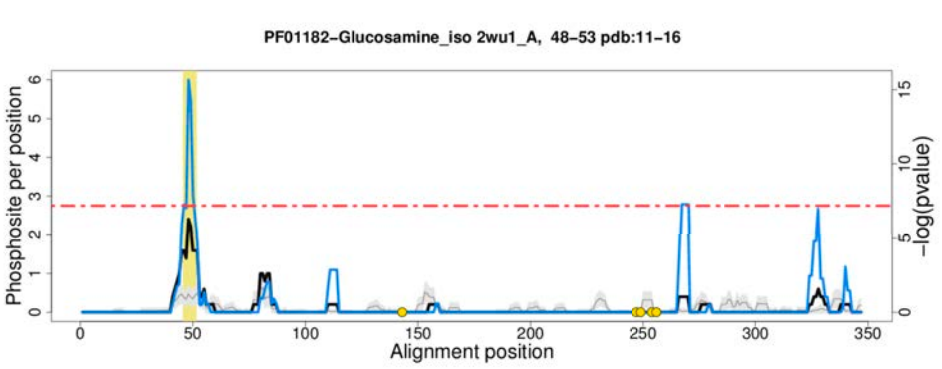


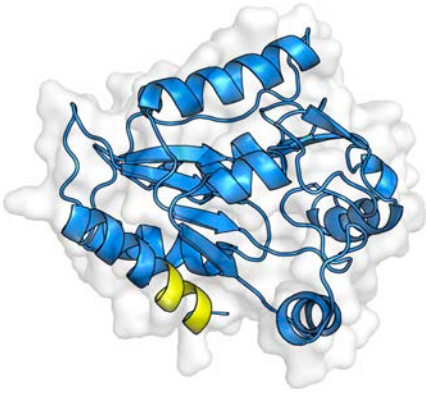

PF01044–Vinculin 3s90_A, 778–784,797–806 pdb:NA,NA
Phosphosite per position
−log(pvalue)
Alignment position

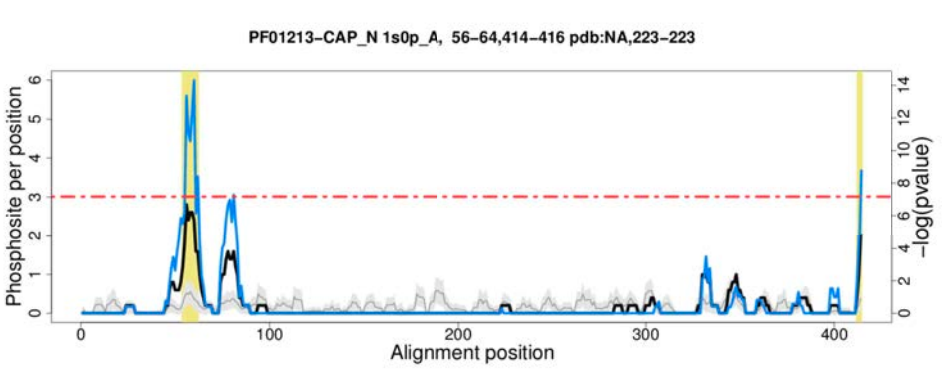


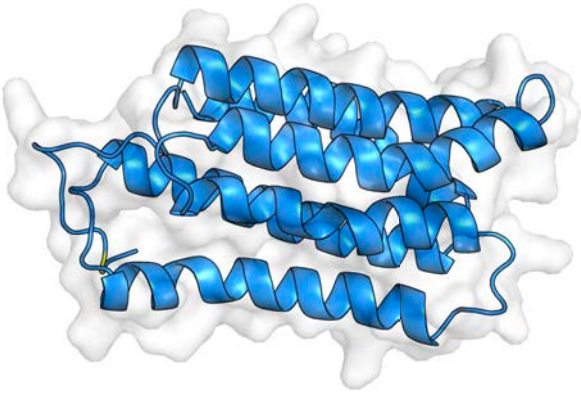

PF01088-Peptidase_C12 2etl_A, 125-133 pdb:121-129
Phosphosite per position
-log(pvalue)
Alignment position

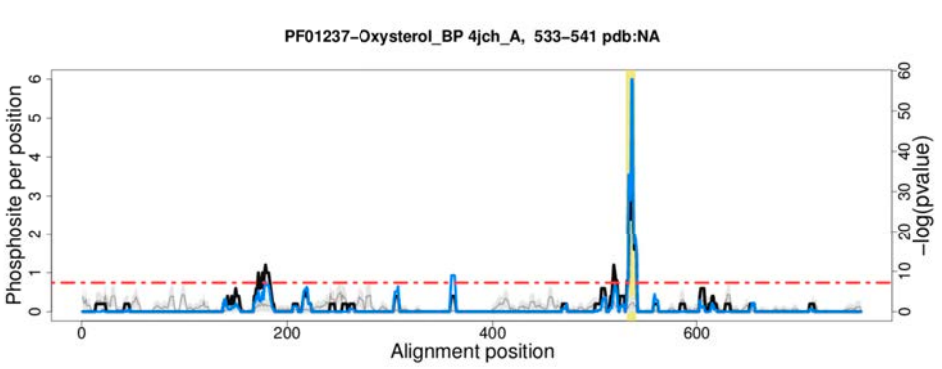


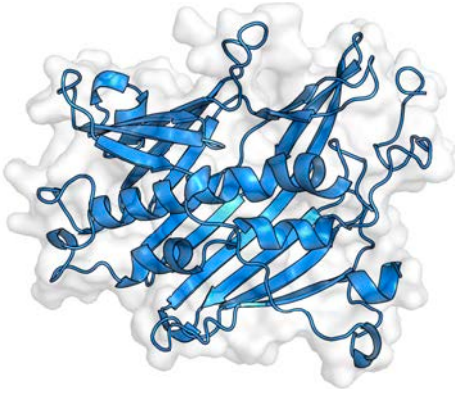

PF01182-Glucosamine_iso 2wu1_A, 48-53 pdb:11-16
Phosphosite per position
-log(pvalue)
Alignment position

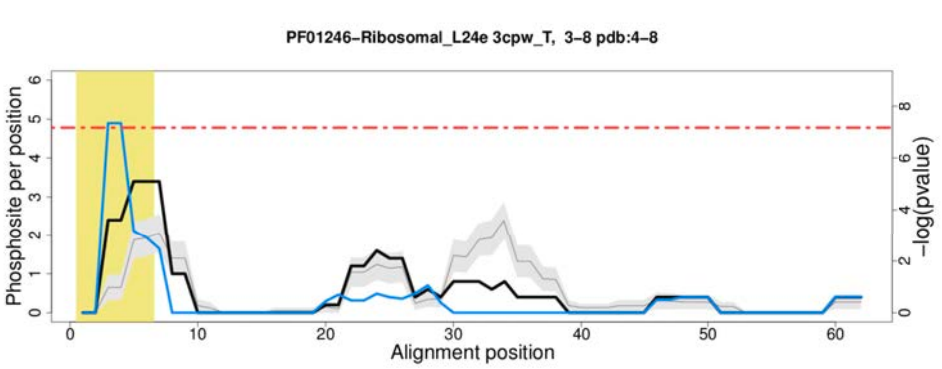


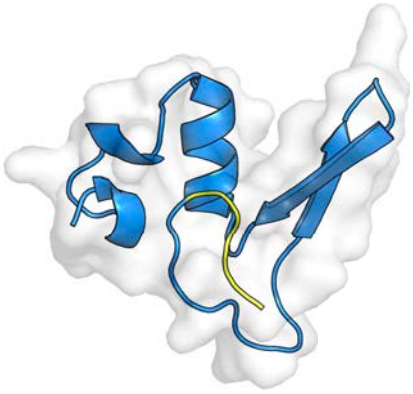

PF01213-CAP_N 1s0p_A, 56-64,414-416 pdb:NA,223-223
Phosphosite per position
-log(pvalue)
Alignment position

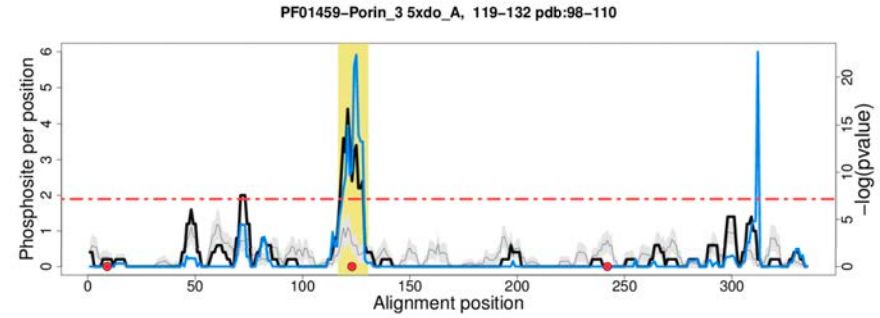


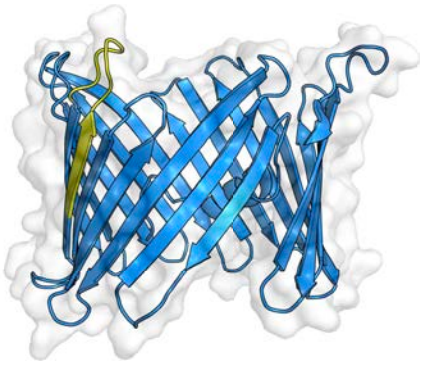

PF01237-Oxysterol_BP 4jch_A, 533-541 pdb:NA
Phosphosite per position
-log(pvalue)
Alignment position

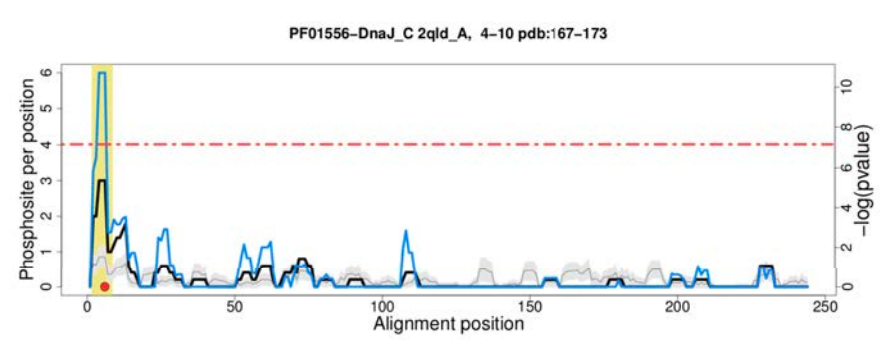


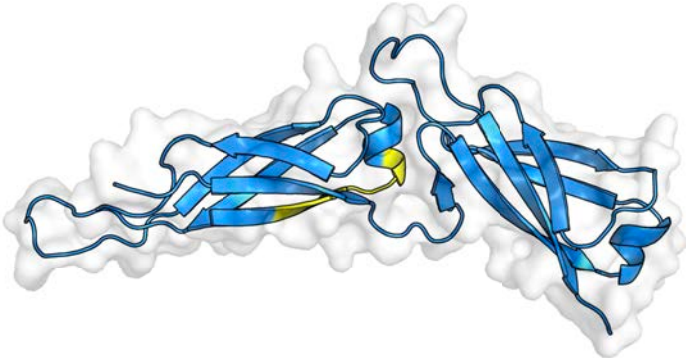

PF01246-Ribosomal_L24e 3cpw_T, 3-8 pdb:4-8
Phosphosite per position
-log(pvalue)
Alignment position

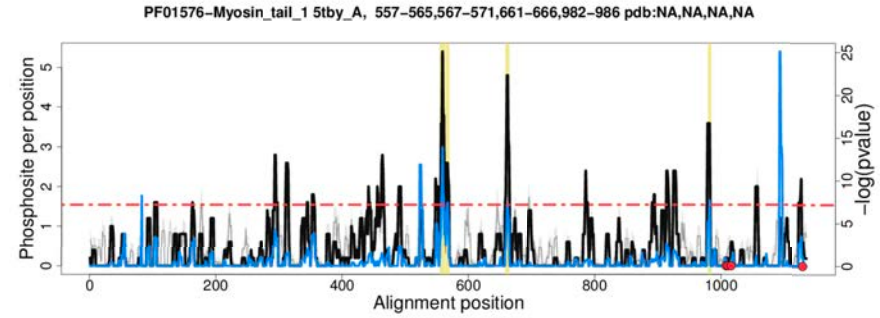


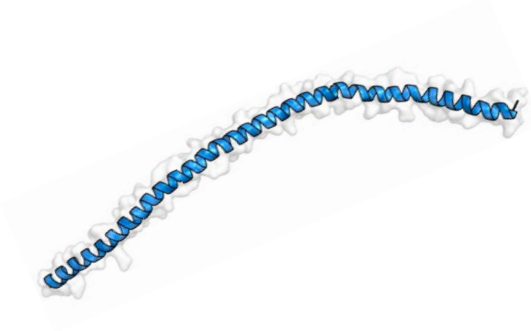

PF01459-Porin_3 5xdo_A, 119-132 pdb:98-110
Phosphosite per position
−log(pvalue)
Alignment position

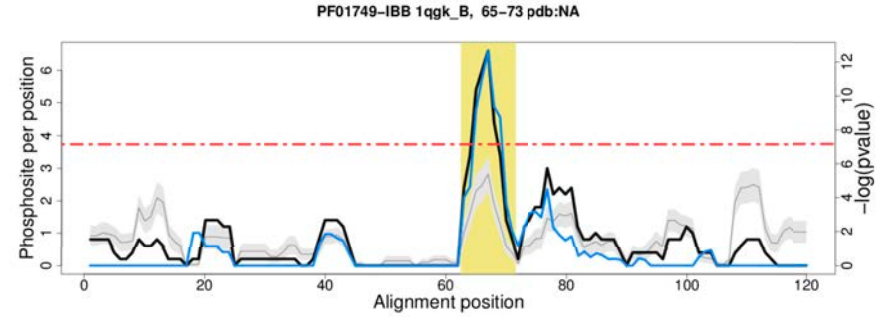


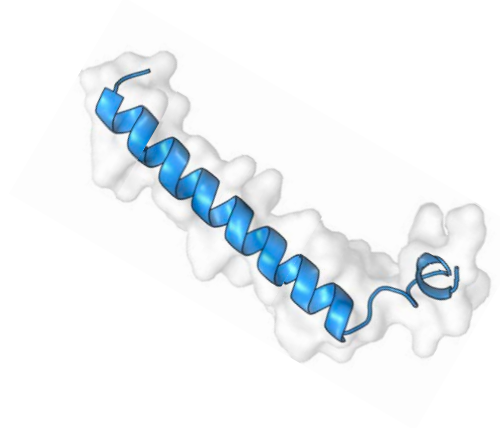

PF01556-DnaJ_C 2qld_A, 4-10 pdb:167-173
Phosphosite per position
−log(pvalue)
Alignment position

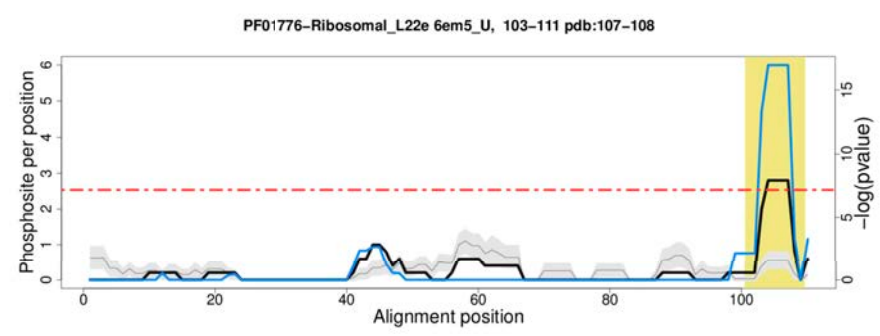


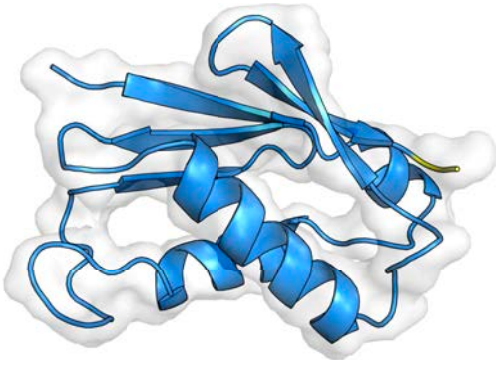

PF01576-Myosin_tail_1 5tby_A, 557-565,567-571,661-666,982-986 pdb:NA,NA,NA,NA
Phosphosite per position
−log(pvalue)
Alignment position

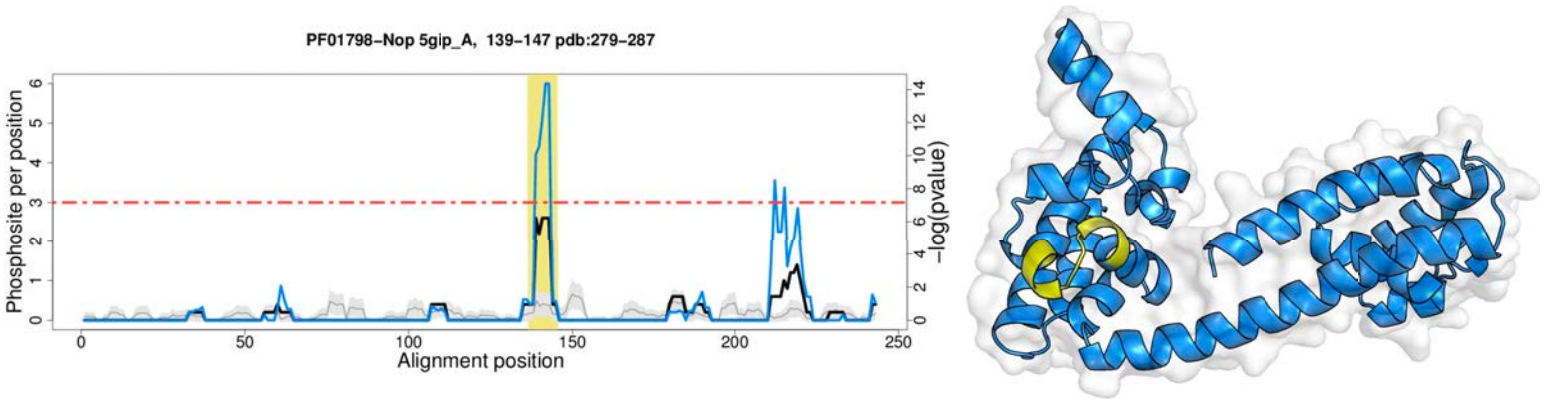


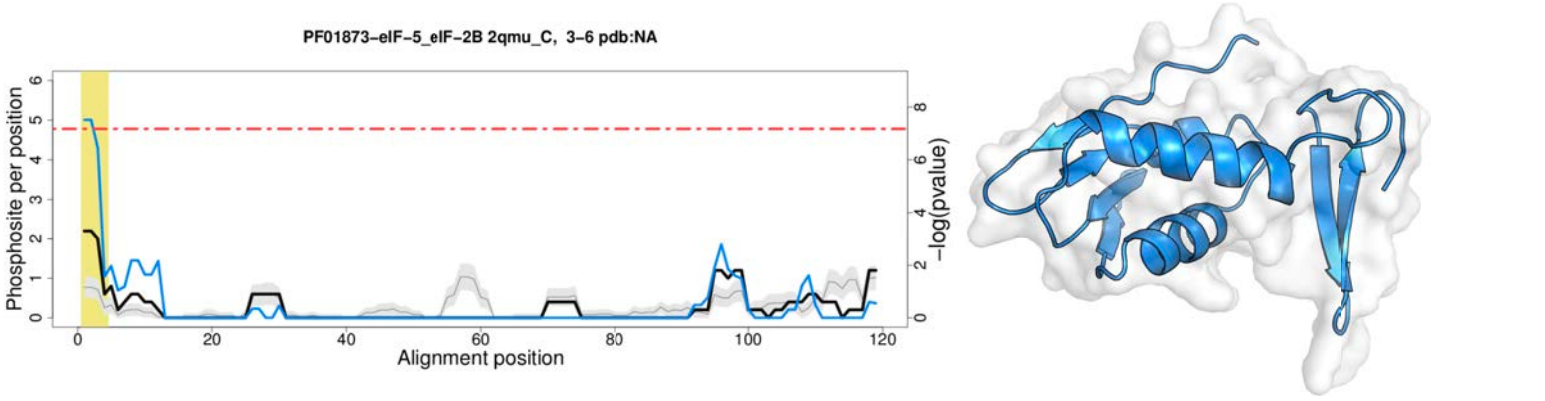

PF01749-IBB 1qgk_B, 65-73 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position

