

UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI INGEGNERIA

Corso di Laurea Magistrale in Ingegneria Informatica



CHARACTERIZATION OF PHOSPHORYLATION SITES USING
3D STRUCTURAL INFORMATION

Relatore: Chiar.ma Prof.ssa Concettina Guerra

Correlatore: Dott.ssa Cinzia Pizzi

Tesi di Laurea di
Alessandro Del Bonifro

Anno Accademico 2009-2010

Ringraziamenti

Prima di tutto vorrei ringraziare la prof.ssa Guerra per la sua disponibilità a seguirmi, consigliarmi e anche un po' assecondarmi nel lavoro di questa tesi. Vorrei ringraziare anche la dott.ssa Pizzi per la sua collaborazione, il suo aiuto e la sua, sempre puntuale, disponibilità. Un grazie anche alla dottoressa Luisa per il suo incitamento e i suoi consigli per questa tesi.

Desidero poi ringraziare la mia famiglia per essermi stata vicina in questi anni ed in particolare i miei genitori, senza il loro apporto e supporto non sarei qui ora.

Un ringraziamento a tutti gli amici con cui ho condiviso l'esperienza universitaria e che mi hanno sopportato in questi 5 anni, Diego, Moreno, Lorenzo, Federico, Silvia, Irene e poi Giovanni, Rossella, Carlo Alberto e Marco (per quest'ultimi uno speciale ringraziamento per i "divertenti" progetti svolti durante la magistrale!). Un ringraziamento speciale a Silvia per essermi stata particolarmente vicina e a Diego la cui amicizia è sempre stata presente in questi anni.

Infine un ringraziamento a mio nonno per tutto ciò che mi ha insegnato e mi ha trasmesso, sarei la persona più felice del mondo se potessi essere qui con me ora ma mi ritengo la persona più fortunata del mondo per aver condiviso 23 anni della mia vita con te. Grazie

Abstract

Protein Phosphorylation is one of the most important and most studied modifications that a protein can undergo after its translation (post translational modifications). This is also a field in which bioinformatics efforts are most helpful understanding this fundamental process and, in particular, to predict protein phosphorylation. Since this problem has been an important bioinformatics research topic, about 20 years ago, for a number of reasons (lack of data, misleading assumption, etc) it has, almost always, been treated mostly with sequence-based methods. Although the local amino acid sequence may contain a significant part of the information contents related to phosphorylation, the local spatial environment may contribute significantly to the specificity of molecular event. Based on this consideration, the aim of this thesis is to explore ways in which three dimensional information can help to better understand and characterize protein phosphorylation mechanisms.

This thesis is organized as follows. First, in Chapter 1 a simple biological basis is provided to understand the argument treated. Chapter 2, after a review of current bioinformatics literature, presents the aims of this thesis in details, and the methods used to solve them. In Chapter 3 the software developed and implemented is analyzed and described. Then in Chapter 4 the results obtained with the software developed are presented with a detailed analysis. In the end in Chapter 5 a short summary of the work done and a discussion about future developments are given.

Contents

RINGRAZIAMENTI	III
ABSTRACT	V
CONTENTS	VII
1. INTRODUCTION	1
1.1 POST-TRANSLATIONAL MODIFICATIONS.....	1
1.2 PHOSPHORYLATION	2
1.3 PROTEIN KINASES.....	3
1.4 PREDICTION METHODS	4
1.5 SEQUENCE BASED APPROACHES.....	6
1.5.1 <i>SCANSITE</i>	6
1.5.2 <i>GPS</i>	8
1.5.3 <i>DISPHOS</i>	9
1.5.4 <i>KinasePhos</i>	10
1.5.5 <i>KinasePhos 2.0</i>	10
1.6 STRUCTURAL BASED APPROACHES	11
1.6.1 <i>Phos 3D</i>	11
1.6.2 <i>NETPHOS</i>	13
1.7 WEB DATABASES	14
1.7.1 <i>Phospho3D</i>	15
1.7.2 <i>UniProt (UNP)</i>	16

1.7.3	<i>Protein Data Bank (PDB)</i>	17
1.7.4	<i>Phospho.ELM</i>	18
1.7.5	<i>PhosphoSitePlus</i>	19
2	AIM & METHODS	21
2.1	AIM OF THE THESIS	21
2.2	METHODS.....	23
2.2.1	<i>Kinase(s) Characterization</i>	23
2.2.2	<i>Phosphorylation site(s) characterization</i>	24
2.2.3	<i>Flanking Sequence</i>	25
2.2.4	<i>Similarity of Histograms</i>	25
2.2.5	<i>Prediction Method</i>	28
3	SOFTWARE IMPLEMENTATION	31
3.1	LANGUAGE	31
3.1.1	<i>External libraries: BioJava</i>	32
3.2	ARCHITECTURE.....	33
3.2.1	<i>PSite</i>	33
3.2.2	<i>Kinase</i>	34
3.2.3	<i>PP3D</i>	35
3.3	MAIN FORM.....	36
3.3.1	<i>Load multiple site</i>	36
3.3.2	<i>Analyze single pdb substrate</i>	37
3.3.3	<i>Current Objects</i>	38
3.3.4	<i>Score possible Sites</i>	38
3.3.5	<i>Analyze loaded sites</i>	39
3.3.6	<i>Print</i>	41
3.3.7	<i>Add site</i>	43
3.4	IMPLEMENTATION DETAILS	43
3.4.1	<i>Dataset Loading</i>	43