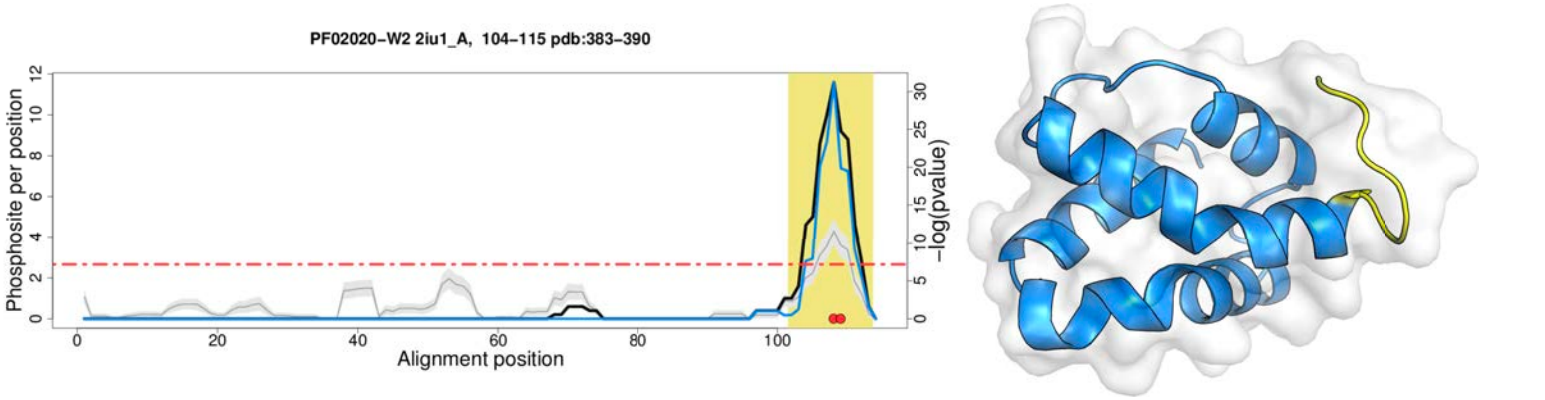


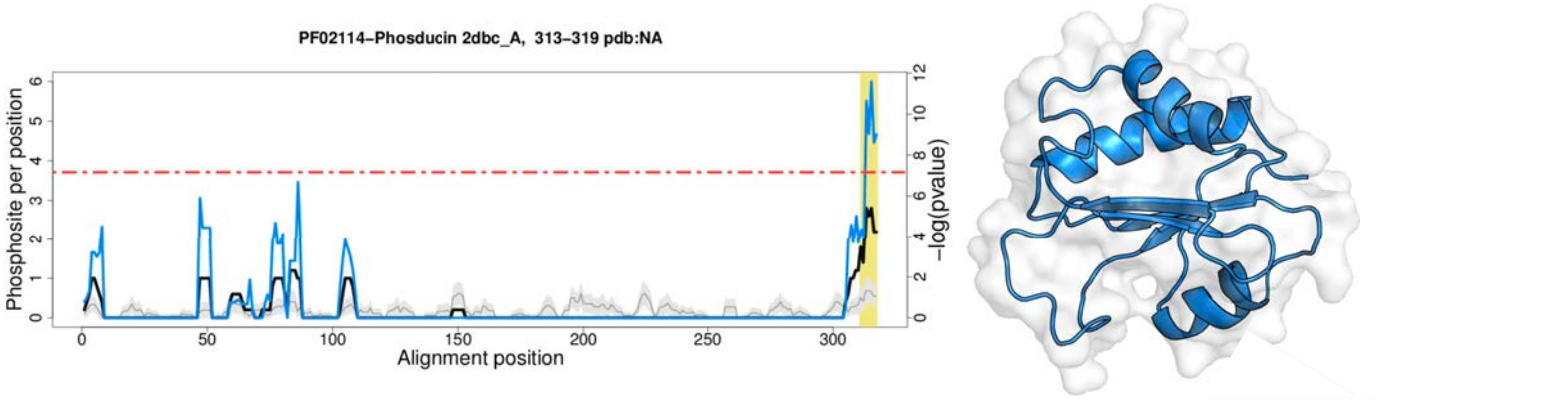

PF01776-Ribosomal_L22e 6em5_U, 103-111 pdb:107-108
Phosphosite per position
−log(pvalue)
Alignment position

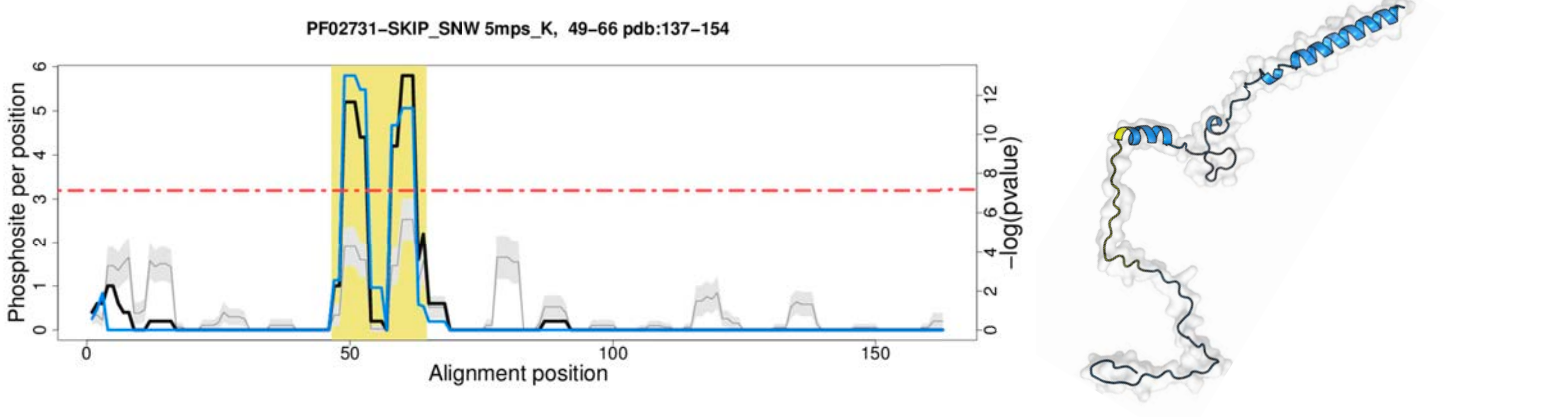


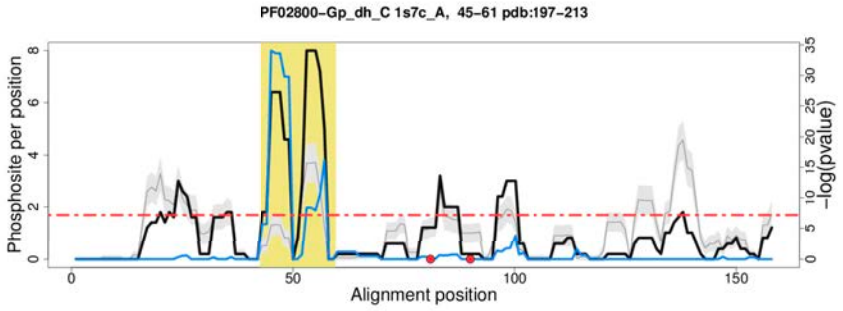

PF01798–Nop 5gip_A, 139–147 pdb:279–287
Phosphosite per position
–log(pvalue)
Alignment position

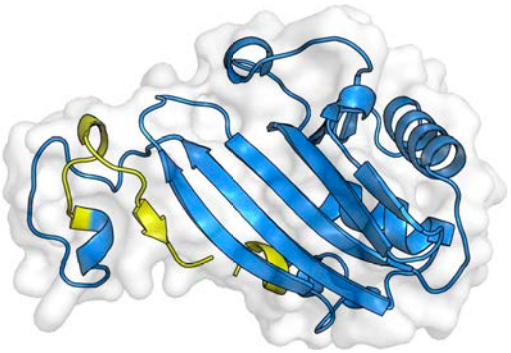

PF01873–eIF–5_eIF–2B 2qmu_C, 3–6 pdb:NA
Phosphosite per position
–log(pvalue)
Alignment position

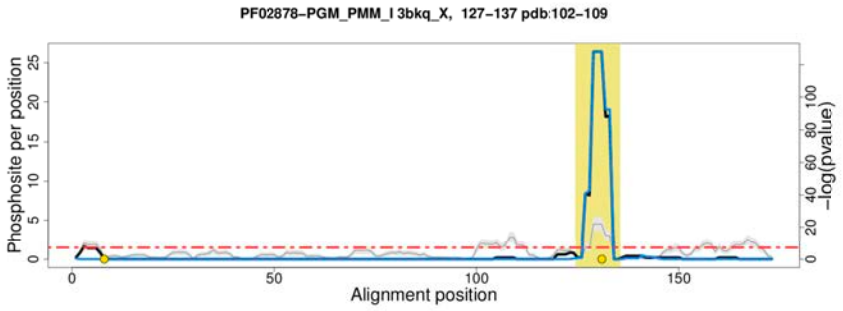

PF02020–W2 2iu1_A, 104–115 pdb:383–390
Phosphosite per position
–log(pvalue)
Alignment position

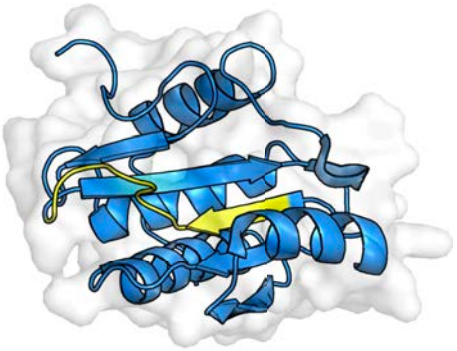

PF02114–Phosducin 2dbc_A, 313–319 pdb:NA
Phosphosite per position
–log(pvalue)
Alignment position

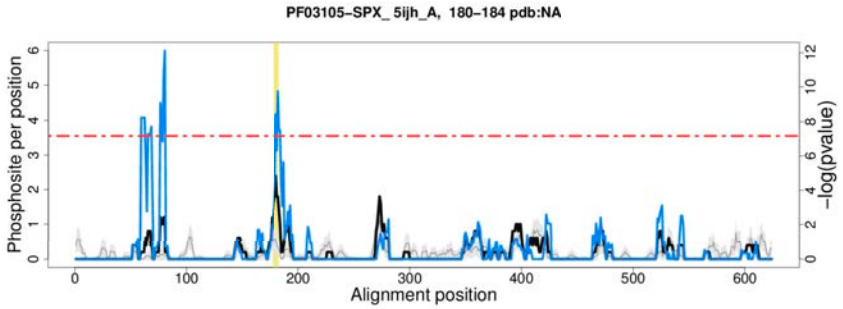

PF02731–SKIP_SNW 5mps_K, 49–66 pdb:137–154
Phosphosite per position
–log(pvalue)
Alignment position

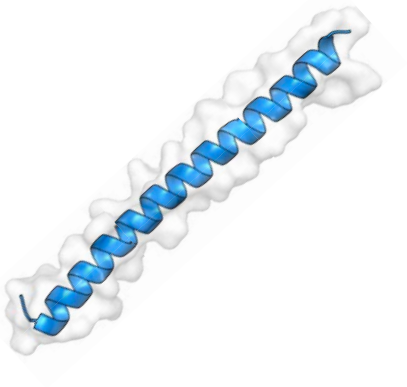

PF02800–Gp_dh_C 1s7c_A, 45–61 pdb:197–213
Phosphosite per position
–log(pvalue)
Alignment position

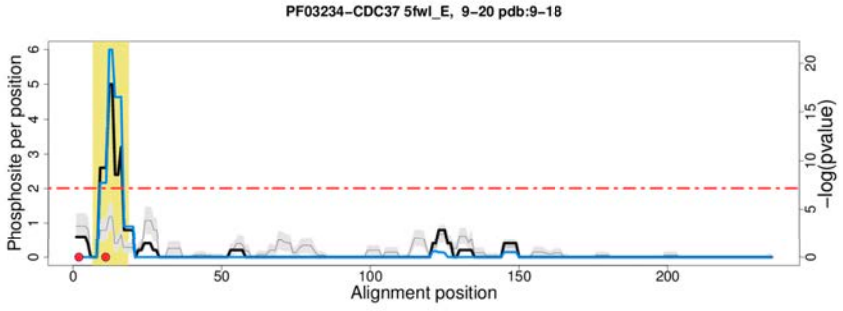


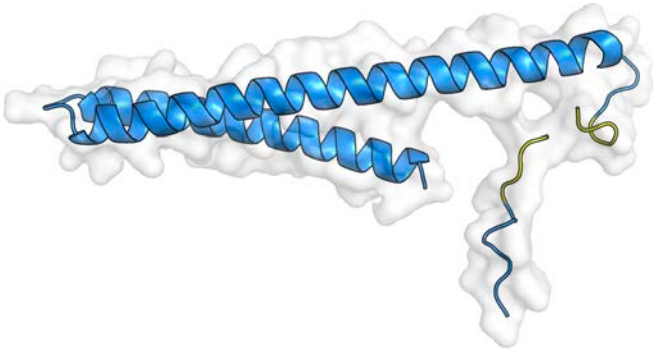

PF02878–PGM_PMM_I 3bkq_X, 127–137 pdb:102–109
Phosphosite per position
–log(pvalue)
Alignment position

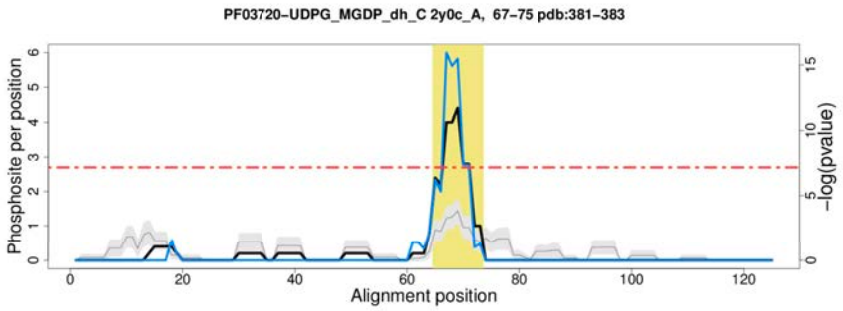


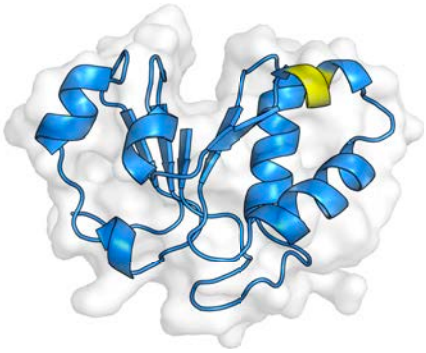

PF03105–SPX_ 5ijh_A, 180–184 pdb:NA
Phosphosite per position
–log(pvalue)
Alignment position

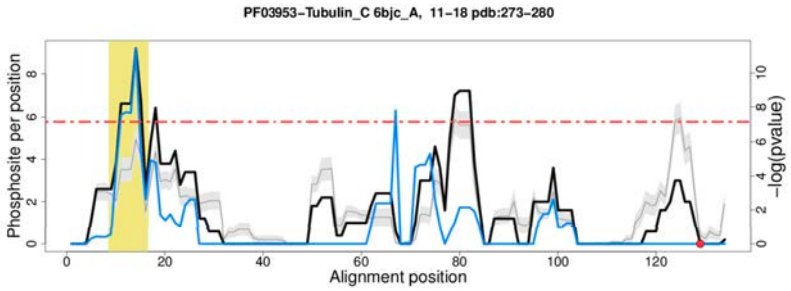


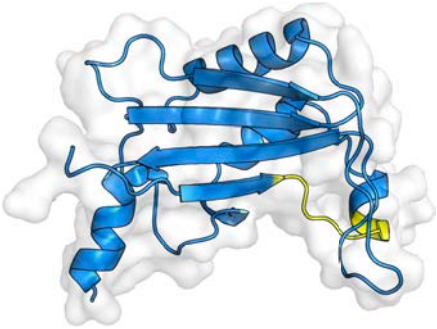

PF03234–CDC37 5fwl_E, 9–20 pdb:9–18
Phosphosite per position
–log(pvalue)
Alignment position

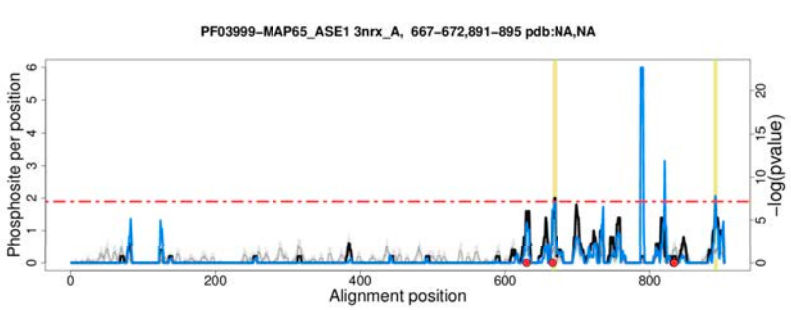


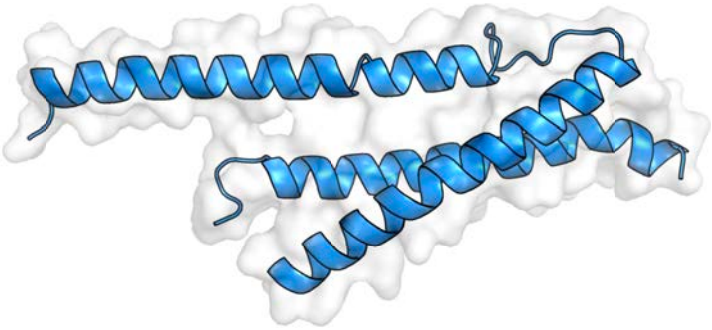

PF03720–UDPG_MGDP_dh_C 2y0c_A, 67–75 pdb:381–383
Phosphosite per position
–log(pvalue)
Alignment position

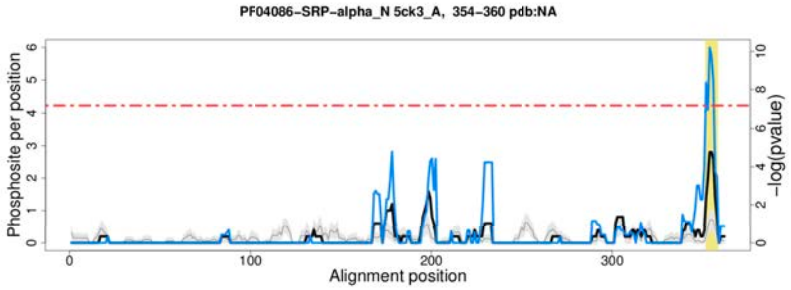


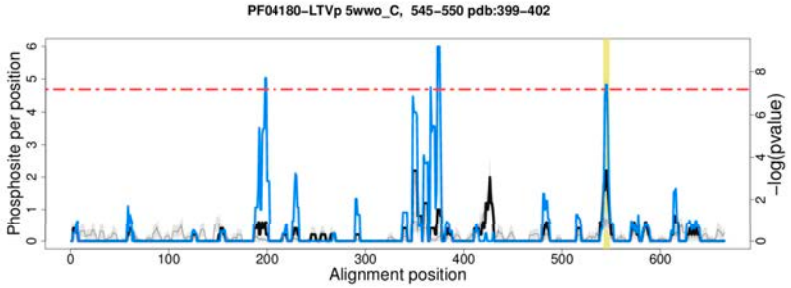

PF03953-Tubulin_C 6bjc_A, 11-18 pdb:273-280
Phosphosite per position
-log(pvalue)
Alignment position

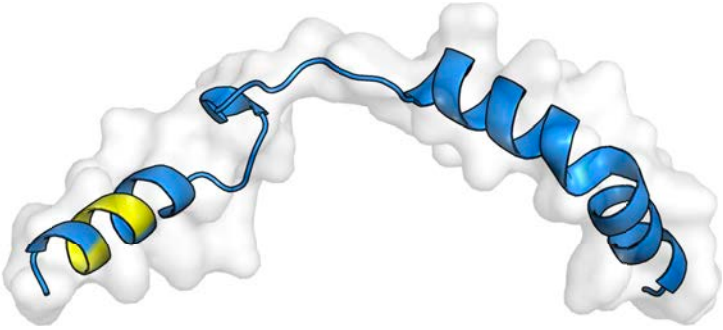


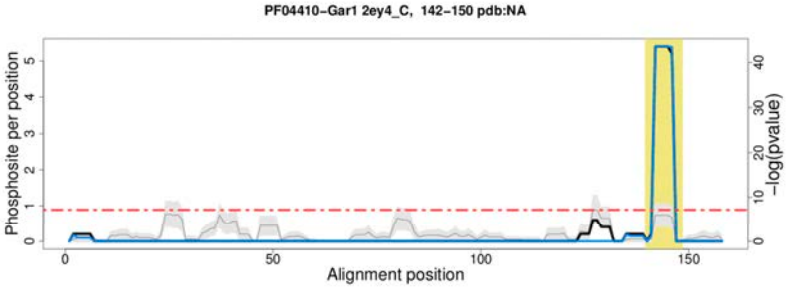

PF03999-MAP65_ASE1 3nrx_A, 667-672,891-895 pdb:NA,NA
Phosphosite per position
-log(pvalue)
Alignment position

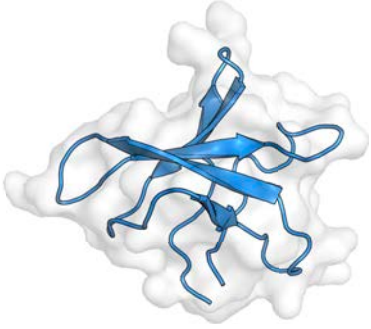


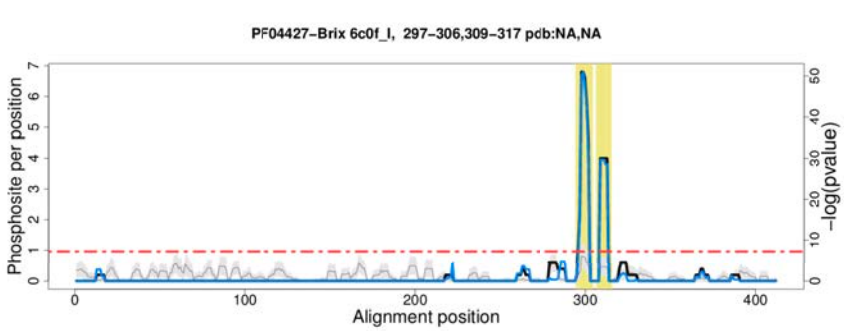

PF04086-SRP-alpha_N 5ck3_A, 354-360 pdb:NA
Phosphosite per position
-log(pvalue)
Alignment position

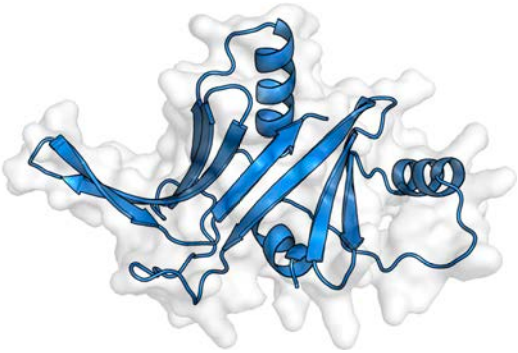

PF04180-LTVp 5wwo_C, 545-550 pdb:399-402
Phosphosite per position
-log(pvalue)
Alignment position