3.4.2	<i>Prediction Algorithm</i>	47
3.4.3	<i>CutOff Algorithm</i>	48
4	RESULTS	55
4.1	THE DATASETS.....	55
4.2	KINASE SPECIFIC DATASET (KS DATASET).....	56
4.3	KINASE FAMILY SPECIFIC DATASET (KFS DATASET)	57
4.4	IMPORTANCE OF CONSIDERING KINASE SPECIFIC PHOSPHORYLATION SITES.....	61
4.5	COMPARATIVE ANALYSIS OF THE DISTANCE METRICS USED	65
4.5.1	<i>Not relevance of Flanking Distance</i>	65
4.5.2	<i>Importance of using different radii in phosphorylation site characterization</i>	66
5	CONCLUSION	69
5.1	WORK PERFORMED.....	69
5.2	FUTURE DEVELOPMENTS	71

1 Introduction

In this chapter the biological basis of this thesis is first provided. The importance of Post-Translational Modifications is discussed as well as the fundamental aspects of protein phosphorylation and its crucial role in cellular processes, especially in signaling pathways.

Bioinformatics efforts and computer science applications about protein phosphorylation are then motivated and discussed. Last, an analysis of the current literature regarding this field is presented.

1.1 Post-Translational Modifications

Post-Translational Modifications (PTMs) are chemical modifications of proteins after their translation. PTMs can extend the range of functions of proteins by attaching other biochemical functional group such as acetate, phosphate, various lipids and carbohydrates, by changing the chemical nature of an amino acid or by making structural changes, like the formation of disulfide bridges.

Various PTMs regulate the dynamics of proteins and are implicated in almost all cellular process, especially in the connections of cellular pathways. To understand these networks and process fully it is crucial to comprehend how their connections are regu-

lated, for example by means of PTMs, and which proteins can be modified as well as the effects and lifetime of the PTMs.

The constant growing amount of available data combined with the labor-intensive and expensive experimental methods have made, in recent year, *in silicio* prediction and study of PTMs as a popular alternative and complementary approach.

1.2 Phosphorylation

Among PTMs protein phosphorylation is the most studied example. Phosphorylation is the addition of a phosphate group to a protein or other organic molecule (Figure 1.1). Reversible protein phosphorylation is a ubiquitously occurring post-translational modification influencing many molecular processes in all complex cells. The most commonly observed phosphorylation affects serine, threonine and tyrosine residues [1] [2], although phosphorylation of aspartame has also been reported [3]. It is estimated that about one third of eukaryotic proteins undergo this reversible post-translational modification [4] [5].

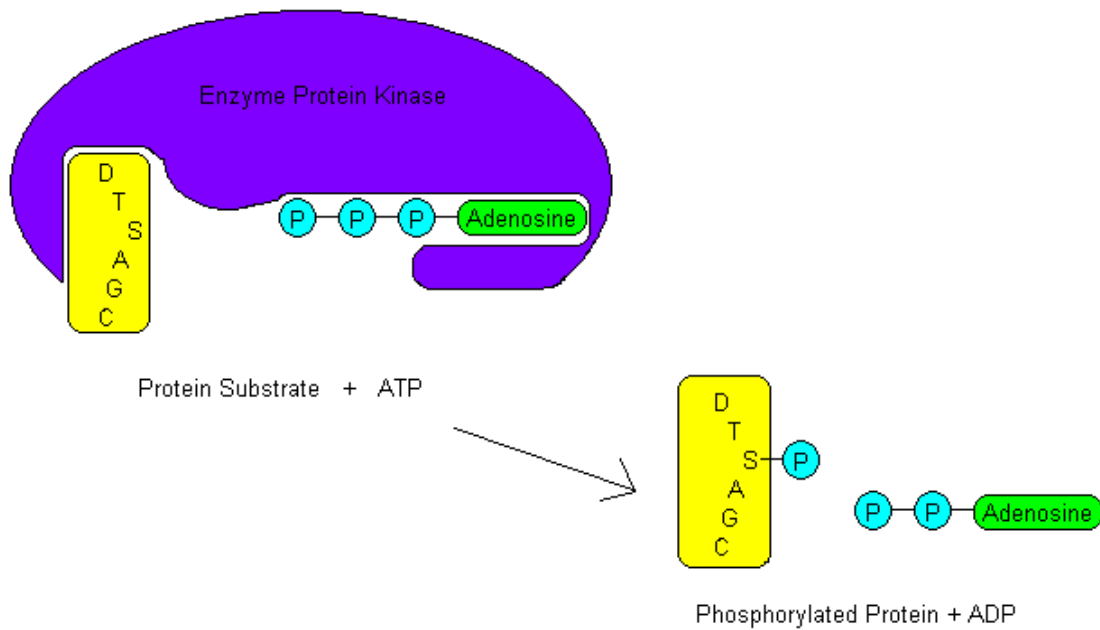


Figure 1.1 - Protein phosphorylation

1.3 Protein Kinases

Protein phosphorylation is catalyzed by enzymes called protein kinases, which are usually specific for either tyrosine or serine/threonine, with few of them being able to modify all three residues indistinguishably [6] [7] [8]. The protein kinases form one of the largest family of genes in eukaryotic [9] [10] and have been intensively studied. Protein kinases are related by virtue of their kinase domains (also known as catalytic domains) which encompass 250-300 amino acids [11] [12] [13]. More than 90% of protein kinases contain the eukaryotic protein kinase (ePK) catalytic domain that characterizes a single superfamily. 13 atypical protein kinase (aPK) families have been also identified. These contain proteins reported to have biochemical kinase activity, but which lack sequence similarity to the ePK domain. Nevertheless some aPK have structural similarity to ePK domain. As shown in Figure 1.2 from [9], 518 human protein kinases have been in-

identified, 478 of which ePK and 40 of which as aPK, constituting about 1.7% of all human genes.

Group	Families	Subfamilies	Yeast kinases	Worm kinases	Fly kinases	Human kinases	Human pseudogenes	Novel human kinases
AGC	14	21	17	30	30	63	6	7
CAMK	17	33	21	46	32	74	39	10
CK1	3	5	4	85	10	12	5	2
CMGC	8	24	21	49	33	61	12	3
Other	37	39	38	67	45	83	21	23
STE	3	13	14	25	18	47	6	4
Tyrosine kinase	30	30	0	90	32	90	5	5
Tyrosine kinase-like	7	13	0	15	17	43	6	5
RGC	1	1	0	27	6	5	3	0
Atypical-PDHK	1	1	2	1	1	5	0	0
Atypical-Alpha	1	2	0	4	1	6	0	0
Atypical-RIO	1	3	2	3	3	3	1	2
Atypical-A6	1	1	1	2	1	2	2	0
Atypical-Other	7	7	2	1	2	0	0	4
Atypical-ABC1	1	1	3	3	3	5	0	5
Atypical-BRD	1	1	0	1	1	4	0	1
Atypical-PIKK	1	6	5	5	5	6	0	0
Total	134	201	130	454	240	518	106	71

Figure 1.2 – Families of Protein Kinases

Protein kinases mediate most of the signal transduction in eukaryotic cells; by modification of substrate activity, protein kinases also control many other cellular processes, including metabolism, transcription, cell cycle progression, cytoskeletal rearrangement and cell movement, apoptosis, and differentiation. Protein phosphorylation also plays a critical role in intracellular communication during development, in physiological responses and in homeostasis, and in functioning of nervous and immune systems.

1.4 Prediction methods

The importance of the role of protein phosphorylation in cellular process and signal transduction, described above, brings the study of this PTM to be a central research topic in molecular biology. Given the high number of candidate phosphorylation sites (pSites), efforts to experimentally identify and verify them all remain challenging. These challenges combined with the high costs in time and money to experimentally find and study phosphorylation sites, and with the constant growing amount of available data of

protein sequence and also 3D structure, motivated the development of many prediction methods *in silicio*.

There are two main type of phosphorylation sites predictor: predictor of non-specific or organism-specific phosphorylation site, whose aim is simply to predict putative pSites of a given protein query, and of kinase-specific phosphorylation sites or phospho-binding motifs, that instead of predict general pSites search for specific kinase sites and/or for known phospho-binding motifs trying, in that way, to predict not only about the site but also about the kinase involved in the phosphorylation.

There are also two main types of approaches: sequence based approaches and structural based approaches. Sequence based approaches are more numerous, more accurate and especially more mature than the structural ones. This is easily explained by the fact that sequence [14] and domain [15] information and data of almost all proteins are known and available; moreover, by virtue of this amount of data, most of computer science application in molecular biology are involved in or based on sequence (protein sequence or DNA sequence) methods, so this kind of approach is the most common and solid used.

Besides, lack of structural data of proteins and, in particularly, of phosphorylation sites, together with reports that phosphorylation sites appear to be preferentially located in unstructured regions of proteins, have suggested a limited relevance of any structurally binding epitopes for the specific recognition of kinases and their substrate proteins [16].