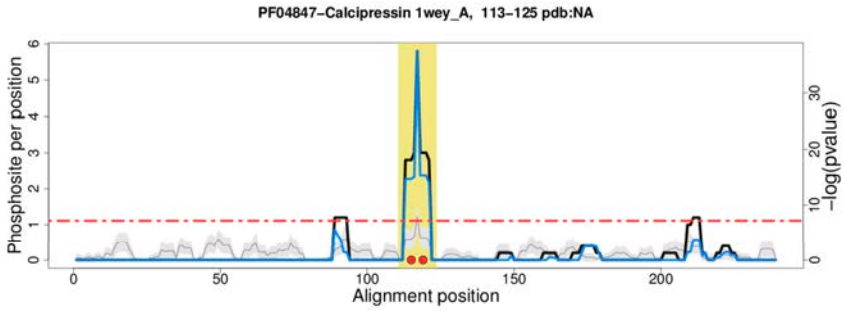


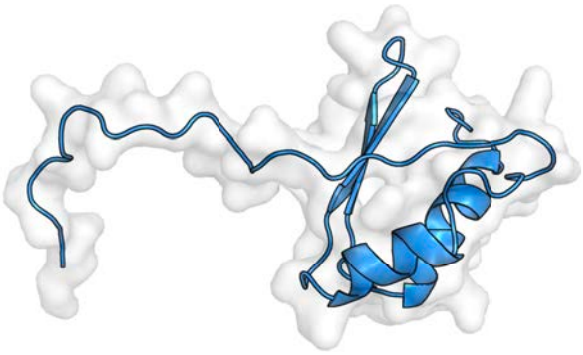

PF04410-Gar1 2ey4_C, 142-150 pdb:NA
Phosphosite per position
-log(pvalue)
Alignment position

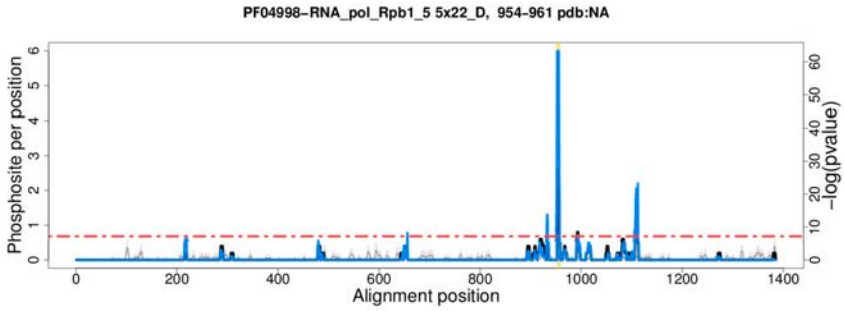


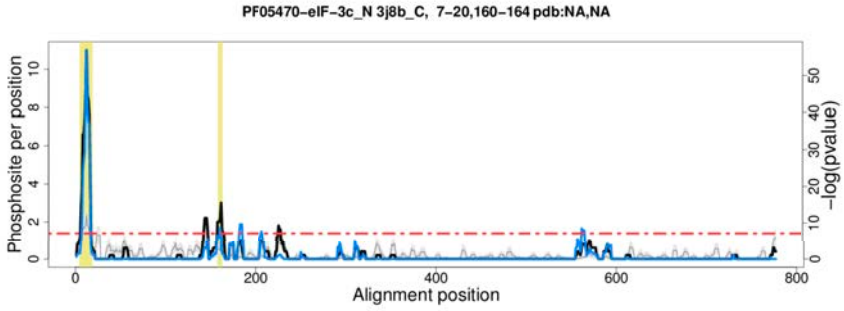

PF04427–Brix 6c0f_I, 297–306,309–317 pdb:NA,NA
Phosphosite per position
–log(pvalue)
Alignment position

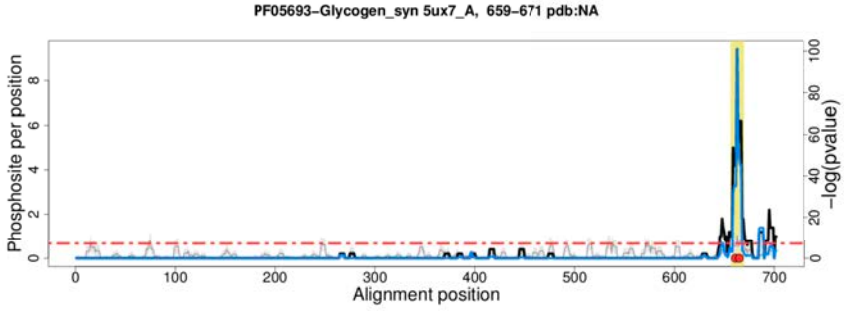


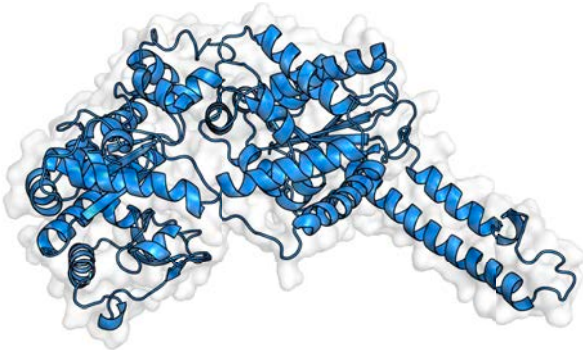

PF04847–Calcipressin 1wey_A, 113–125 pdb:NA
Phosphosite per position
–log(pvalue)
Alignment position

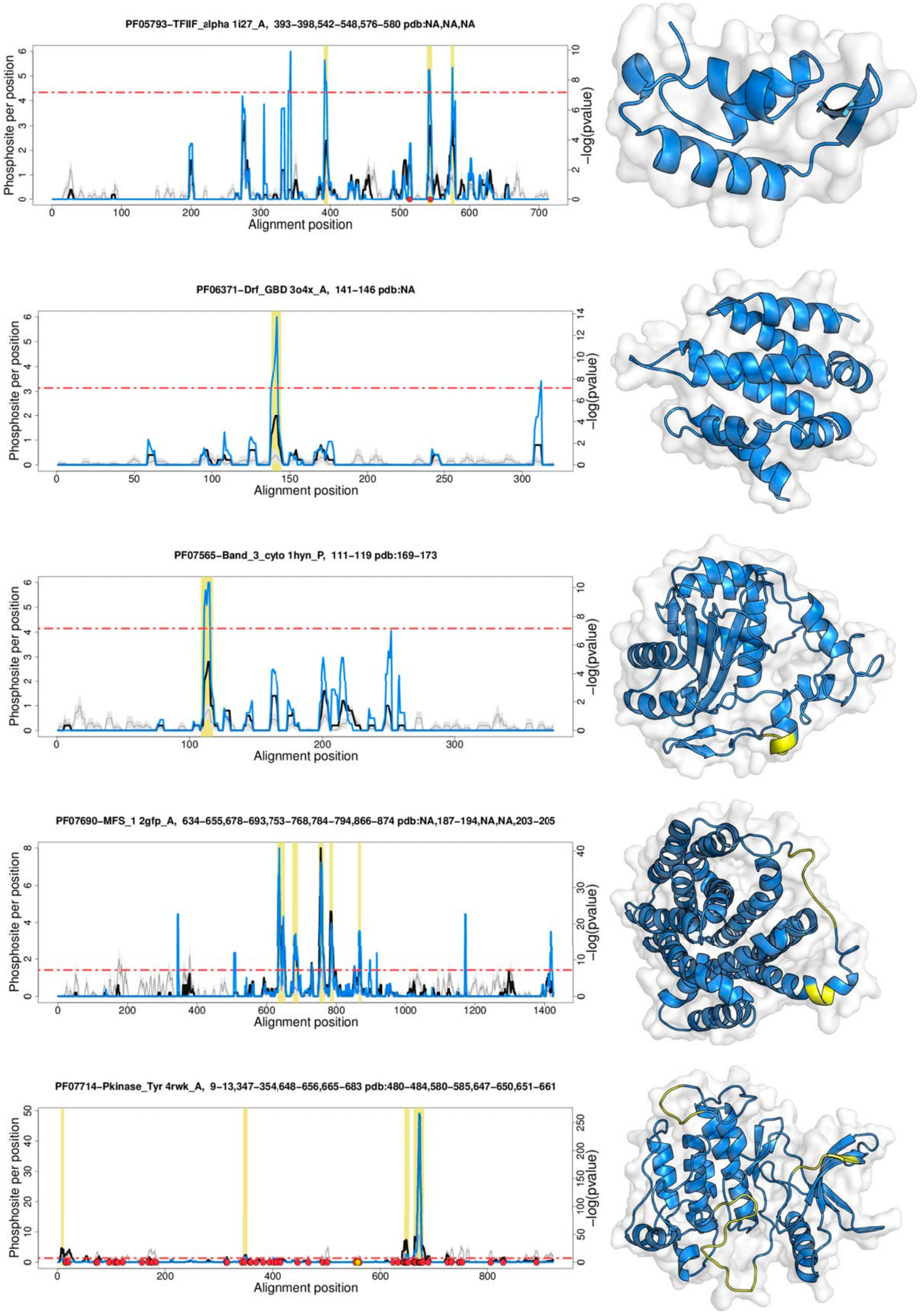


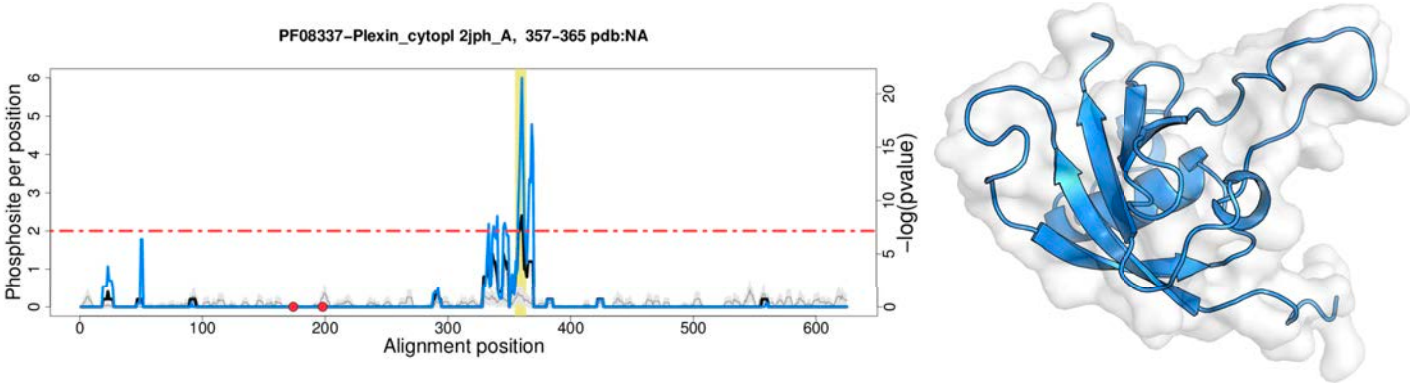

PF04998–RNA_pol_Rpb1_5 5x22_D, 954–961 pdb:NA
Phosphosite per position
–log(pvalue)
Alignment position

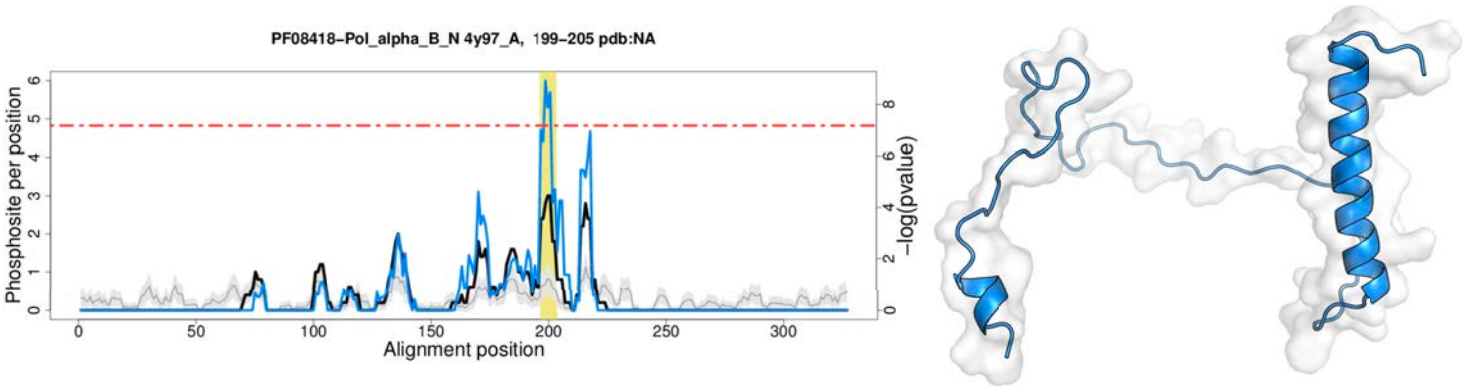

PF05470–eIF–3c_N 3j8b_C, 7–20,160–164 pdb:NA,NA
Phosphosite per position
–log(pvalue)
Alignment position

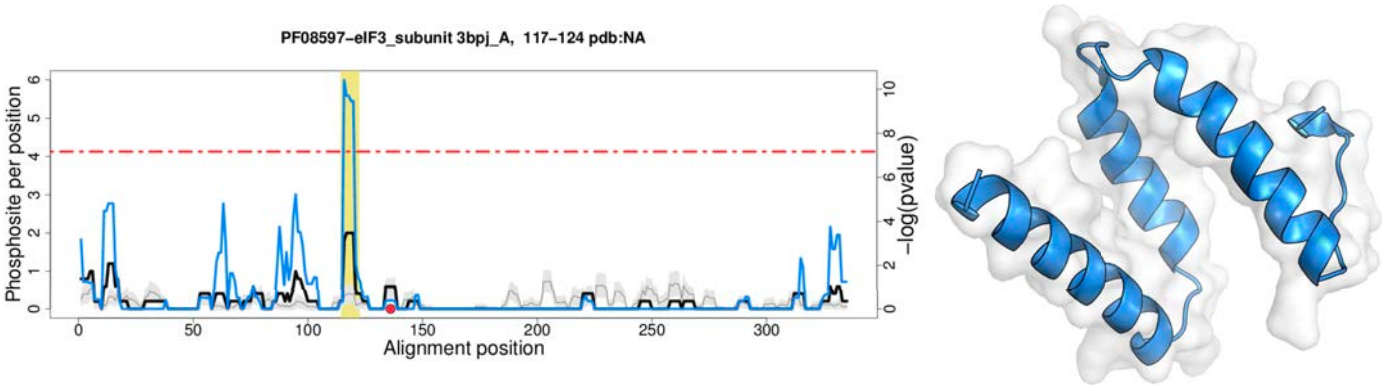

PF05693–Glycogen_syn 5ux7_A, 659–671 pdb:NA
Phosphosite per position
–log(pvalue)
Alignment position

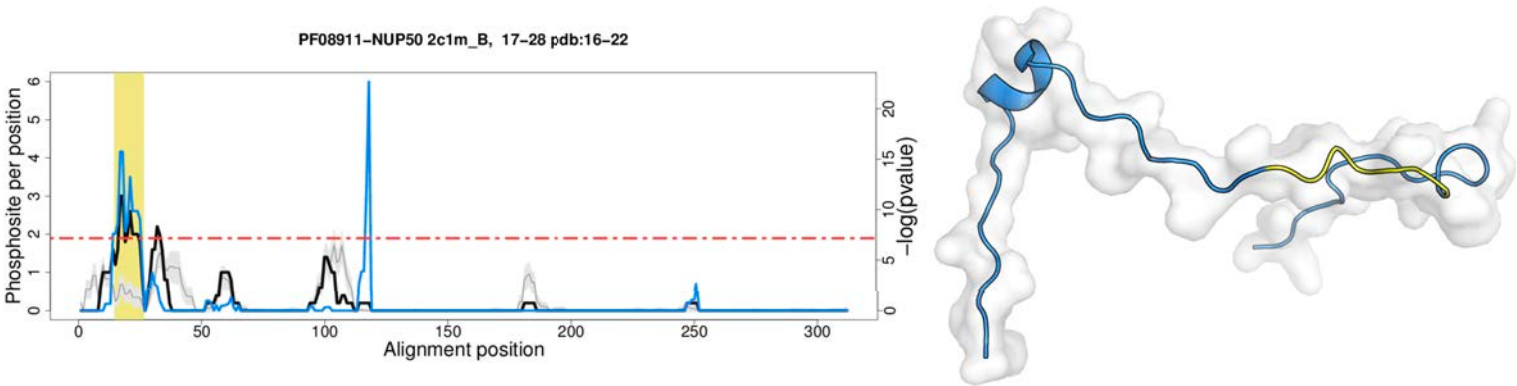


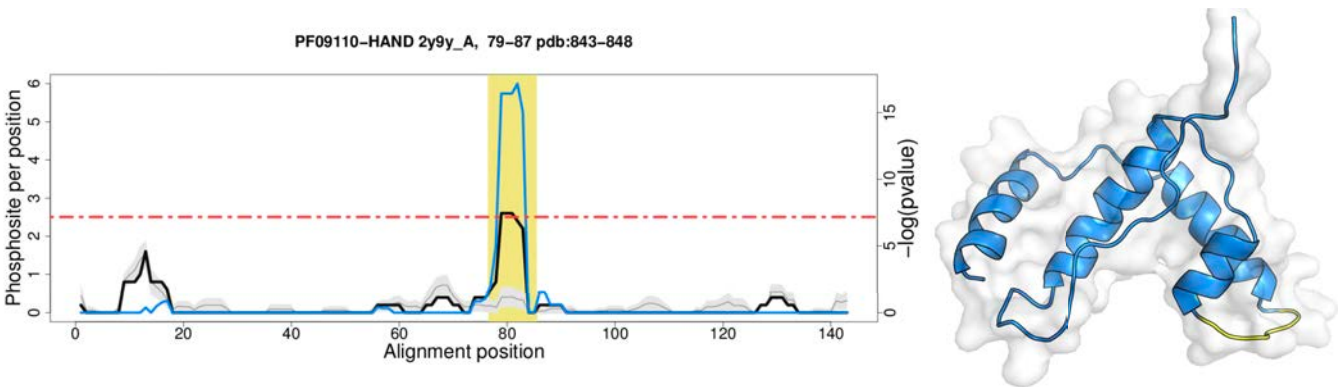

PF05793-TFIIF_alpha 1i27_A, 393-398,542-548,576-580 pdb:NA,NA,NA
Phosphosite per position
-log(pvalue)
Alignment position
PF06371-Drf_GBD 3o4x_A, 141-146 pdb:NA
Phosphosite per position
-log(pvalue)
Alignment position
PF07565-Band_3_cyto 1hyn_P, 111-119 pdb:169-173
Phosphosite per position
-log(pvalue)
Alignment position
PF07690-MFS_1 2gfp_A, 634-655,678-693,753-768,784-794,866-874 pdb:NA,187-194,NA,NA,203-205
Phosphosite per position
-log(pvalue)
Alignment position
PF07714-Pkinase_Tyr 4rwk_A, 9-13,347-354,648-656,665-683 pdb:480-484,580-585,647-650,651-661
Phosphosite per position
-log(pvalue)
Alignment position

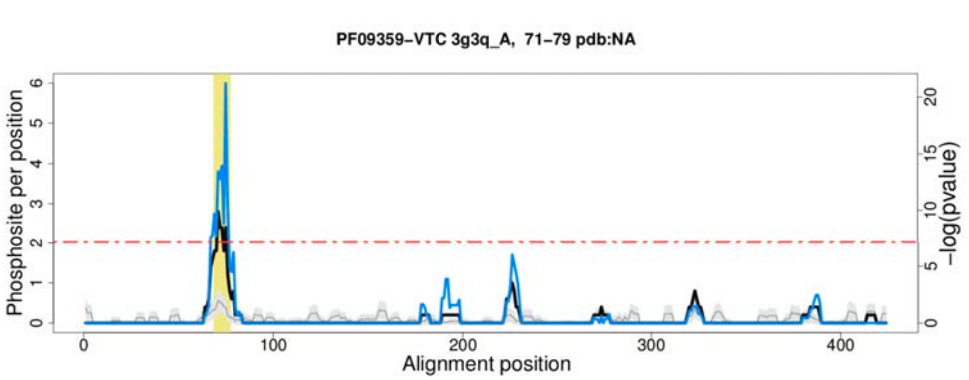

PF08337–Plexin_cytopl 2jph_A, 357–365 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position

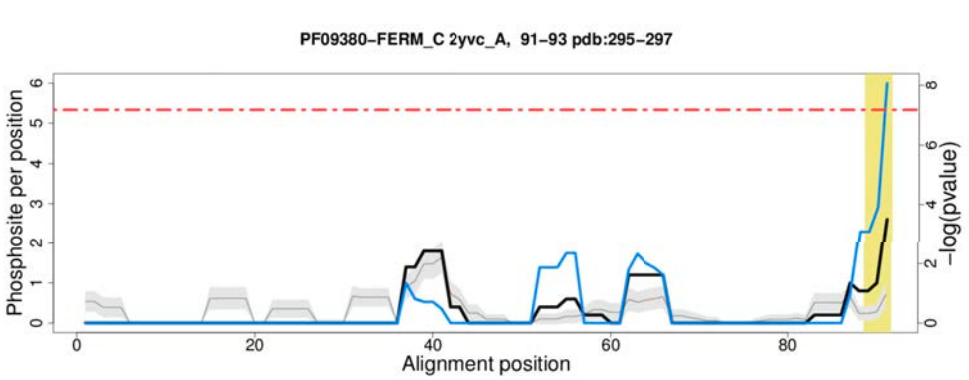

PF08418–Pol_alpha_B_N 4y97_A, 199–205 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position

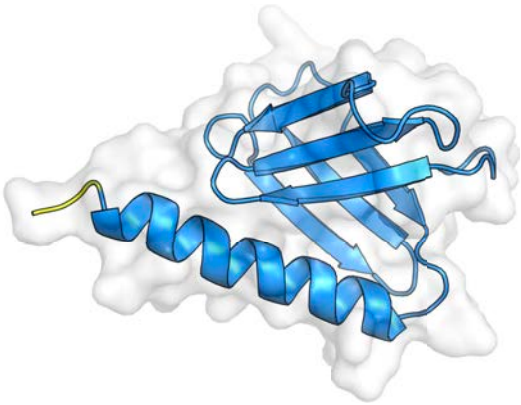

PF08597–eIF3_subunit 3bpj_A, 117–124 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position