However, in the last years, the significantly increased number of experimentally determined phosphorylation sites by proteomics technologies with simultaneously growing, but still insufficient, available 3D structures of the associated proteins [17], motivated to re-investigate the role of 3D-structural information for the specific recognition of kinases and their substrate proteins. To support this theory recent analyses [18] [19] suggest that the target site may very well assume defined structural conformations and, furthermore that phosphorylation sites may be surrounded by specific 3D-structural environments. Perhaps 3D structural information of proteins is very limited compared to the huge number of proteins in public database so these approaches still remain in their infancy.

Some of the more interesting and incisive developed methods and algorithms for the prediction and analysis of phosphorylation sites will be now presented and discussed. Methods discussed are differentiated between sequence based approaches and structural based approaches.

1.5 Sequence based approaches

Among sequence based approaches two of them, Scansite [20] and GPS [21], will be deeply analyzed while others predictors will be only cited and briefly presented.

1.5.1 SCANSITE

Scansite [20], searches for short sequence motifs within proteins that are likely to be phosphorylated by specific protein kinases or bind to domains such as SH2 domains, 14-3-3 domains or PDZ domains. Many of the motifs in Scansite were determined using oriented peptide library experiments.

Optimal phosphorylation sites for particular protein Ser/Thr kinases or protein Tyr kinases are predicted using the matrix of selectivity values for amino acids at each position relative to the phosphorylation site as determined from the oriented peptide library technique described in [22].

Optimal binding sites for SH2 domains, PDZ domains, 14-3-3 domains and other domains are determined using the matrix of selectivity values for amino acids at each position relative to an orienting residue as determined by the oriented peptide library technique described in [23].

These scoring matrices, which quantitatively indicate the preference for each amino acid type at each position within the domain's recognition motif, can then be used to score entire database of protein sequences to find a small number of proteins with high-ranking motif matches, indicating possible protein-protein interaction. The Motif scanner program utilizes an entropy approach that assesses the probability of a site matching the motif using the selectivity values and sums the logs of the probability values for each amino acid in the candidate sequence. The program then indicates the percentile ranking of the candidate motif in respect to all potential motifs in proteins of a protein database. When available, percentile scores of some confirmed phosphorylation sites for the kinase of interests or confirmed binding sites of the domain of interest are provided for comparison with the scores of the candidate motifs.

In the graphical output, the candidate motifs are superimposed on the predicted domain structure of the protein, a hot link to the domain families via Pfam is provided as well as the primary sequence at the motif and the percentile score. The program also

provides information about the surface probability of the region of the protein around the motif of interest.

1.5.2 GPS

GPS (Group-based Prediction System) [21] is a tool for the identification of protein phosphorylation sites with their cognate protein kinases. The chief hypothesis of the algorithm is that if two short peptides share high sequence homology, they may also bear similar 3D structures and biochemical properties. The dataset for the classification is retrieved from Phospho.ELM [24] and the protein kinases are classified using the *Manning et al* [9] method into a hierarchical structure of four levels, including group, family, sub-family and single protein kinase.

To predict kinase-specific phosphorylation site firstly, a phosphorylation site peptide PSP (m, n) is defined as a serine (S), threonine (T) or tyrosine (Y) amino acid flanked by m residues upstream and n residues downstream. Then the amino acid substitution matrix BLOSUM62 is used to calculate the similarity between two PSP (7, 7) peptides.

Given a putative PSP (7, 7) peptide, it will be compared with all known sites pairwise to calculate the substitution scores, separately. Then the average value of the substitution score is computed as the final prediction score of the given site.

To improve performance, because the BLOSUM62 and other matrices are optimized to evaluate the similarity between homologous proteins but may not be optimized for the similarity of two PSPs, modifications of the BLOSUM62 matrix are applied to optimize it for each protein kinase group.

Apart from the algorithm *Xue et al.* propose an interesting method for the prediction of potential substrates of a given kinase combining protein-protein interaction (verified and predicted) and detection of phosphorylation sites. Because only a short peptide flanking of a site is not sufficient for providing full specificity for a PK modification in vivo [25] [26] they argue that a kinase must at least “kiss” its substrate and then “say farewell” by direct or indirect interactions. So they adopted this “kiss and farewell” model to predict protein kinase Aurora-B substrates with their site from its interacting proteins.

So taking from DIP [27], BioGrid [28], MINT [29] and other interaction protein databases experimental protein-protein interactions and from STRING [30] predicted protein-protein interactions they constitute a database of interactions with protein kinase Aurora-B. This dataset was then made non redundant (containing in the end 51,529 entry) and was analyzed with the GPS algorithm searching for phosphorylation sites of Aurora-B protein kinase. 21 of 26 experimentally verified Aurora-B sites were found, in addition several novel substrates with potential sites were predicted.

Other available predictor tools will be now briefly presented.

1.5.3 DISPHOS

DISPHOS (DISorder-enhanced PHOSphorylation predictor) [16] is a machine learning method of prediction of phosphorylation sites based on the role of the disorder in the phosphorylation process. Analyzing more than 1500 experimentally verified phosphorylation sites, Iakoucheva and co-workers pointed out that the similarity in sequence complexity, amino acid composition, exhibility parameters, and other properties between phosphorylation sites and disordered protein regions suggests that intrinsic disorder in

and around the potential phosphorylation target site is an essential common feature for eukaryotic serine, threonine and tyrosine phosphorylation sites. Therefore this information is used in the feature vectors of the machine learning method.

1.5.4 KinasePhos

KinasePhos [31] is a web server for computationally identifying catalytic kinase-specific phosphorylation sites. Firstly, experimentally verified phosphorylation sites are extracted from available web database as positive sets and non-phosphorylated sites as negative sets. Then positive sets are categorized by catalytic kinases. Therefore profile Hidden Markov Models (HMMs) are learned from the site sequences in the positive sets to detect relationships between amino acids sequences.

After the models are learned they are evaluated with two cross validation methods, such as k-fold cross validation and leave-one-out cross validation. For each kinase-specific positive set of phosphorylation sites, the best performed model is selected and used to identify the phosphorylation sites within the input protein sequences.

1.5.5 KinasePhos 2.0

KinasePhos 2.0 [32] is the current release of KinasePhos [31], it adapts the sequence-based amino acid coupling-pattern analysis and solvent accessibility as new features for SVM (support vector machine) to characterize the phosphorylation site.

The feature of coupling-pattern $[XdZ]$ denotes the amino acid coupling-pattern of amino acid types X and Z that are separated by d amino acids. The coupling strength of $[XdZ]$, defined by coupling-pattern analysis, indicates the positive or negative correlation

of amino acids X and Z with respect to the distance d . Therefore the coupling strength is used to characterize phosphorylation sites also computing the differences between positive and negative set of phosphorylation proteins. Then SVM (support vector machine) to build the models and the cross validation are applied.

1.6 Structural based approaches

As discussed before, structural based approaches are still in their infancy though a big step as been made by Durek's and co-workers' work, Phos3D [33], which will be deeply used and cited during this thesis.

1.6.1 Phos 3D

Phos3D [33] is a web server for the prediction of phosphorylation sites in proteins, originally designed to investigate the advantages of including spatial information in pSites prediction. The approach is based on Support Vector Machines trained on sequence profiles enhanced by information from the spatial context of experimentally identified pSites. Phos3D is capable to predict kinase-specific phosphorylations by the serine kinases PKA, PKC, MAPK, and CKII, as well as by the tyrosine kinase SRC. The goal is to characterize phosphorylation sites by spatial amino acid propensity distributions to generate spatial signature motif and the subsequent assessment of this information to improve the predictors of phosphorylation sites in proteins. Because previous studies have shown that "one fits all" approaches, i.e. parameterization of the prediction method irrespective of kinase family, have led to only modest success rates, they investigate whether considering kinase family specific 3D-motifs may reveal improved prediction results.

The dataset of phosphorylation sites was obtained from the Phospho.ELM database [24]. Serine, threonine, and tyrosine residues that were annotated as phosphorylated were extracted from their native sequence together with six flanking amino acids on either side. To identify associated protein structures and the actual conformations and locations of the peptide motifs within their three-dimensional context, the Protein Data Bank (PDB) is screened for protein structures containing the 13-mer peptide sequence motifs associated with phosphorylation sites based on exact sequence matches.

For classification, Phos3D uses Support Vector Machines (SVM) implemented in the kernlab R-package by Alexandros et al. [34]. Two major themes were pursued: the analysis of phosphorylation in a kinase family specific fashion and the investigation on whether phosphorylation sites are characterized by specific three-dimensional structural motif, composed by amino acid not necessarily close in sequence.

The feature-vector (FV) used for the Support Vector Machines consist of chemical-physical amino acid properties for the sequence-information and spatial information based on amino acid distribution patterns in the spatial context of putatively phosphorylated sites. For the amino acid property components of the FV, values from a collection of 530 commonly used indices provided by the AAindex database [35] including hydrophobicity, solvent accessibility preferences, secondary and tertiary structure preferences, polarity, volume, and solvent accessibility, structural disorder indices and others are used. The vector consisted of 530 x 12 dimensions for every index and position around the central serine, threonine, or tyrosine, where the components were values from the respective index and 530 dimensions for the average index value of the particular sequence motif.

The spatial information component consists of the normalized distribution ratios. The ratios of amino acid residues within the local sequence, outside the local sequence, and irrespective of the position in the protein sequence were used for distances in a range of 2 to 10 Å between the putatively activated oxygen (β -hydrogen) in case of a central serine and threonine, or γ -carbon in case of tyrosine and the closest atom of all other amino acid residues, or between the interaction centers proposed by Park et al in [36]. Two different radial profiles are used: sequence local, i.e. residues that are outside the local sequence environment (> 6 residue position) and the general sequence profile, irrespective of the amino acid position in the protein sequence.

The performance of Phos3D was evaluated by the area under the Receiver Operator Characteristic-curve resulting from a 10-fold cross-validation.