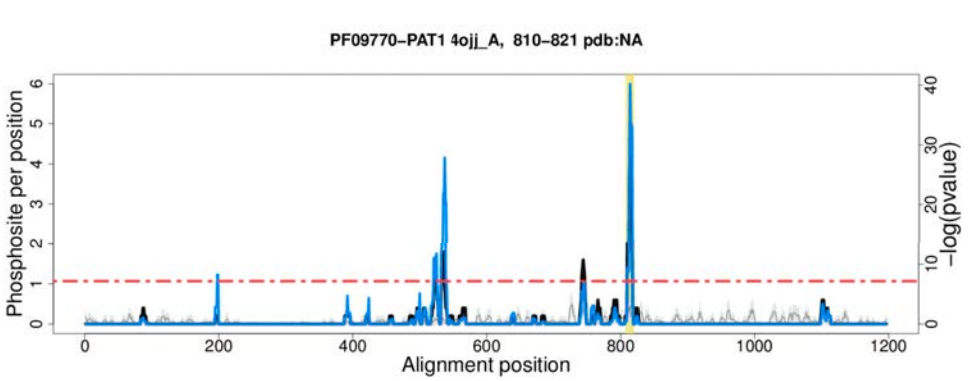

PF08911–NUP50 2c1m_B, 17–28 pdb:16–22
Phosphosite per position
−log(pvalue)
Alignment position

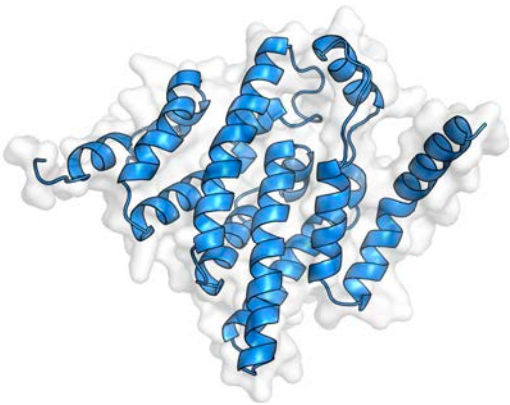

PF09110–HAND 2y9y_A, 79–87 pdb:843–848
Phosphosite per position
−log(pvalue)
Alignment position

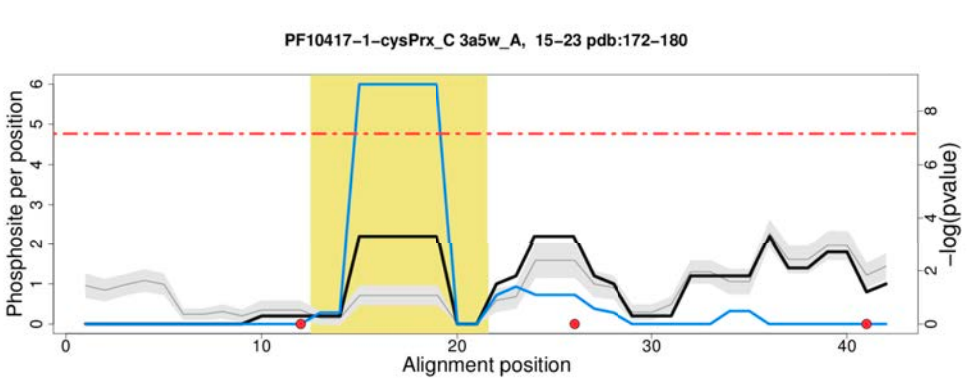

PF09359-VTC 3g3q_A, 71-79 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position

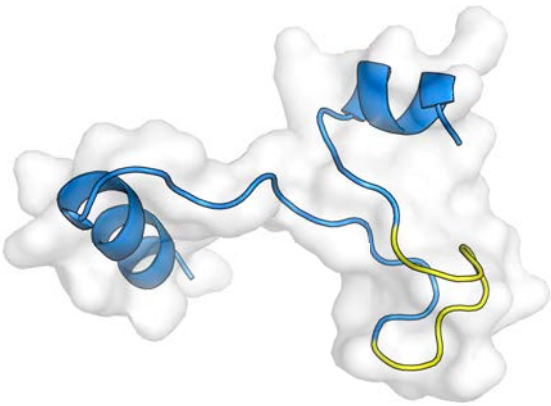

PF09380-FERM_C 2yvc_A, 91-93 pdb:295-297
Phosphosite per position
−log(pvalue)
Alignment position

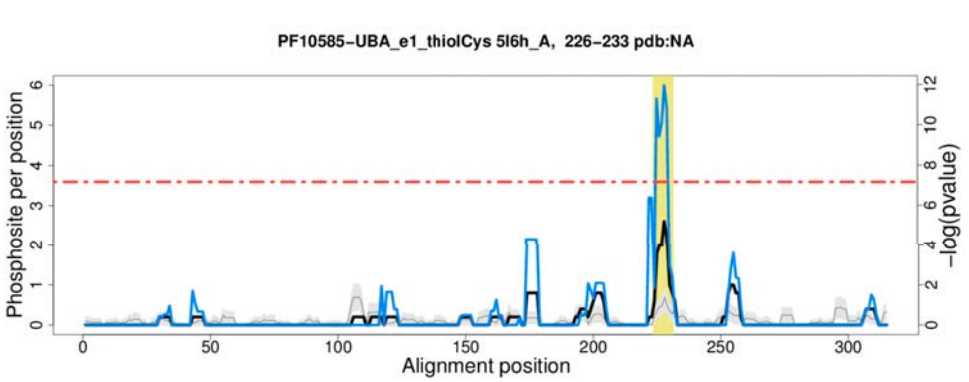


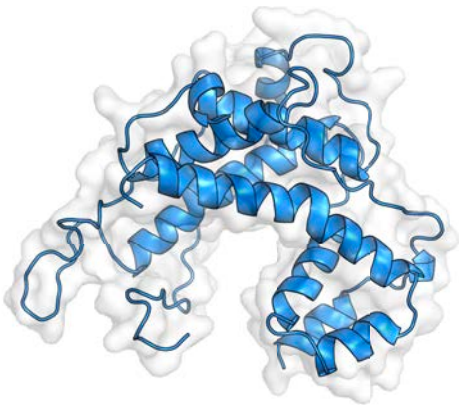

PF09770-PAT1 4ojj_A, 810-821 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position

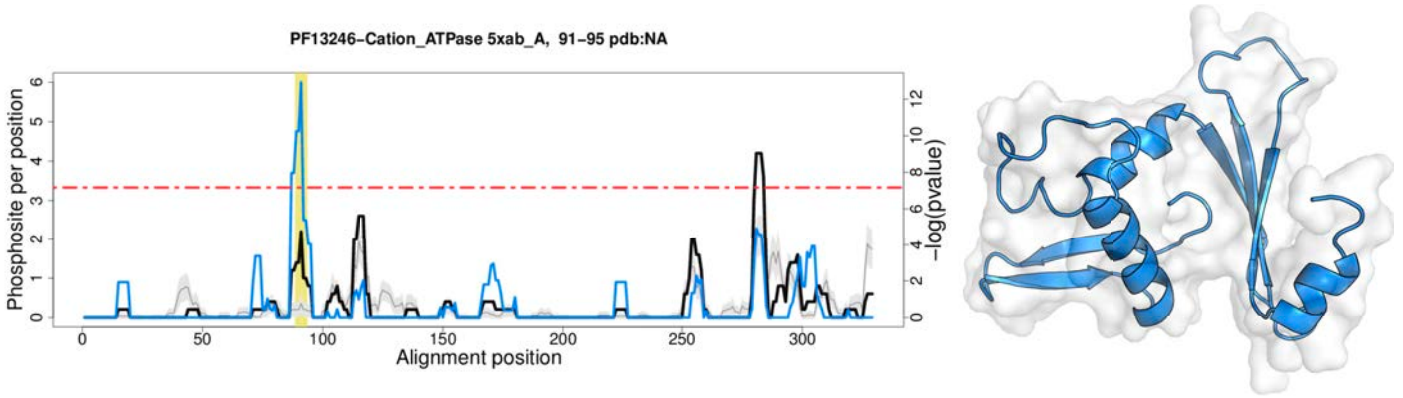


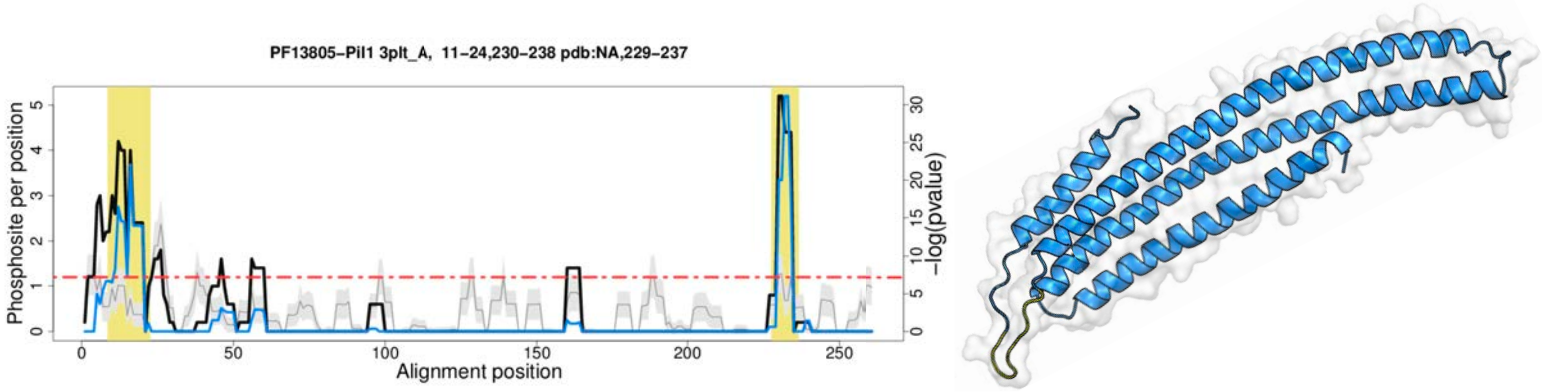

PF10417-1-cysPrx_C 3a5w_A, 15-23 pdb:172-180
Phosphosite per position
−log(pvalue)
Alignment position

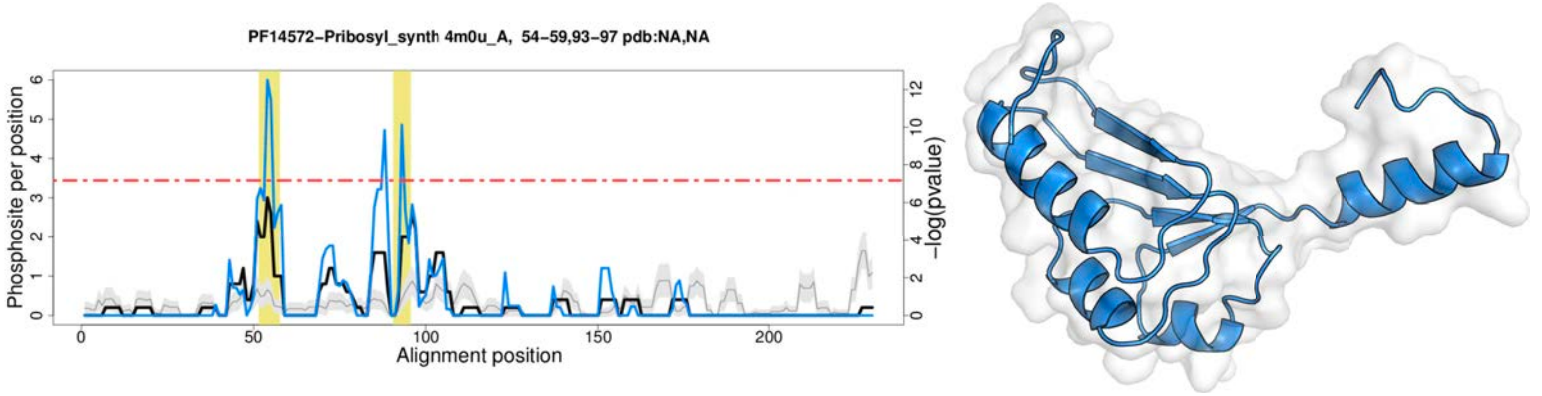


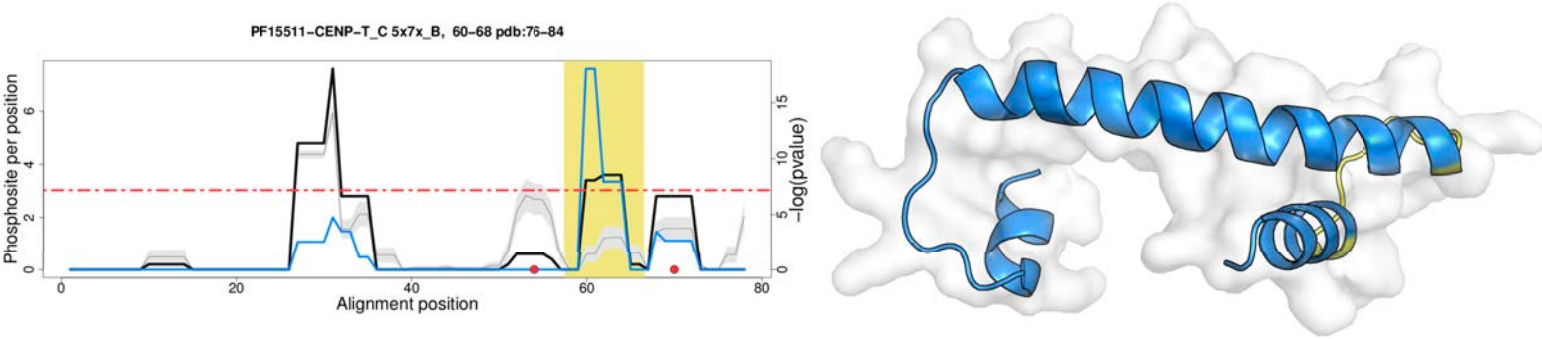

PF10585-UBA_e1_thiolCys 5l6h_A, 226-233 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position

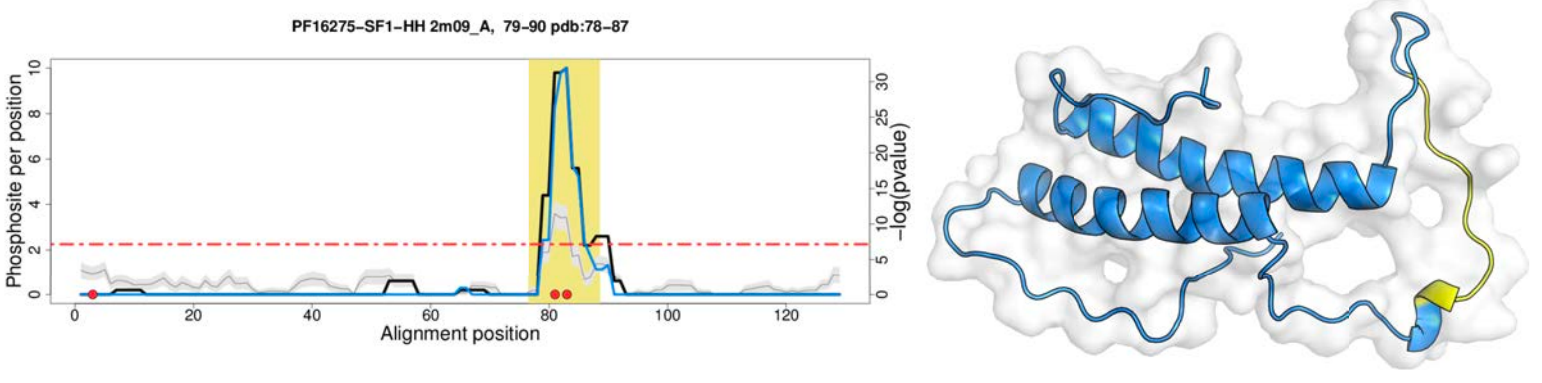



PF13246-Cation_ATPase 5xab_A, 91-95 pdb:NA
Phosphosite per position
−log(pvalue)
Alignment position


PF13805-Pil1 3plt_A, 11-24,230-238 pdb:NA,229-237
Phosphosite per position
−log(pvalue)
Alignment position

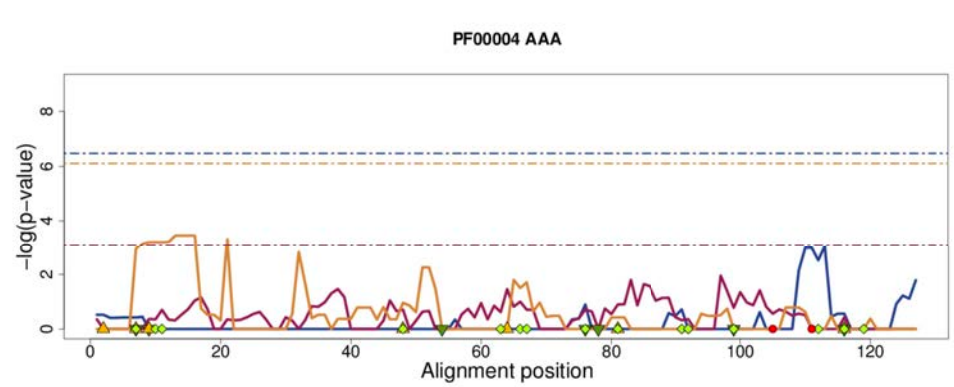

PF14572-Pribosyl_synth 4m0u_A, 54-59,93-97 pdb:NA,NA
Phosphosite per position
−log(pvalue)
Alignment position

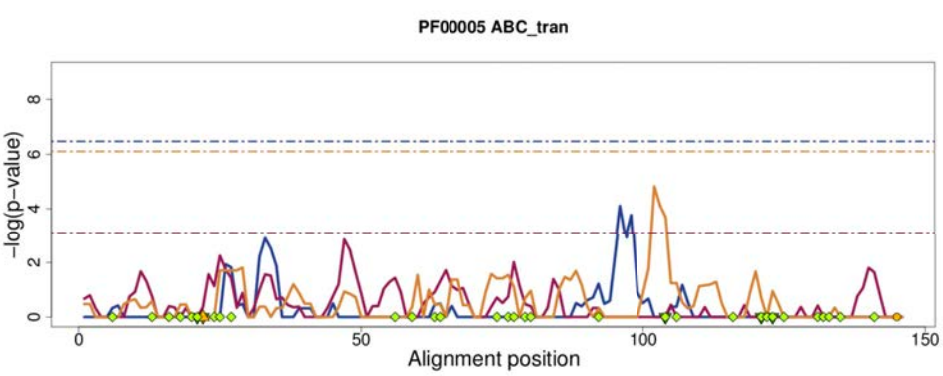

PF15511-CENP-T_C 5x7x_B, 60-68 pdb:76-84
Phosphosite per position
−log(pvalue)
Alignment position

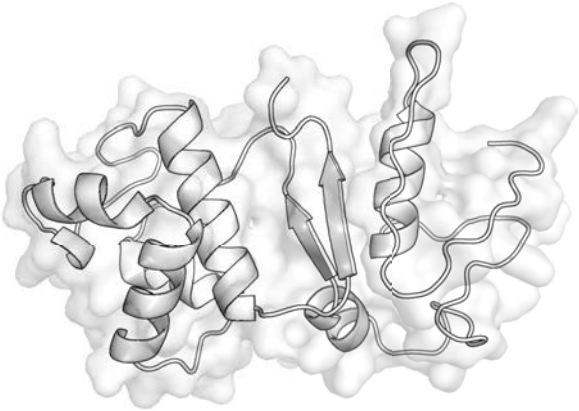

PF16275-SF1-HH 2m09_A, 79-90 pdb:78-87
Phosphosite per position
−log(pvalue)
Alignment position

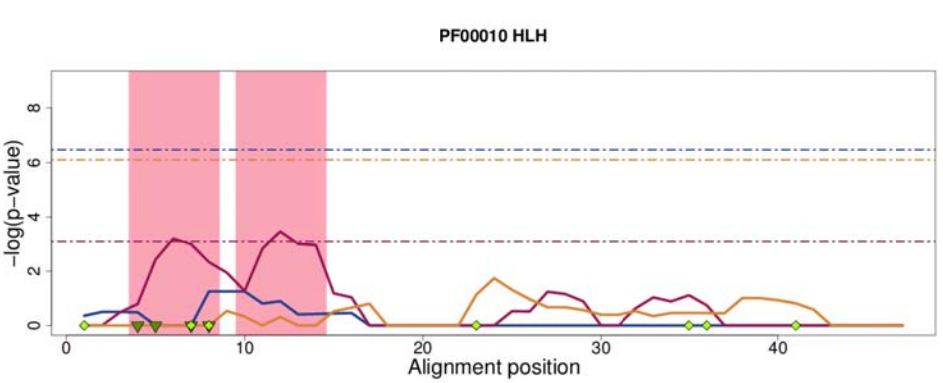
Appendix B

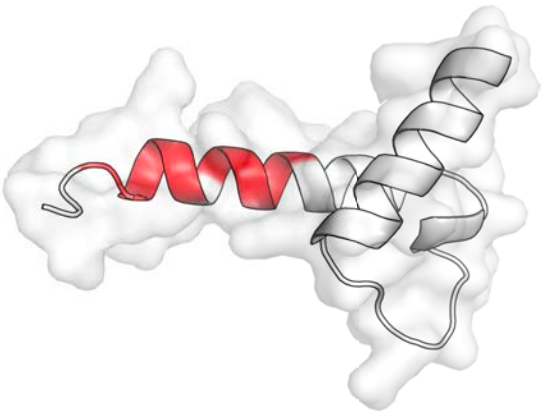
Remaining Phosphorylation, Ubiquitination and Mutation hotspots