Durek and co-workers reported that adding 3D-information to using only sequence information resulted in modest (up to 5%) improved prediction results for all three target amino acid types.

Another structural “experimental” approach, used in addition to a well known sequence based approach [37] is now presented in this section.

1.6.2 NETPHOS

NETPHOS [37] is one of the firsts methods of phosphorylation site prediction developed. It is based on a neural network approach that consider for the prediction phylogenetic trees, indicating the relationship between the proteins phosphorylated on the same amino acid (serine, threonine and tyrosine), and sequence logos, used for display-

ing specific features of complex sequence alignments around the phosphorylation site (using windows in the range from 21 to 25 residues).

In addition to the sequence based approach a 3D contact map up to 33 five residues centered in the phosphorylation site considering the probabilities of contacts between C α atoms is used to characterize 3D environment of phosphorylation sites always using a neural network method.

Even if the results reported are not excellent, also due to the scarcity of data at the time where the predictor was developed, this is the first attempt to characterize phosphorylation sites using structural information, suggesting for the first time that structural information could be very useful in phosphorylation sites prediction.

1.7 Web Databases

Some of the first difficulties encountered working on these “relatively” new problems are the lack of data and their dispersion and non homogeneity, when available. Looking for phosphorylation sites the main problem is the dispersion and non homogeneity of data, numerous web database are available with large amount of data overlapping. Looking for sequence and any kind of protein information UniProt [14] is the main and universally recognized access point for all these information. Even searching for 3D structures there is one clear access point, the Protein Data Bank [17], perhaps in this case clearly not all protein structure are available and this is a serious problem when using 3D information to attach a problem.

The main accessible databases of protein sequence and information, protein structure and phosphorylation sites, that have also been used for the developing of this thesis will be now presented.

1.7.1 Phospho3D

Phospho3D [38] is a database of 3D structures of phosphorylation sites. It collects information retrieved from the phospho.ELM database [24] and is enriched with structural information and diverse annotations at the residue level. In addition, the database stores the results of a large scale local structural comparison which suggest functional annotation of phosphorylation sites by 3D similarity. Cases of significant structural similarity between phosphorylation sites may indicate that they are phosphorylated by the same kinase.

The correspondence between phospho.ELM sequences and the Protein Data Bank (PDB) chains was established via the Seq2Struct resource [39], an exhaustive collection of annotated links between SwissProt-TrEMBL and PDB sequences.

The basic information stored in Phospho3D consists of the instance (i.e. the phosphorylation site), its flanking sequence (10 residues) and any residue whose distance from the instance does not exceed 12 Å, thus defining a 3D neighborhood, which is defined as zone.

In addition, for each zone the results of a large-scale local structural comparison versus a representative dataset of PDB protein chains from eukaryotic organisms are also given. The comparison was carried out using the Query3D sequence/fold indepen-

dent algorithm [40]. Structural matches are assessed by two criteria: structural similarity and biochemical similarity.

The structural similarity demands that matching residues have a root mean square deviation (r.m.s.d.) lower than a given threshold, whereas the biochemical similarity is evaluated using a Dayhoff substitution matrix. The score of the match is the number of matching residues which fulfill the similarity criteria. The significance of the score is evaluated by calculating the Z-score over the score distribution of the query zone comparison to the whole dataset.

For each match, the Z-score is computed as the difference between the score of the match and the average score of all the matches for the query patch, divided by the standard deviation.

1.7.2 UniProt (UNP)

The Universal Protein Resource (UniProt) [14] is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). The UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for metagenomic and environmental data.

UniProt is a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). Across the three institutes close to 150 people are involved through different tasks such as database curation, software development and support.

Until a few years ago, EBI and SIB together produced Swiss-Prot and TrEMBL, while PIR produced the Protein Sequence Database (PIR-PSD). These two data sets coexisted with different protein sequence coverage and annotation priorities. TrEMBL (Translated EMBL Nucleotide Sequence Data Library) was originally created because sequence data was being generated at a pace that exceeded Swiss-Prot's ability to keep up. Meanwhile, PIR maintained the PIR-PSD and related databases, including iProClass, a database of protein sequences and curated families. In 2002 the three institutes decided to pool their resources and expertise and formed the UniProt Consortium.

1.7.3 Protein Data Bank (PDB)

The Protein Data Bank [17] archive is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. These are the molecules of life that are found in all organisms including bacteria, yeast, plants, flies, other animals, and humans. The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule. They then deposit this information, which is then annotated and publicly released into the archive by the wwPDB.

The PDB was established in 1971 at Brookhaven National Laboratory and originally contained 7 structures. In 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) became responsible for the management of the PDB. In 2003, the wwPDB was formed to maintain a single PDB archive of macromolecular structural data that is freely and publicly available to the global community. It consists of organizations that act as

deposition, data processing and distribution centers for PDB data. In addition, the RCSB PDB supports a website where visitors can perform simple and complex queries on the data, analyze, and visualize the results.

1.7.4 Phospho.ELM

Phospho.ELM [24] is a resource containing experimentally verified phosphorylation sites manually collected from the literature and is developed as part of ELM (eukaryotic Linear Motif) resource. The latest version of Phospho.ELM, version 8.4 (June 2010), contains 8,718 substrate proteins from different species covering 3,370 tyrosine, 31,754 serine and 7,449 threonine experimentally verified phosphorylation sites.

The key information consists of the phosphorylated site and its flanking sequence (10 residues by each side) within a protein, for which experimental evidence has been found in literature. Moreover, annotations to each instance include (where known) the kinase(s) that phosphorylate(s) the given site, the domain(s) that bind to a phosphorylated motif and a link to ELM server to retrieve further information about the kinase. Furthermore, additional information for each protein kinase substrate includes the sub-cellular compartment, tissue distribution, a list of interaction partners derived from the MINT database and a diagram of a signaling pathway in which the protein is involved. The database can be searched by protein name (for the substrate), kinase name to get a list of known substrate, or by phosphopeptide-binding domain to retrieve all instances interacting with the given domain.

1.7.5 PhosphoSitePlus

PhosphoSitePlus, reengineered from PhosphoSite [41] is an open, dynamic, continuously curated, and highly interactive systems biology resource for studying experimentally observed PTMs in the regulation of biological processes. PhosphoSite was limited to phosphorylation. PhosphoSitePlus, while still providing comprehensive coverage of protein phosphorylation, now includes coverage of other commonly studied PTMs including acetylation, methylation, ubiquitination, and O-glycosylation. PhosphoSitePlus contains 66,894 non redundant phosphorylation sites and 11,252 non redundant proteins.

It includes critical structural and functional information about the topology, biological function and regulatory significance of specific modification sites, and powerful tools for mining and interpreting this data in the context of biological regulation, diseases, tissues, subcellular localization, protein domains, sequences, motifs, etc.

The database can be searched by protein substrate name or by kinase name to get a list of all substrates. Very useful are the PDB links provided for the protein substrate, within information about the chains and the residues present in the PDB structure.

2 Aim & Methods

In this chapter the aim of this thesis and the general goals are presented. The problems and the specific objectives are then analyzed and discussed; last, a structure-based method for studying protein phosphorylation is proposed.

2.1 Aim of the thesis

As discussed above, about one third of human proteins undergo protein phosphorylation and the importance of this PTM, especially in signaling pathways, has been recognized and studied by now for many years. This high importance together with the large costs in time and money to experimentally discover phosphorylation sites have stimulated the development of numerous *in silicio* prediction methods. Because of the lack of data only recently some prediction methods that use also 3D information have been developed [33] [37] but always in conjunction with sequence information. Therefore 3D approaches are still in their infancy and the well known importance of 3D structure in protein function and mechanisms makes them an important research topic.

The work of this thesis started with the aim to better understand mechanisms involved in insulin signaling pathway. In Figure 2.1 is showed a simplified version of the insulin signaling pathway; as can be easily seen, almost all the interactions regard a phosphorylation (+p) or a de-phosphorylation (-p) of a target protein substrates. So the high

number of protein kinases involved suggested to better investigate this fundamental mechanism of cellular signaling transduction, in fact most of the protein kinases collected and used on this work (see Chapter 4) are involved in the insulin signaling pathway.