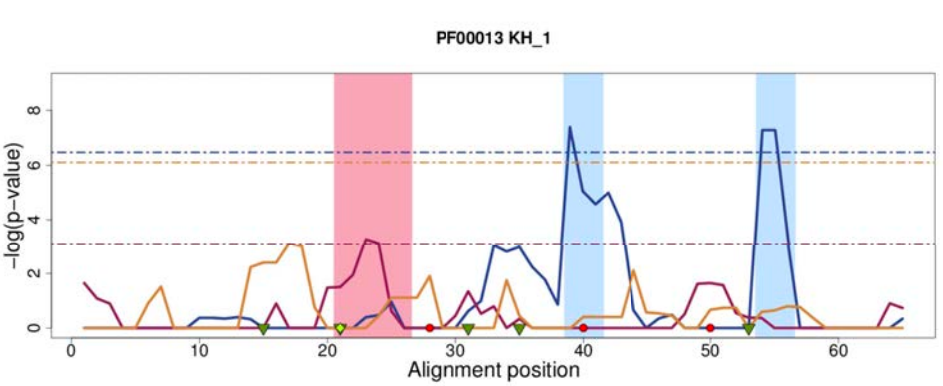

PF00004 AAA
−log(p−value)
Alignment position

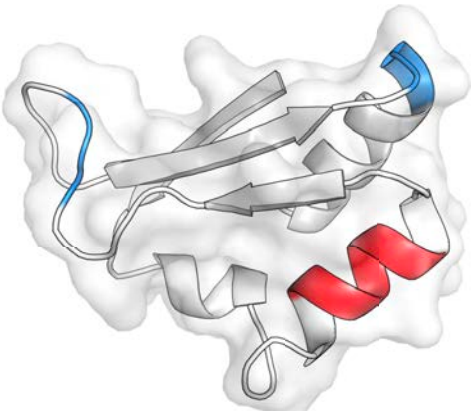

PF00005 ABC_tran
−log(p−value)
Alignment position

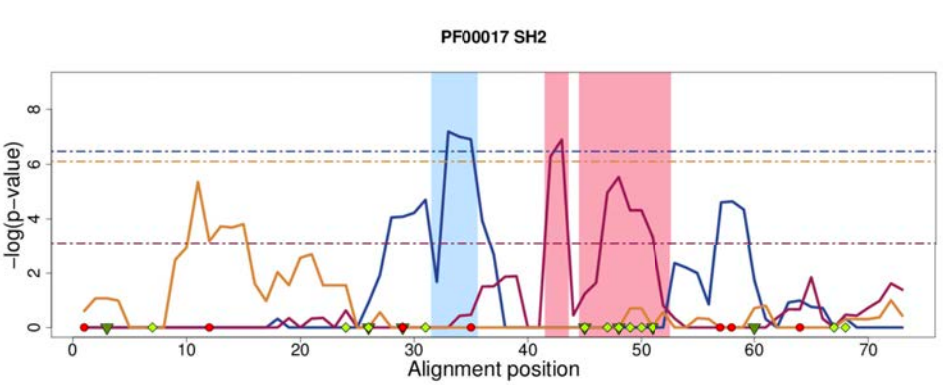


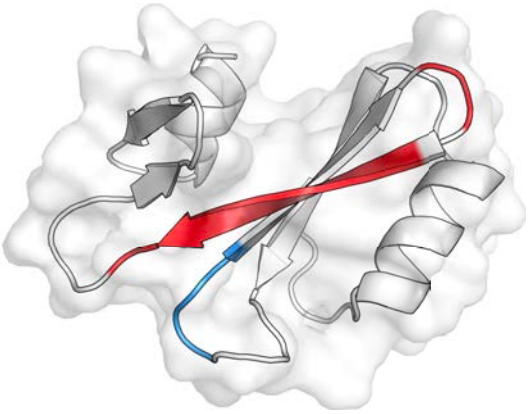

PF00010 HLH
−log(p−value)
Alignment position



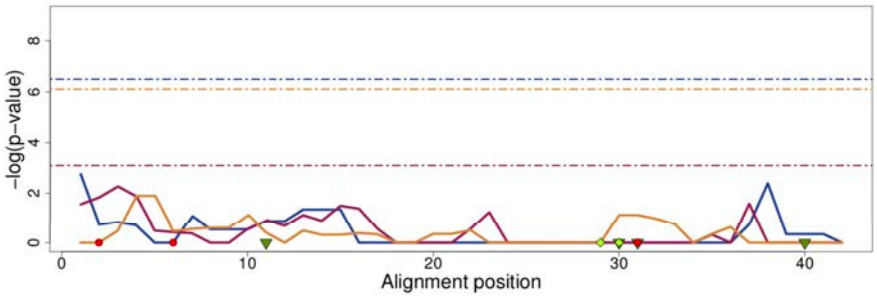

PF00013 KH_1
−log(p−value)
Alignment position

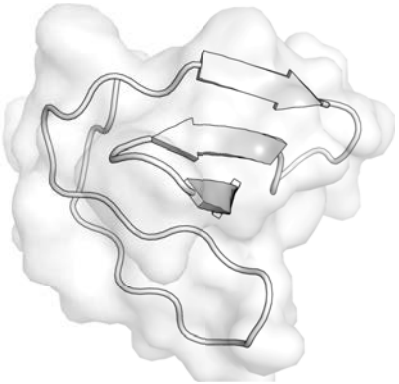



PF00017 SH2
−log(p−value)
Alignment position

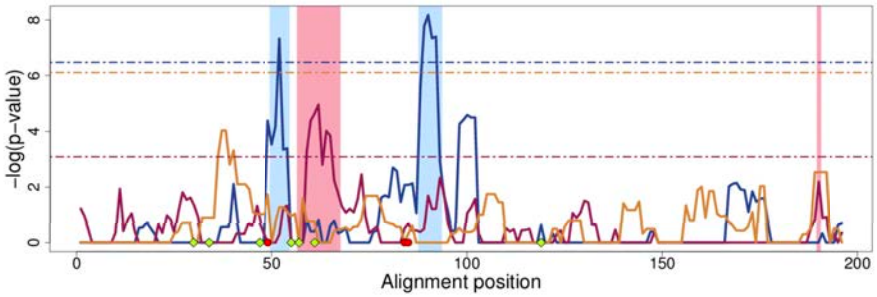


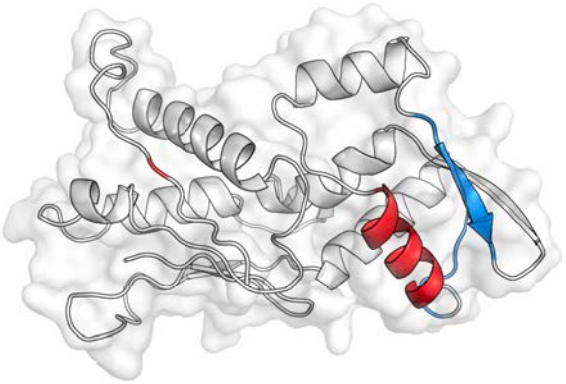
PF00018 SH3_1



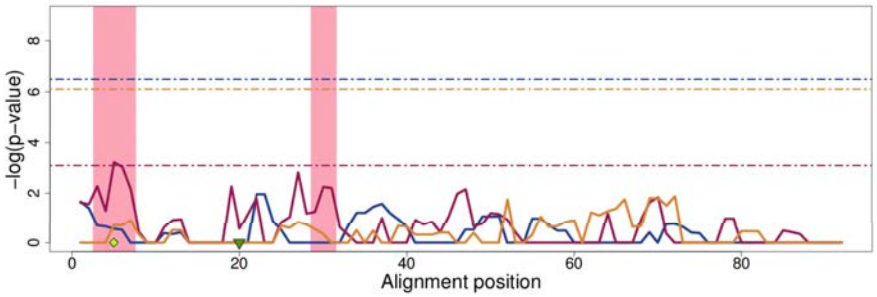


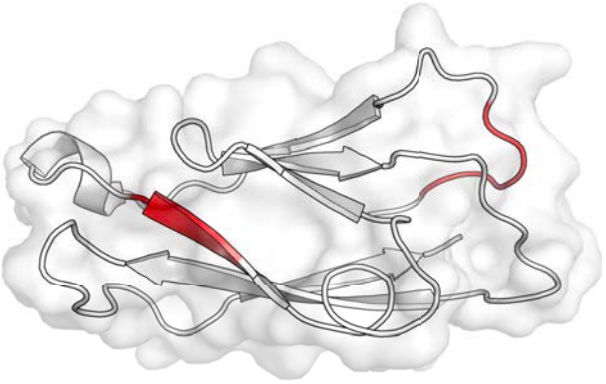
PF00022 Actin



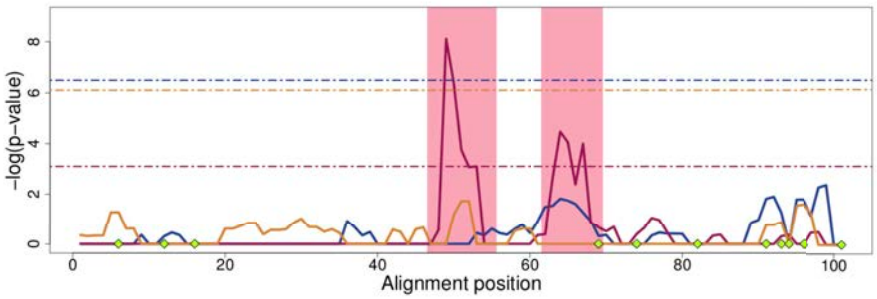


PF00028 Cadherin

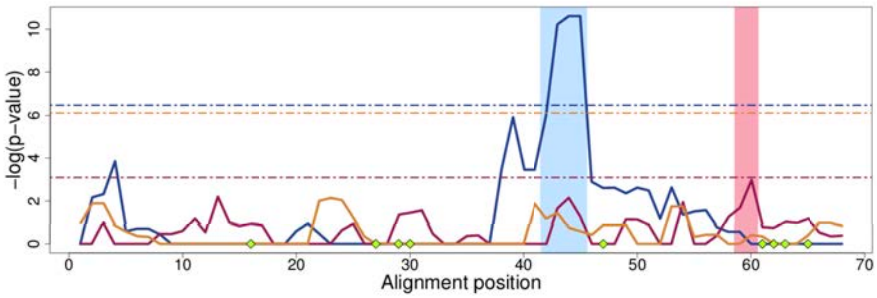


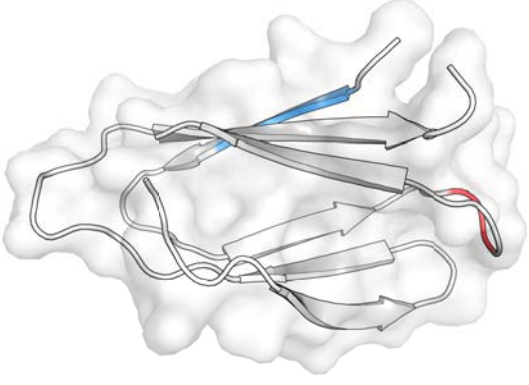


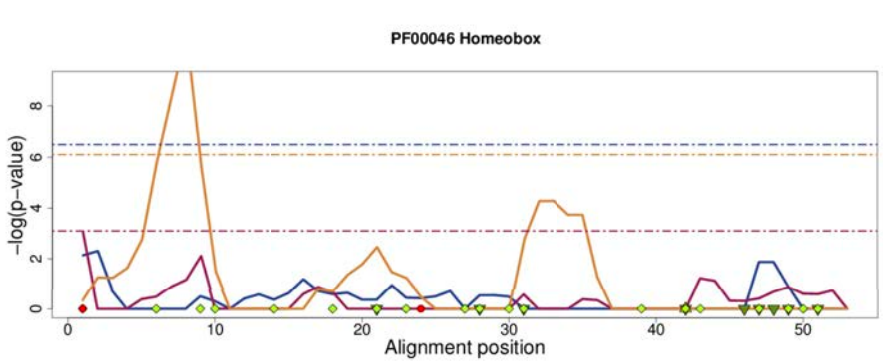
PF00038 Filament

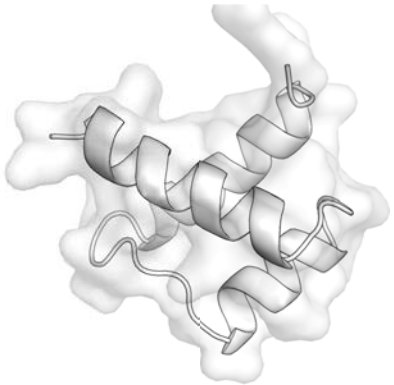


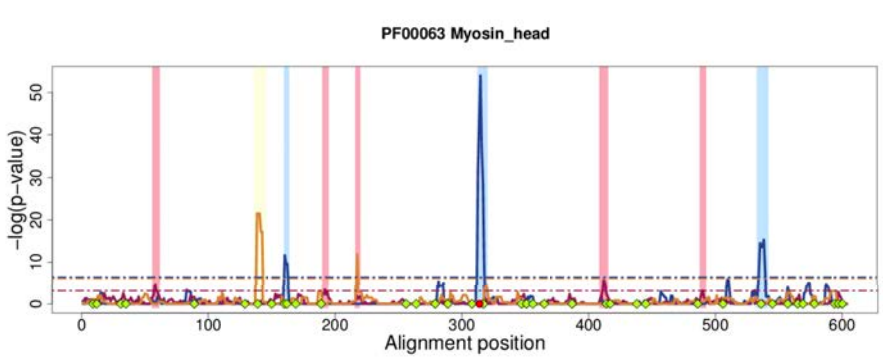
PF00041 fn3

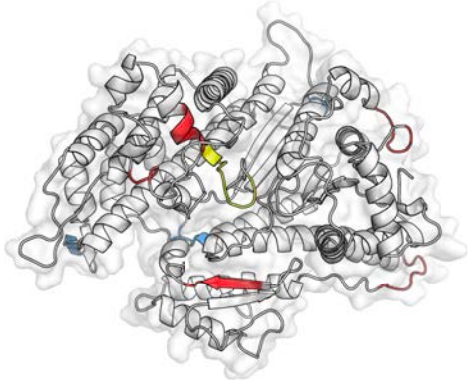


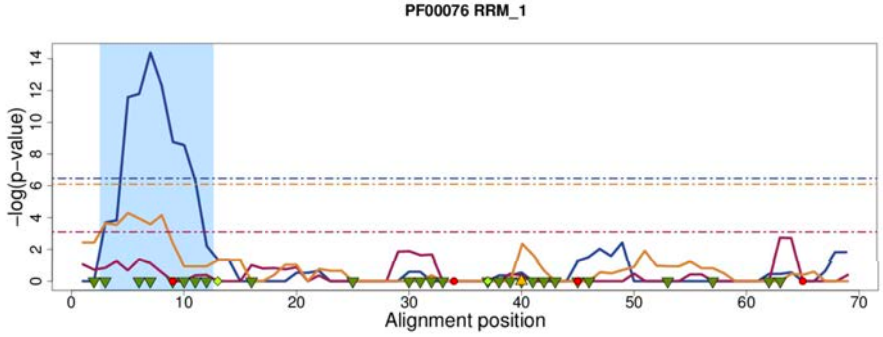


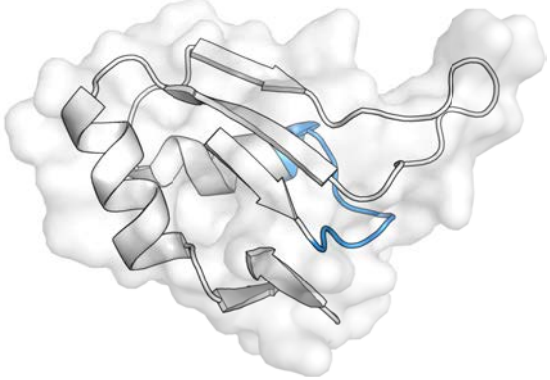

PF00046 Homeobox
−log(p−value)
Alignment position

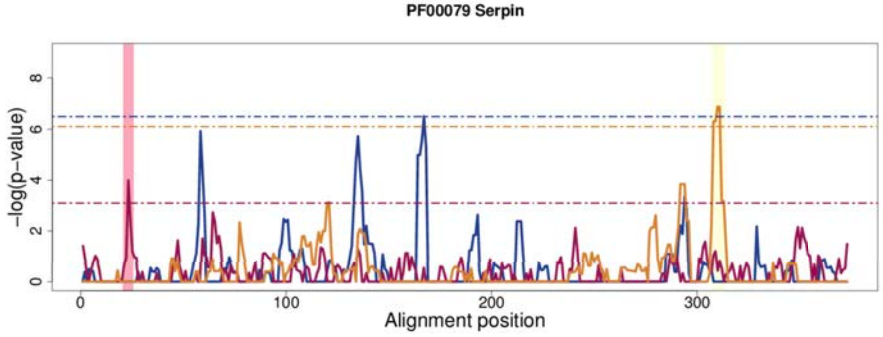


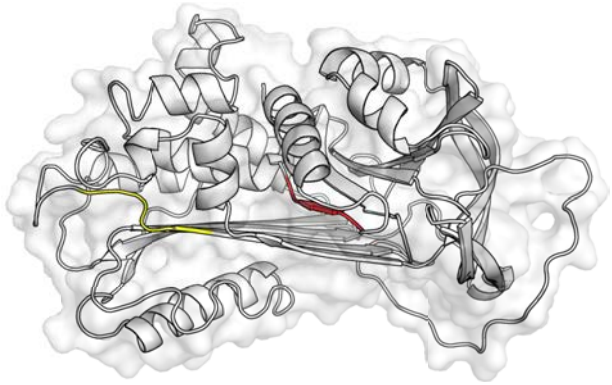

PF00063 Myosin_head
−log(p−value)
Alignment position

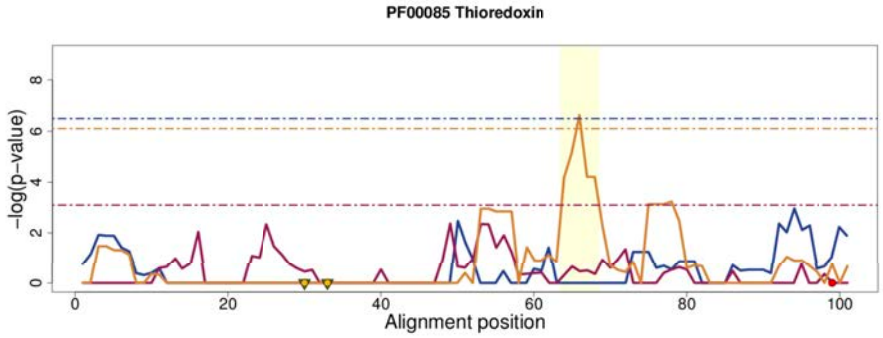


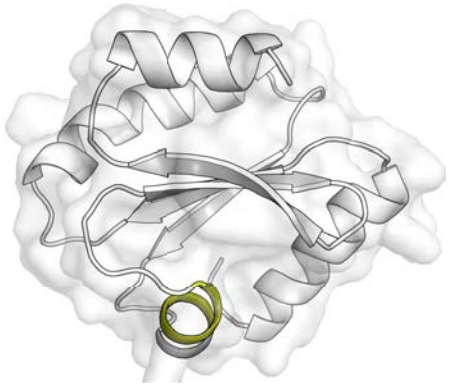

PF00076 RRM_1
−log(p−value)
Alignment position

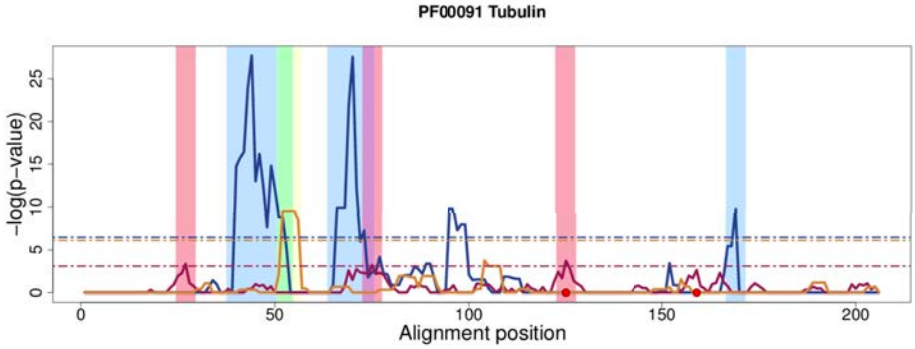


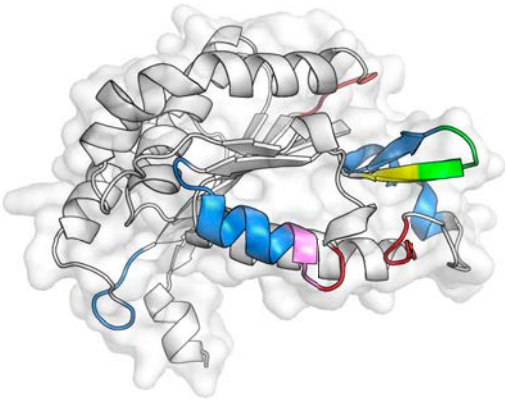

PF00079 Serpin
−log(p−value)
Alignment position

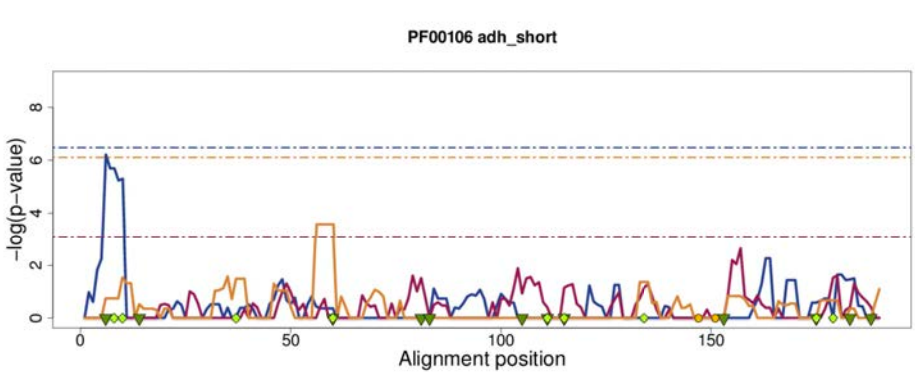


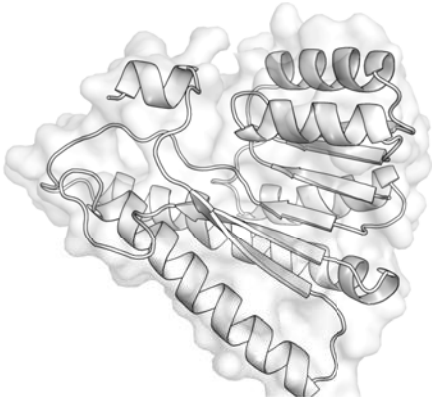

PF00085 Thioredoxin
−log(p−value)
Alignment position

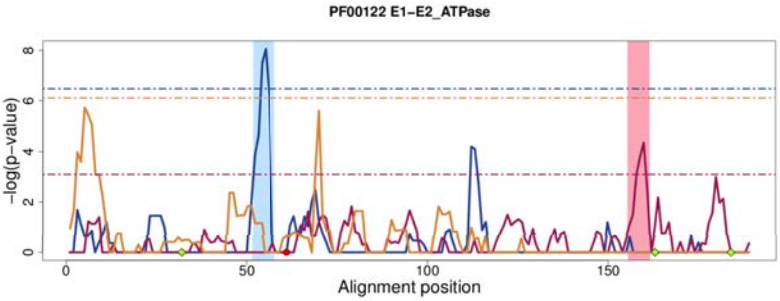


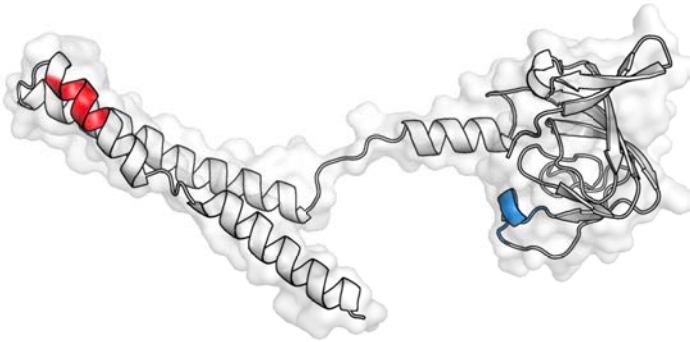

PF00091 Tubulin
−log(p−value)
Alignment position

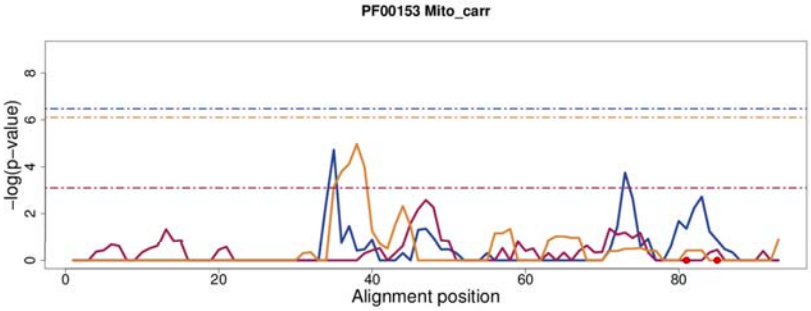


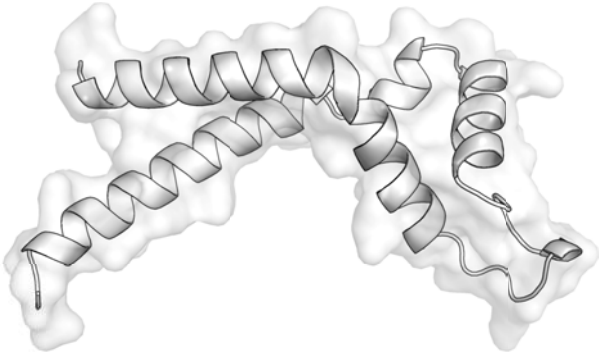

PF00106 adh_short
−log(p−value)
Alignment position

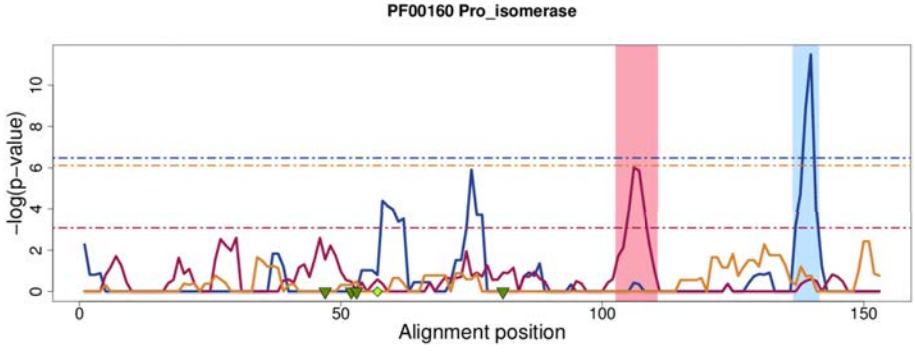


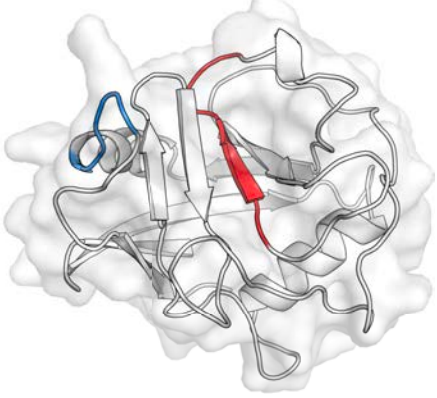

PF00122 E1−E2_ATPase
−log(p−value)
Alignment position

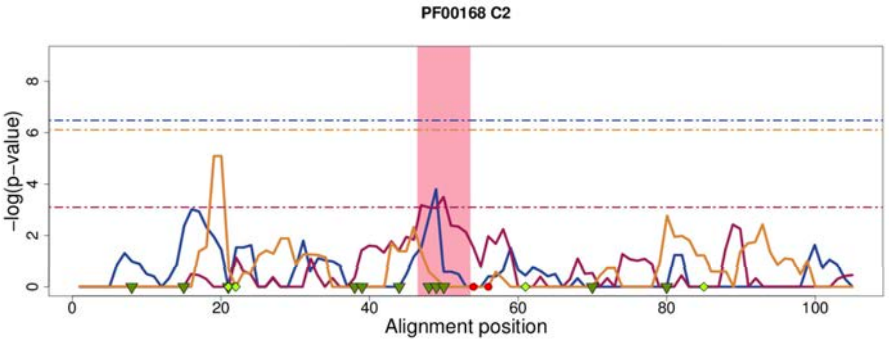


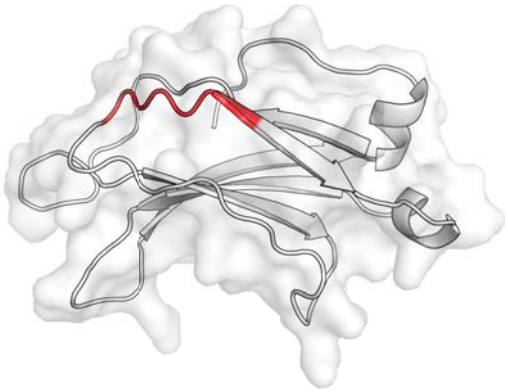

PF00153 Mito_carr
−log(p−value)
Alignment position

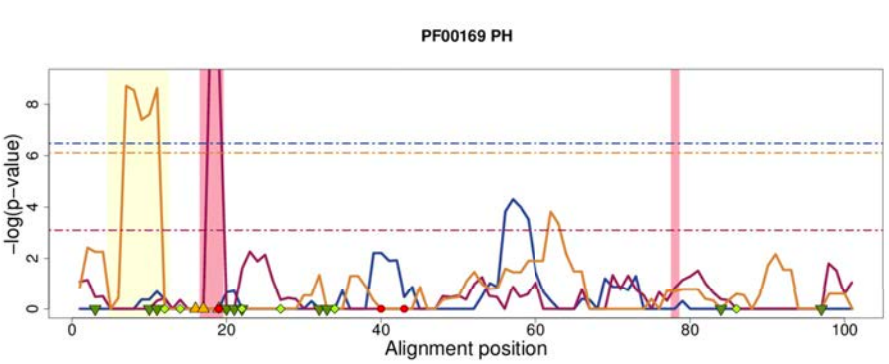


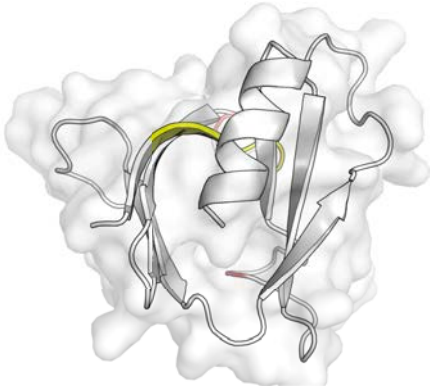

PF00160 Pro_isomerase
−log(p−value)
Alignment position

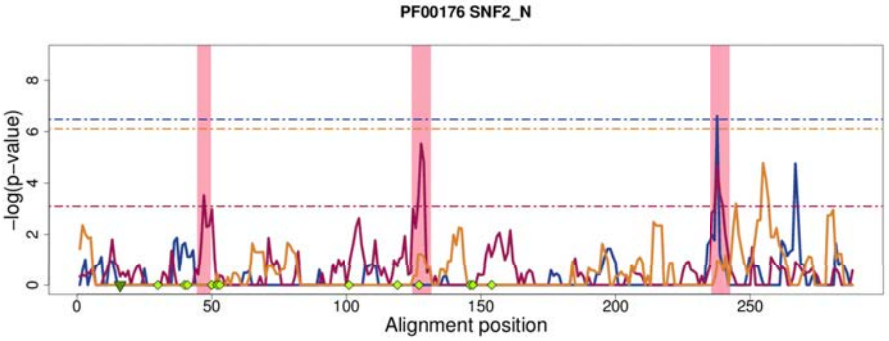


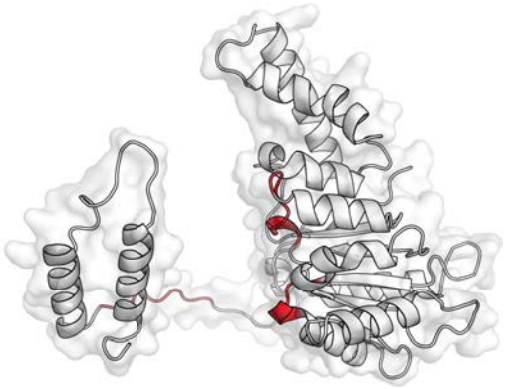

PF00168 C2
−log(p−value)
Alignment position

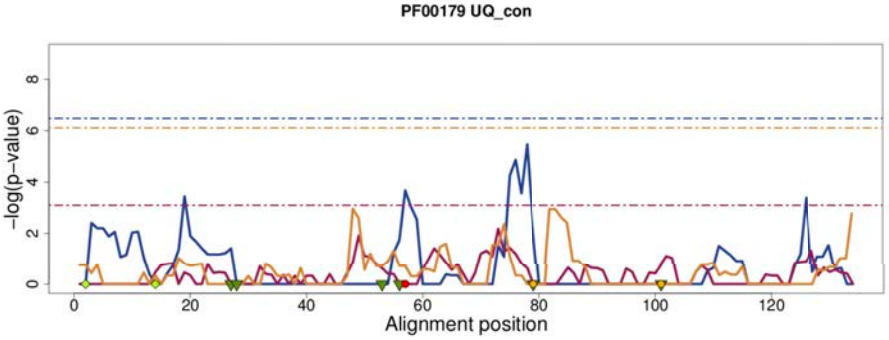


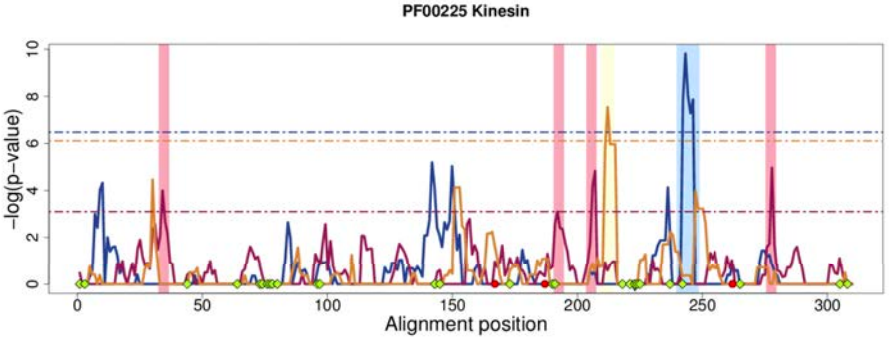

PF00169 PH
−log(p−value)
Alignment position

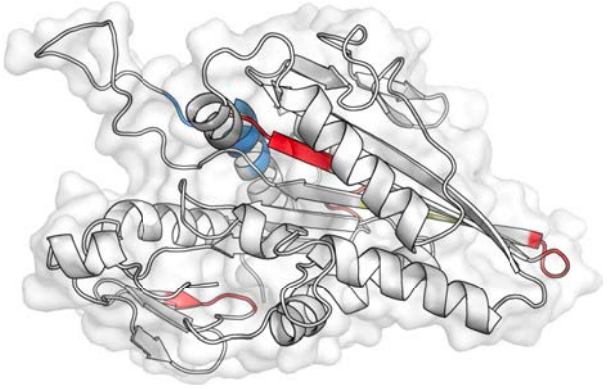



PF00176 SNF2_N
−log(p−value)
Alignment position

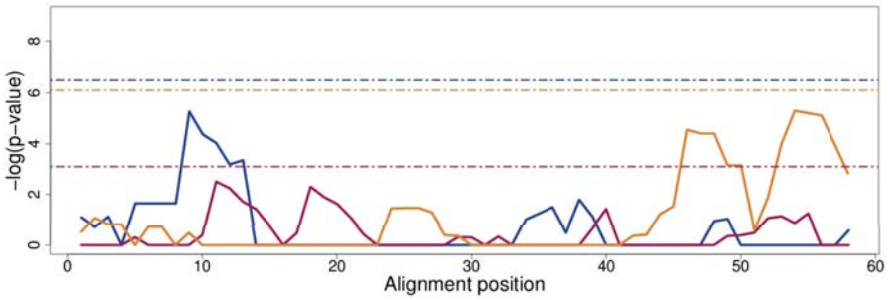


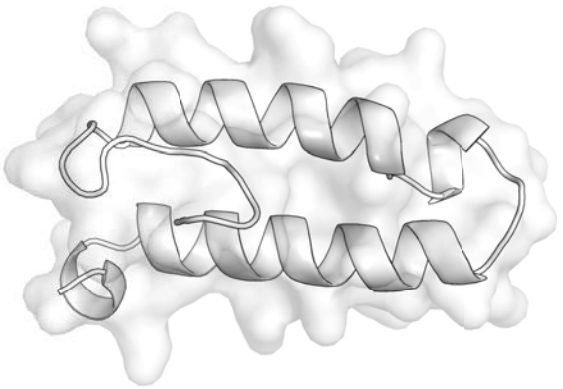

PF00179 UQ_con
−log(p−value)
Alignment position


PF00225 Kinesin
−log(p−value)
Alignment position

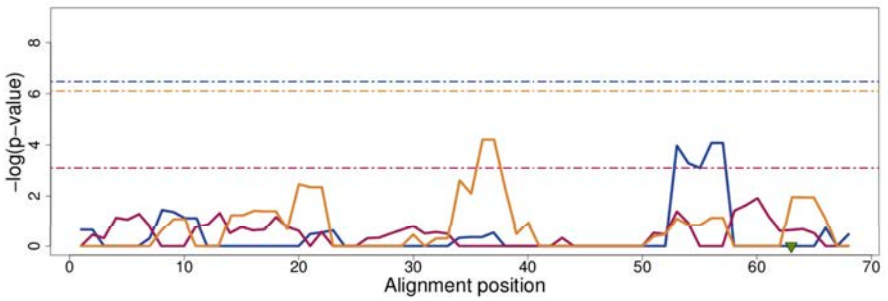


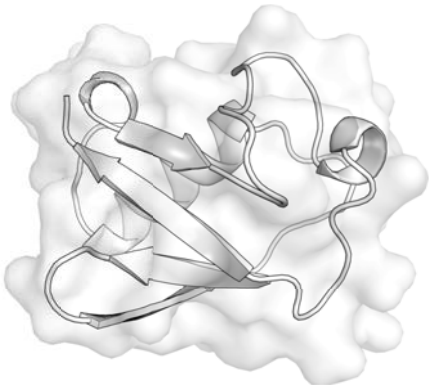
PF00226 DnaJ



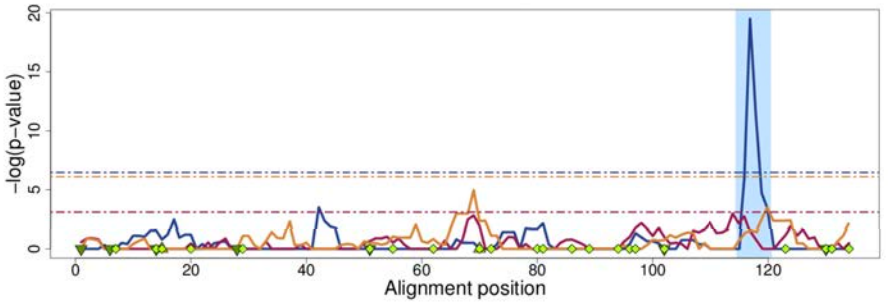


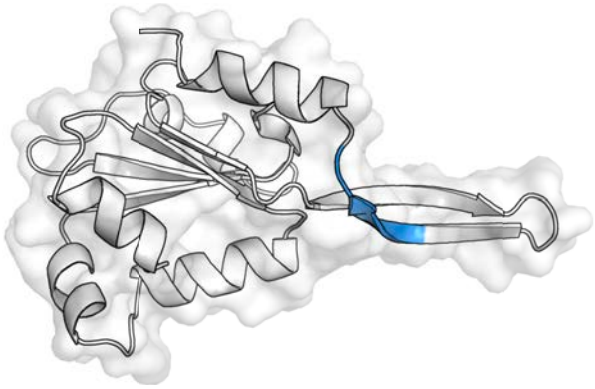
PF00240 ubiquitin



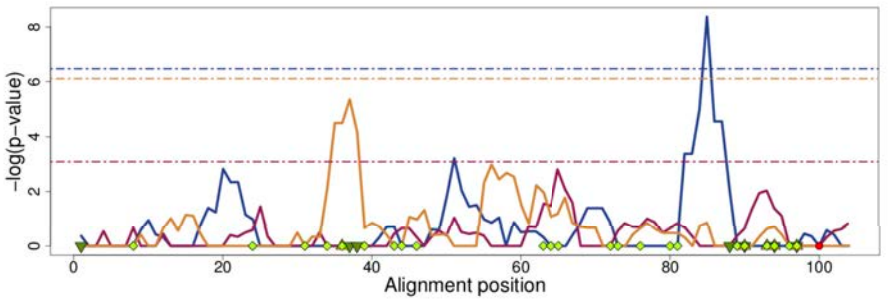


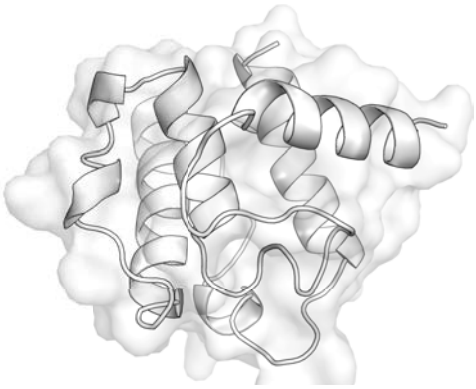
PF00271 Helicase_C



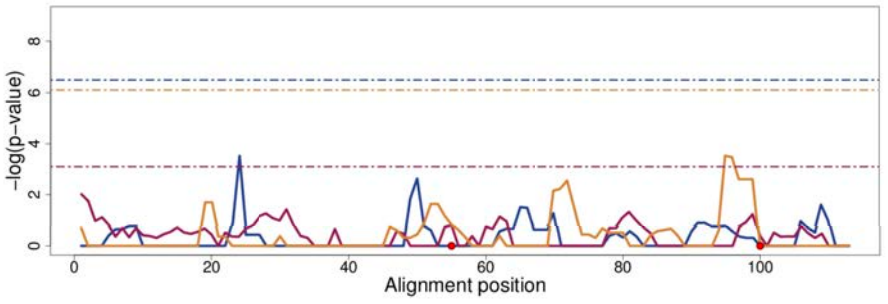


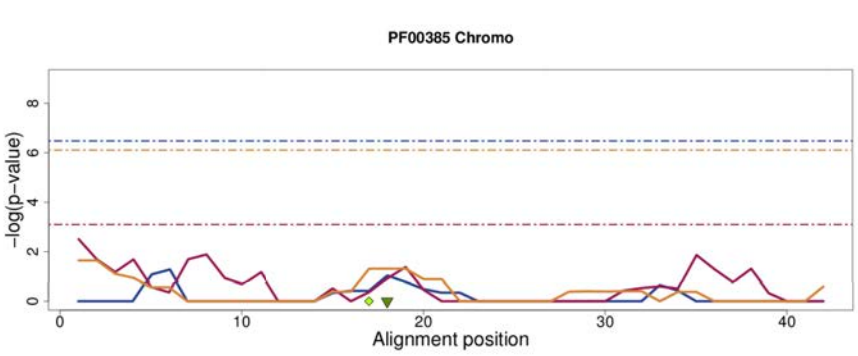
PF00307 CH

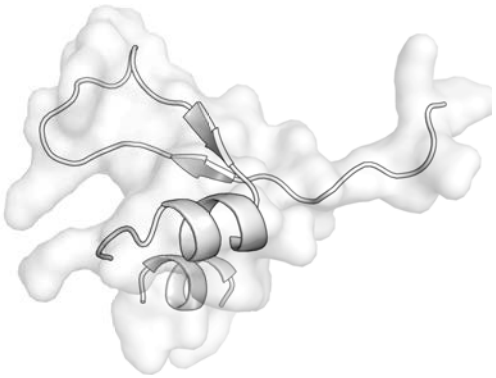


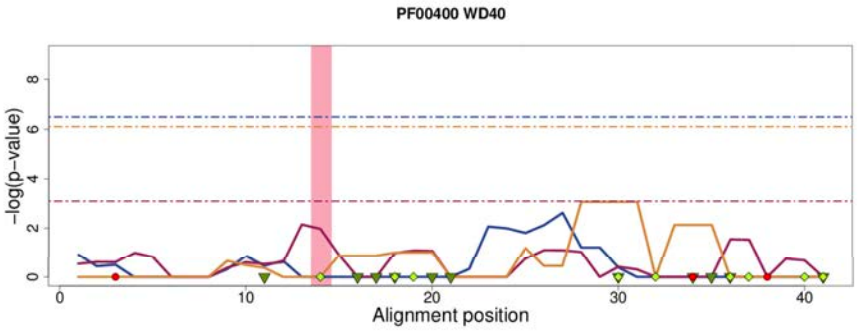


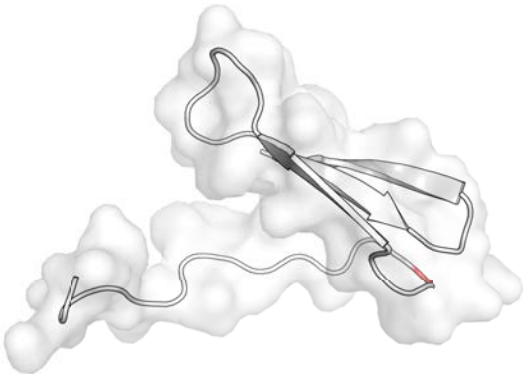
PF00373 FERM_M

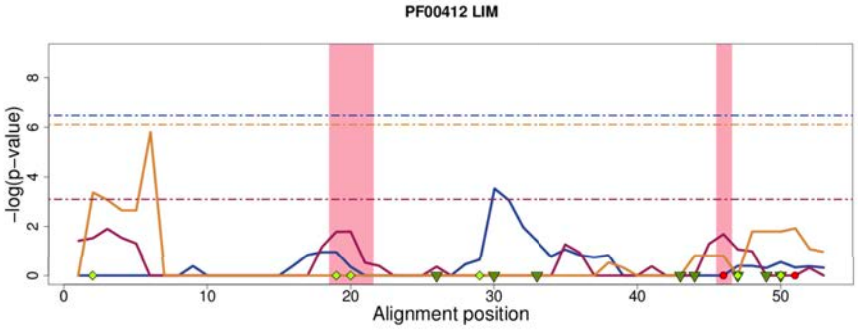


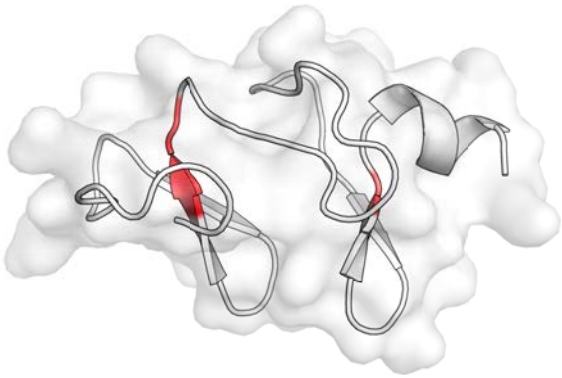

PF00385 Chromo
−log(p−value)
Alignment position

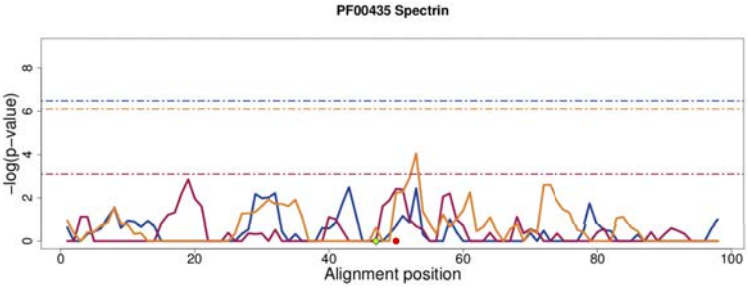


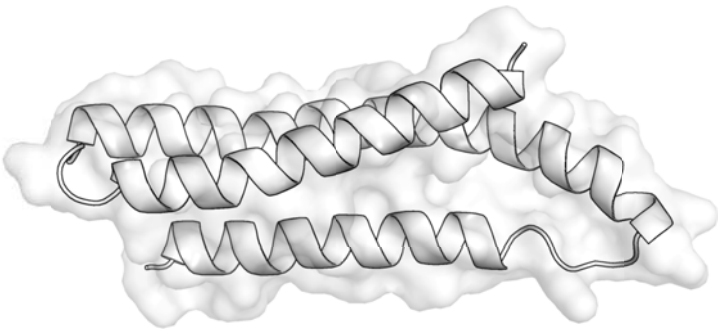

PF00400 WD40
−log(p−value)
Alignment position

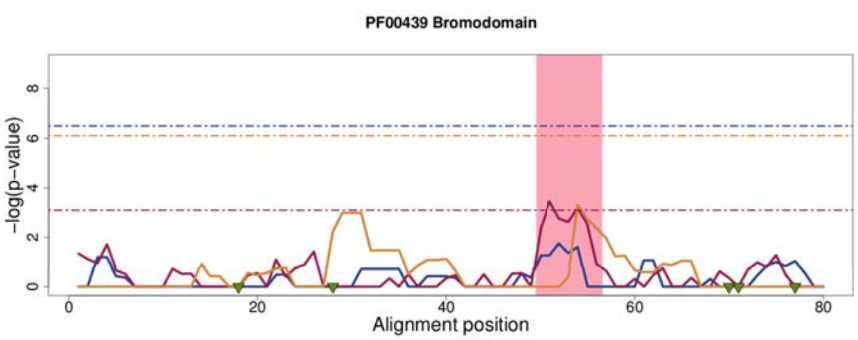


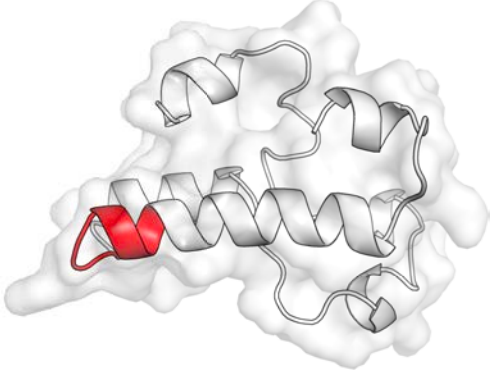

PF00412 LIM
−log(p−value)
Alignment position

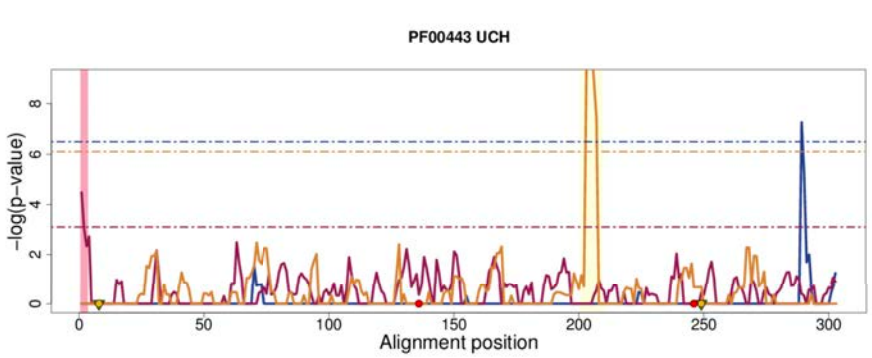


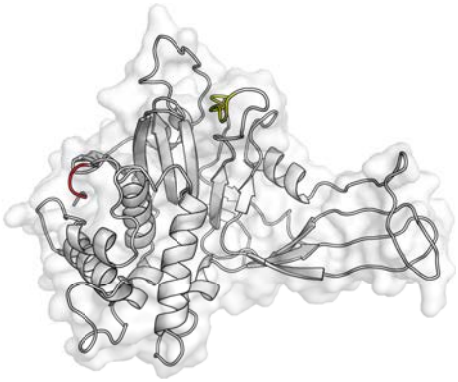

PF00435 Spectrin
−log(p−value)
Alignment position

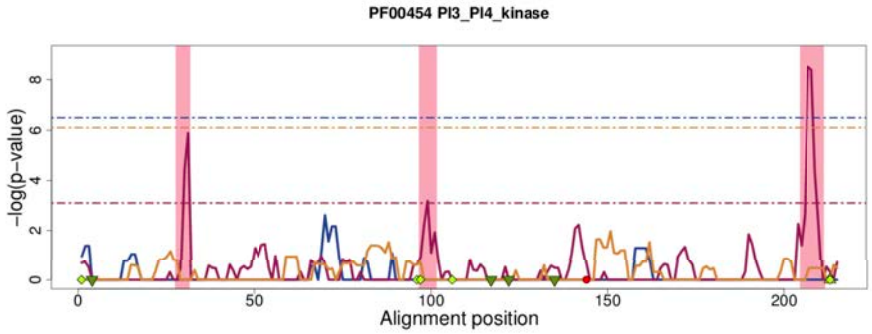


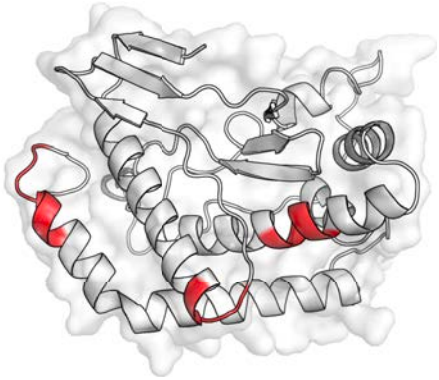

PF00439 Bromodomain
−log(p−value)
Alignment position

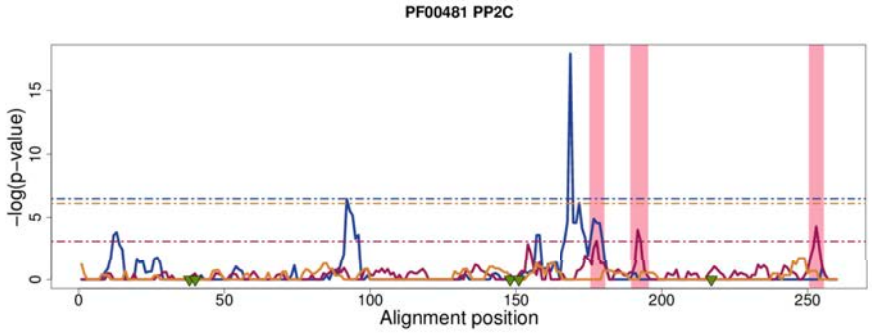


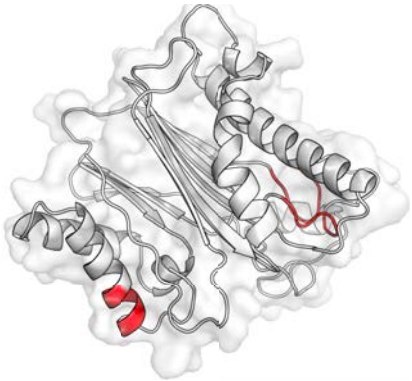

PF00443 UCH
−log(p−value)
Alignment position

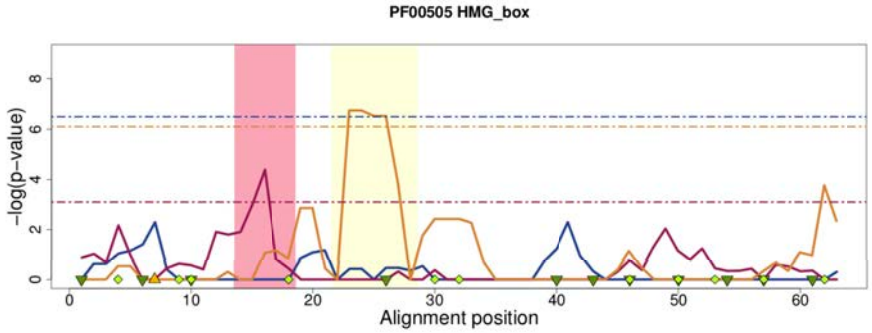


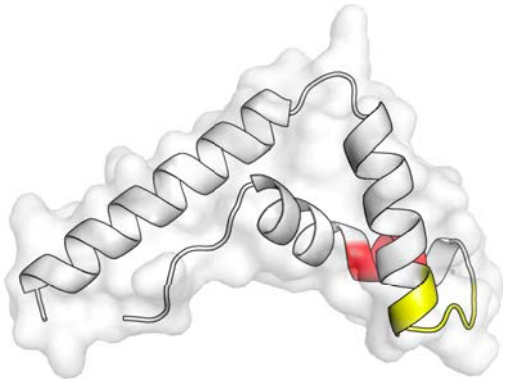

PF00454 PI3_PI4_kinase
−log(p−value)
Alignment position

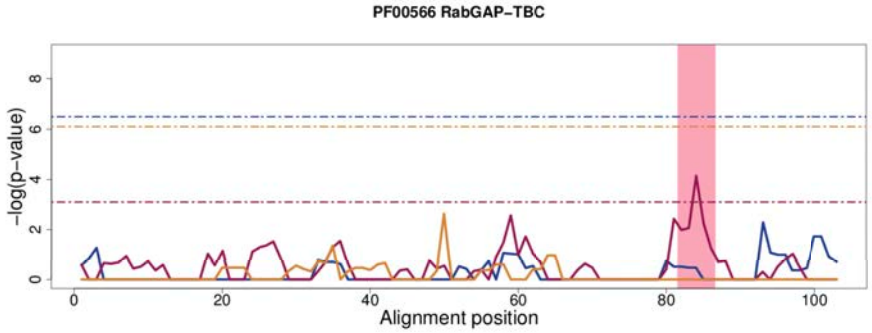


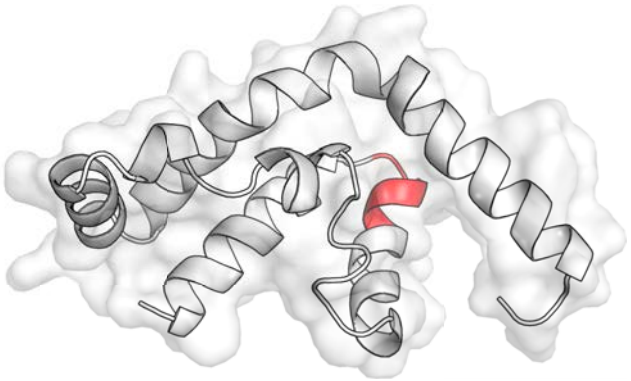

PF00481 PP2C
−log(p−value)
Alignment position

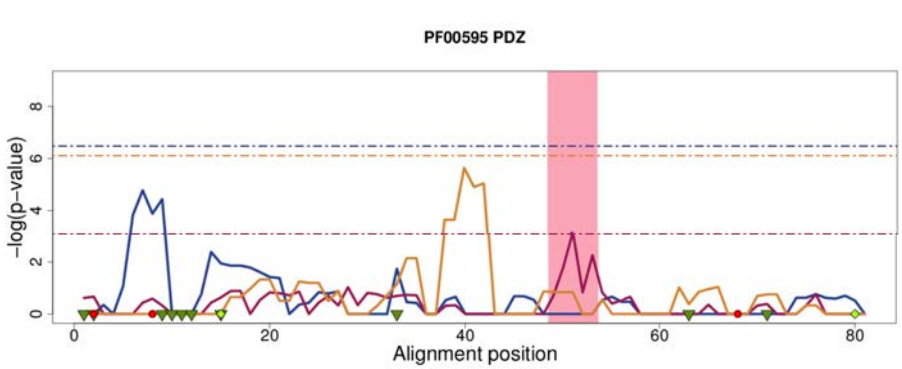


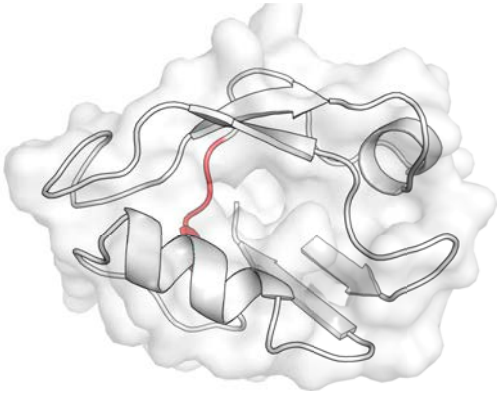

PF00505 HMG_box
−log(p−value)
Alignment position

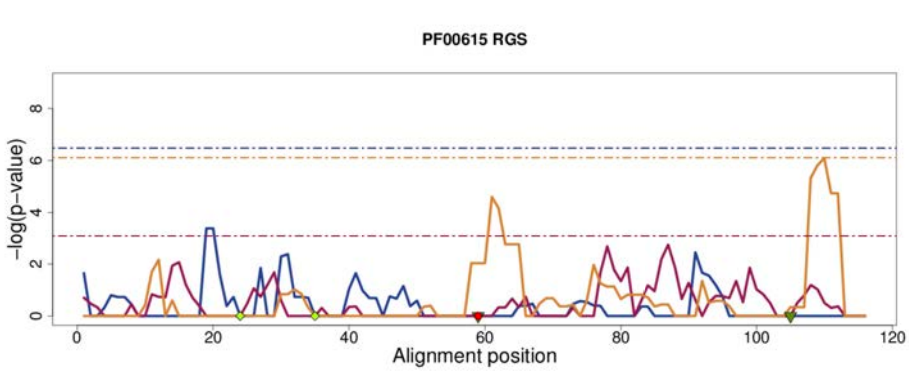


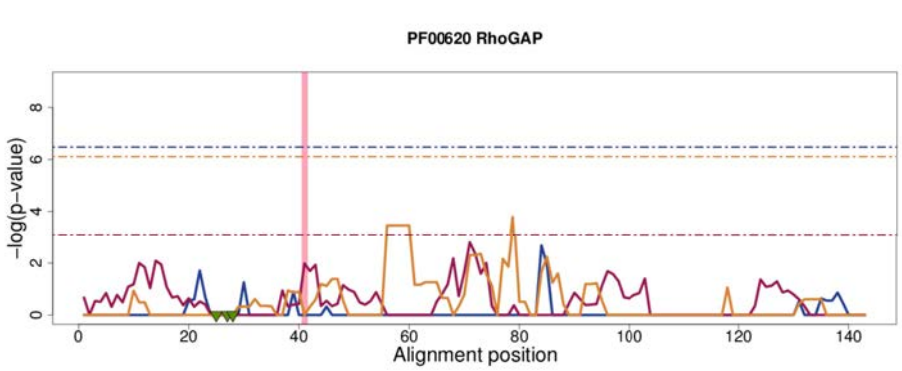

PF00566 RabGAP−TBC
−log(p−value)
Alignment position

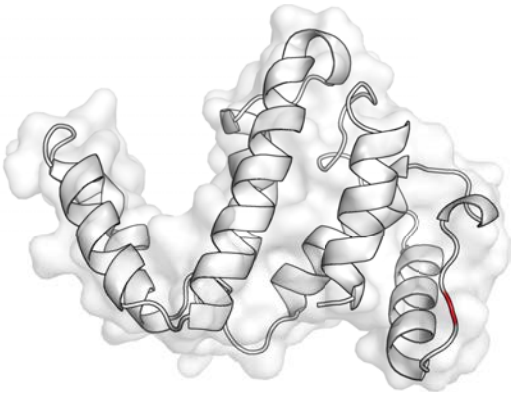


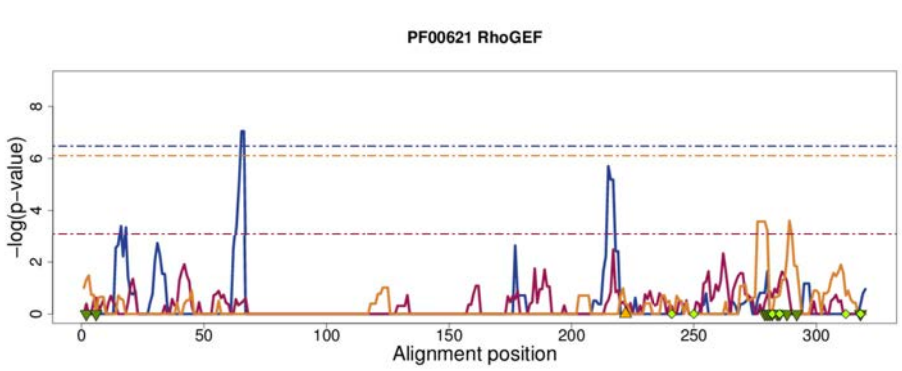

PF00595 PDZ
−log(p−value)
Alignment position

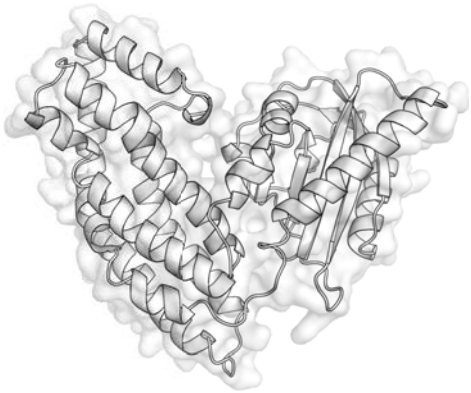


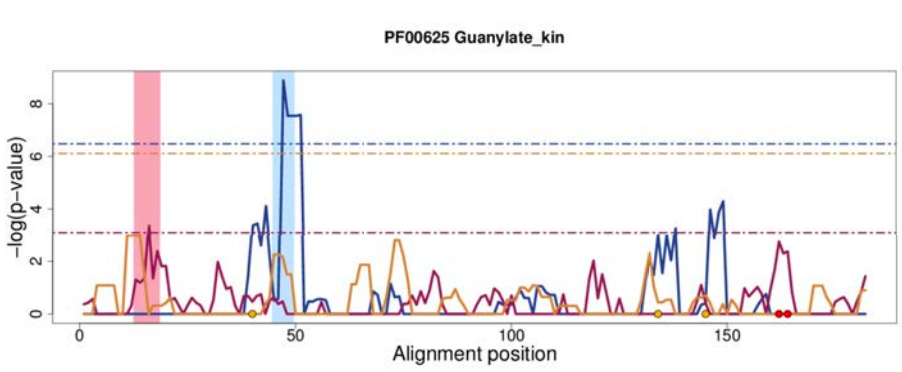

PF00615 RGS
−log(p−value)
Alignment position

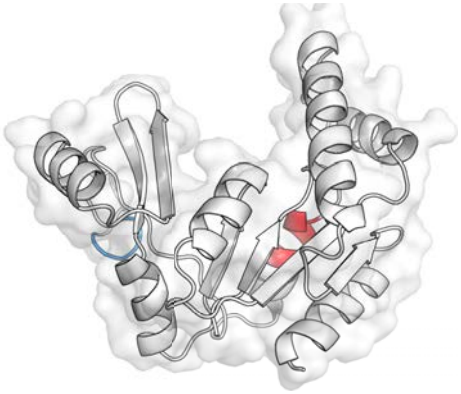

PF00620 RhoGAP
−log(p−value)
Alignment position

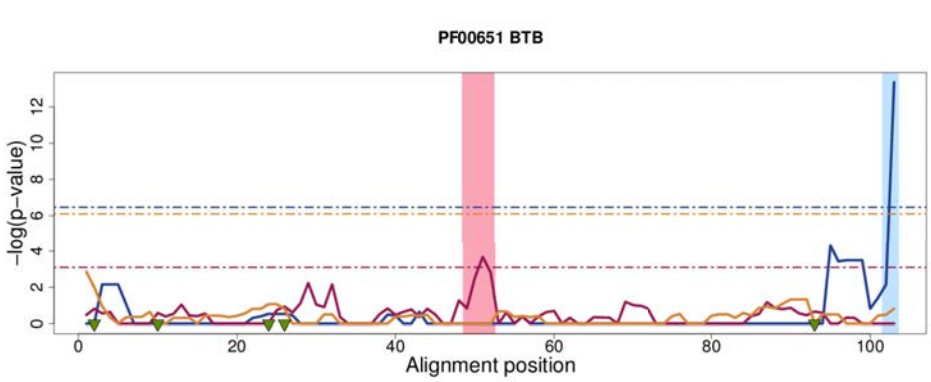


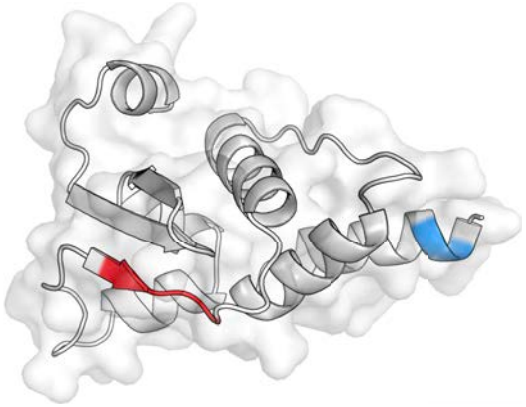

PF00621 RhoGEF
−log(p−value)
Alignment position



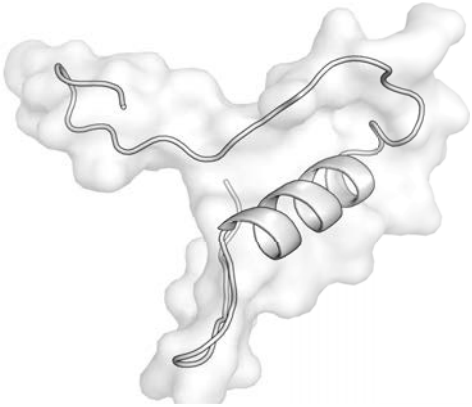

PF00625 Guanylate_kin
−log(p−value)
Alignment position

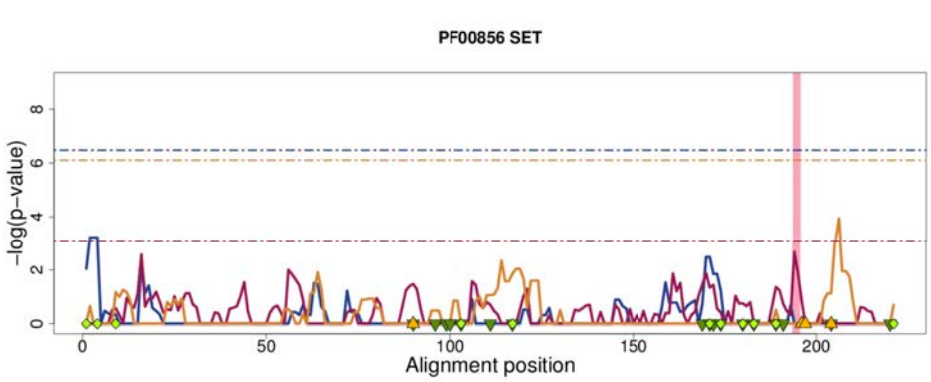


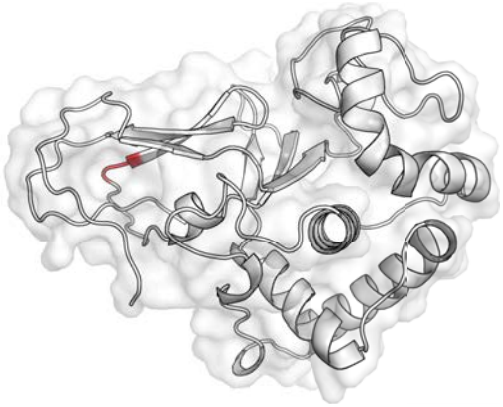

PF00651 BTB
−log(p−value)
Alignment position

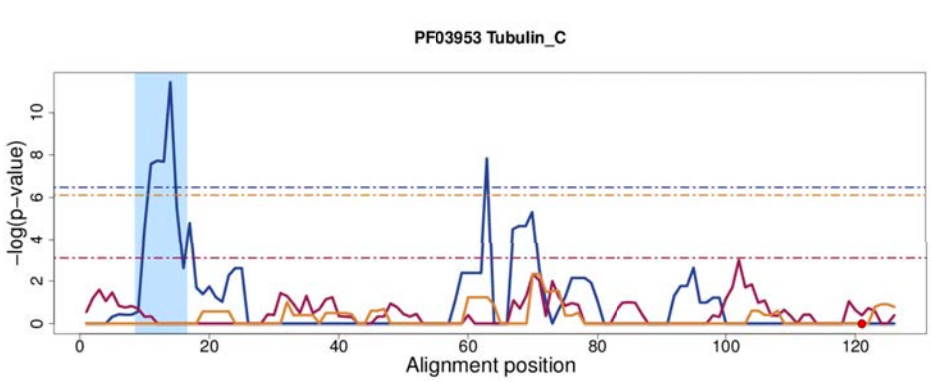


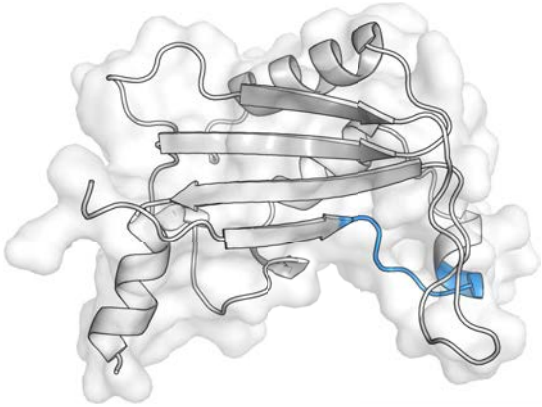

PF00788 RA
−log(p−value)
Alignment position

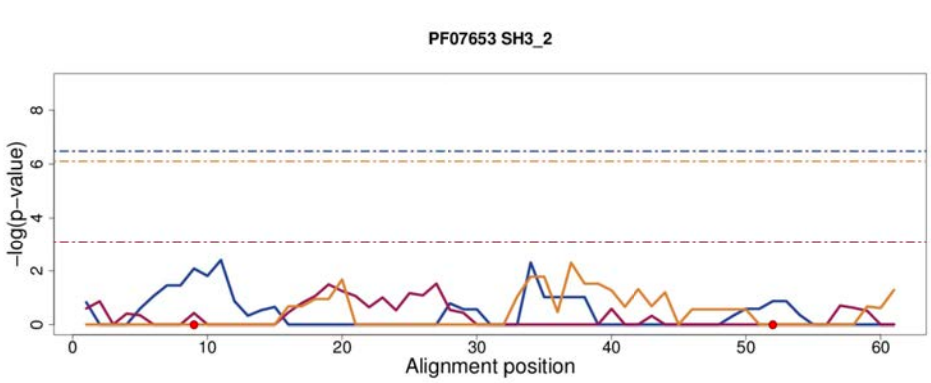


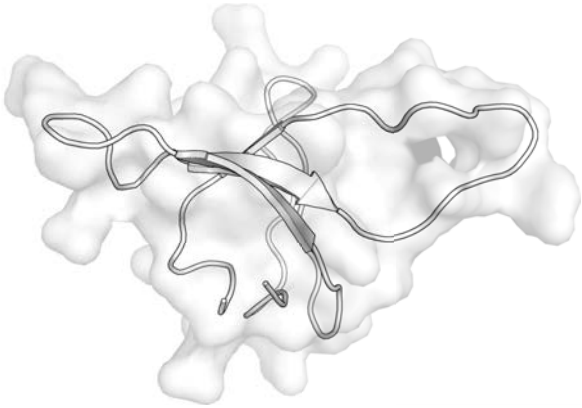

PF00856 SET
−log(p−value)
Alignment position

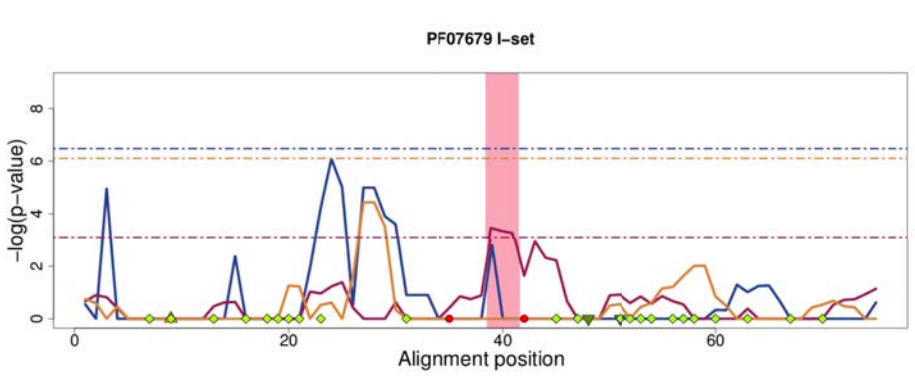


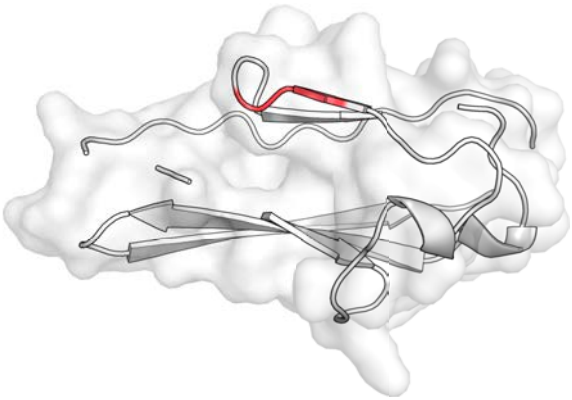

PF03953 Tubulin_C
−log(p−value)
Alignment position

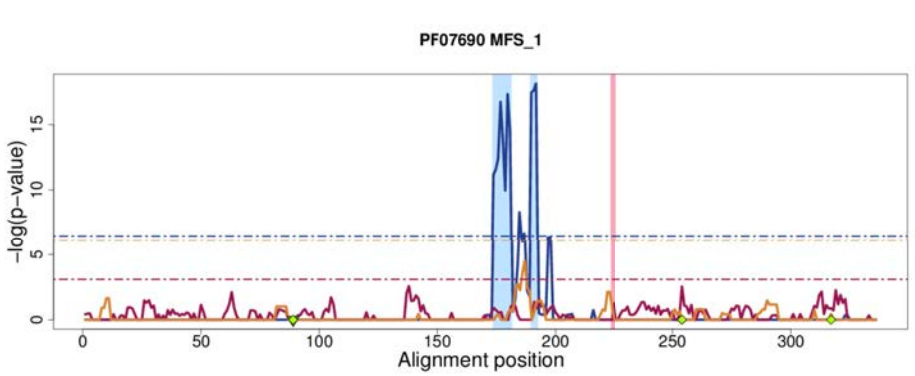


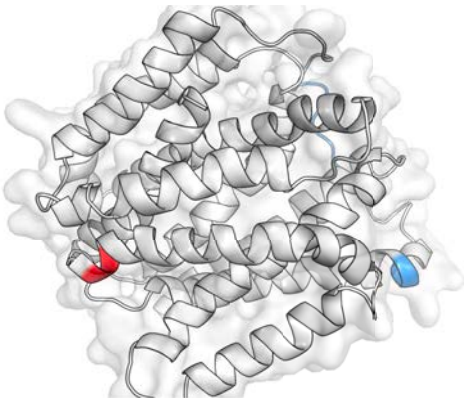

PF07653 SH3_2
−log(p−value)
Alignment position

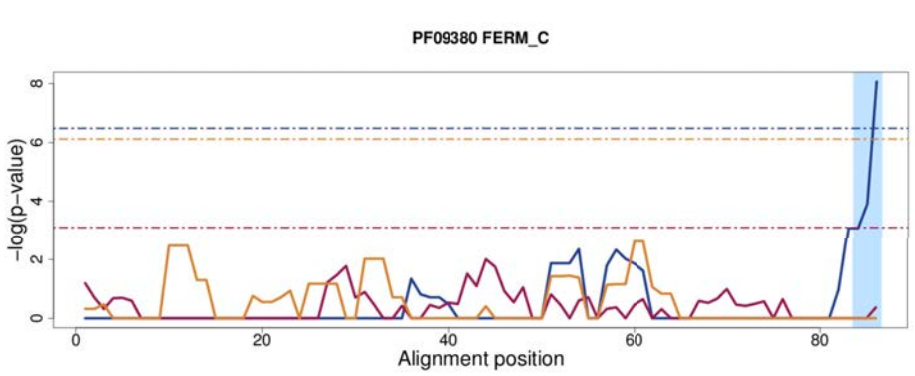


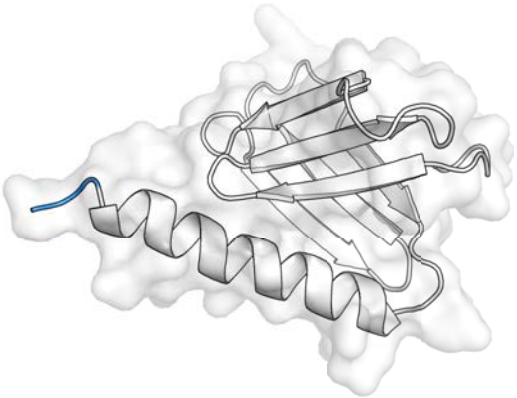

PF07679 I-set
−log(p−value)
Alignment position

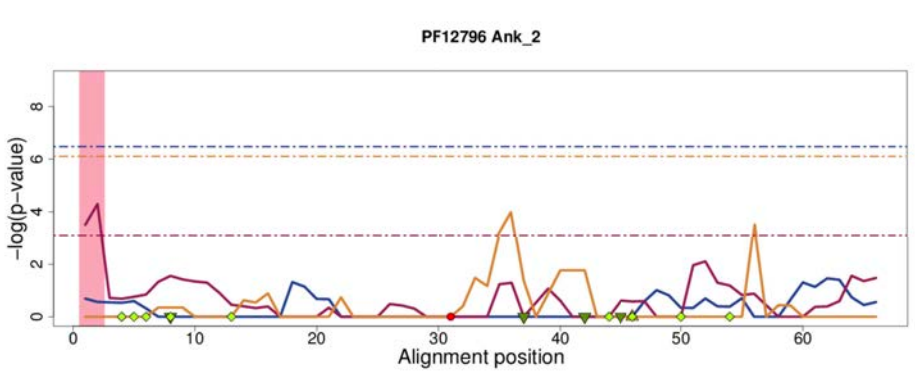


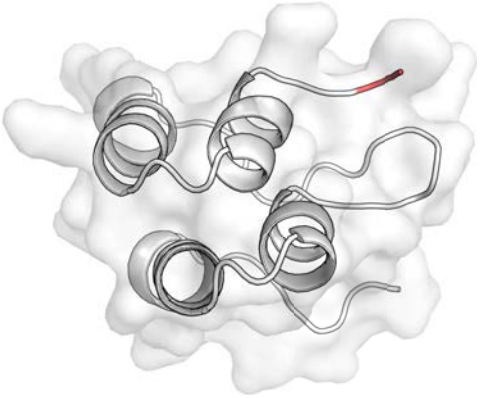

PF07690 MFS_1
−log(p−value)
Alignment position

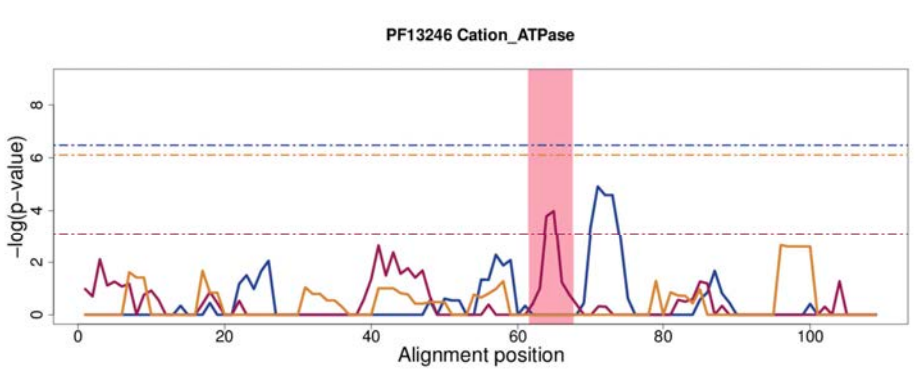


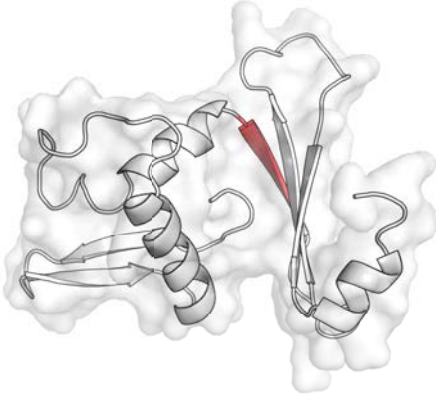

PF09380 FERM_C
−log(p−value)
Alignment position

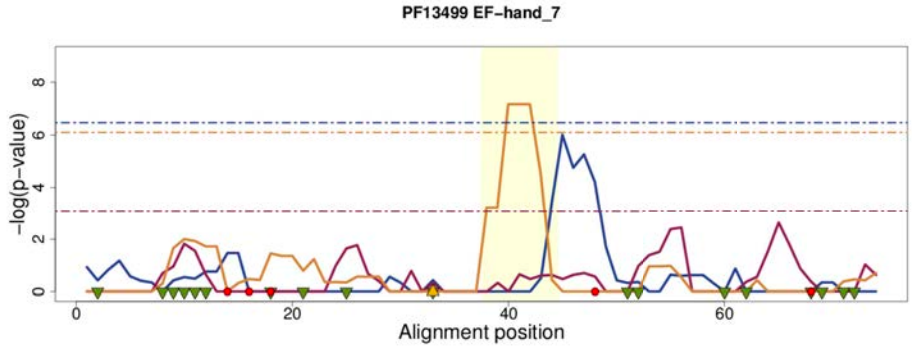


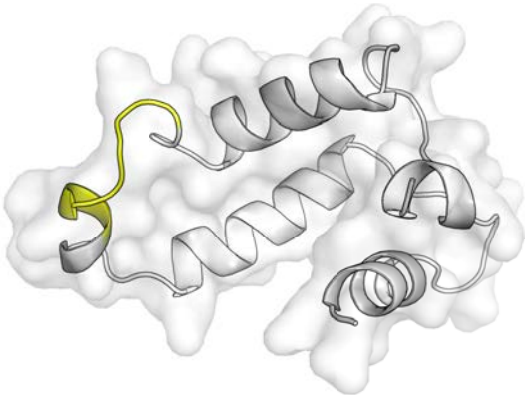

PF12796 Ank_2
−log(p−value)
Alignment position

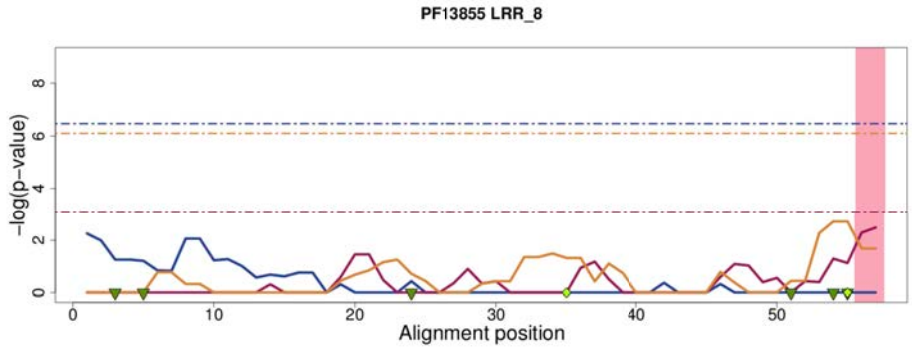



PF13246 Cation_ATPase
−log(p−value)
Alignment position




PF13499 EF−hand_7
−log(p−value)
0
2
4
6
8
0
20
40
60
Alignment position




PF13855 LRR_8
−log(p−value)
0
2
4
6
8
0
10
20
30
40
50
Alignment position