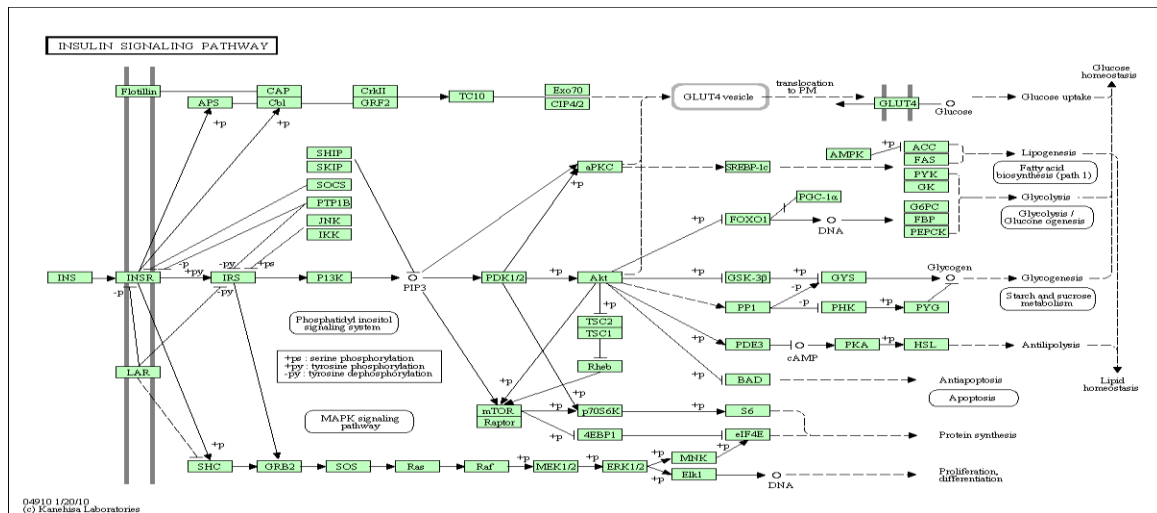


Figure 2.1 - insulin signaling pathway

Based on these considerations the aim of this thesis is to understand how three dimensional information can be used to better understand protein phosphorylation mechanisms, to analyze and find 3D features common to substrates of the same kinase or family of kinases and to improve phosphorylation site prediction in conjunction with developed and well established sequence based approaches.

Then the final aim is to develop a software tool able to characterize protein kinases using 3D information (of their substrates), providing some measures of similarity between phosphorylation sites. This characterization may ultimately lead to the design of methods able to predict possible phosphorylation sites or to find similarities among candidate phosphorylation sites and experimentally determined ones.

2.2 Methods

In the following sections I will discuss the problems, the objectives and the methods used to solve and pursue them.

2.2.1 Kinase(s) Characterization

As Durek and co-workers pointed out [33] previous studies have shown that parameterization of the prediction methods irrespective of kinase-family have led only to modest success rate.

Considering the fact that most of the phosphorylation sites are phosphorylated by a specific protein kinase and not by all protein kinases of a certain family, it would be interesting to investigate, in addition to a family characterization of phosphorylated sites, also a specific protein kinase characterization. In other words it would be interesting to understand if it is possible to find similarities among sites of different protein kinases in the same family.

This work is based on the key idea, common to other existing methods, that sites phosphorylated by the same kinase (or by kinases of the same family) should share common features; thus protein kinases will be analyzed and compared by analyzing and comparing their substrates. It would be an interesting result if it turns out that it is possible to discriminate among spatial environment of substrates of a single protein kinase and substrates of different protein kinases.

2.2.2 Phosphorylation site(s) characterization

Characterization of phosphorylation sites using 3D information can be done in several ways. When working with structural information and with protein modifications that involve a physical binding between the substrate and the kinase, ideally one would like to investigate surface similarity of binding regions. However, this seems not promising because a number of modular domains involved in phosphorylation, such as WW, SH2, SH3, PTB, PDZ and 14-3-3, bind to their ligands through direct interaction with very short amino acid sequences (typically < 10 amino acids). Furthermore, the interaction occurs in regions of the proteins that are often unstructured.

A good characterization already used by Durek et al. [33] is the propensity of amino acids to appear around the phosphorylation site. They proposed to evaluate the propensity of residues to occur in a sphere of radius 10 Å centered at the phosphorylation sites, differentiating between 3 profiles: sequence-local amino acids, i.e. counting only amino acids residues located within 6 residues position in the sequence from the site, non-local amino acids, i.e. counting only residues that are outside the local sequence environment (>6 residues position in the sequence from the site) and the general spatial profile, counting all the amino acids irrespective of their position in the protein sequence.

In this work a similar approach has been used: the spatial environment of a phosphorylation site is characterized by the propensity of the residues to appear in 3 spherical regions of different radii (of 4, 10 and 16 Angstrom) centered at the site.

Therefore the spatial environment around a phosphorylation site is represented as a bi-dimensional histogram 20x3 (20 amino acids for 3 radii).

2.2.3 Flanking Sequence

With flanking sequence only those amino acids near (<8 residues position) the phosphorylation site in the sequence environment are considered. It is interesting to analyze the spatial position of the amino acids of the flanking sequence to check whether they constitute a discernable factor to characterize the similarity of substrates of kinases, as happens considering only the sequence. Considering the importance of the flanking sequence in sequence-based approaches, it is expected to find all the residues of the flanking sequence close to the phosphorylation sites, i.e. in the first and the second sphere of radius 4 and 10 Å around the sites. What it is unknown is if their propensity to appear in the structure local environment tends to assume similar values in sites phosphorylated by the same kinase, or family of kinases, as in the sequence local environment.

Therefore to understand the importance of spatial position of the flanking sequence and the improved precision of considering 3 different radii instead of only one, we have also computed the propensity of sequence local amino acids (<8 residues position) with the three radii and with a single radius of 16 Å and the propensity of all amino acids with a single radius of 16 Å.

2.2.4 Similarity of Histograms

In a more general mathematical sense, a histogram is a function that counts the number of observations that fall into each of the disjoint categories (known as bins).

In our work a histogram is an array of 20 elements (one for each residue) that count the number of times that a particular residue appears in the spatial environment of a

phosphorylation site. More in specific, each cell of the histogram, representing an amino acid, counts the number of that particular amino acid that appear in the spatial environment of the phosphorylation site.

For comparing the spatial environments represented as histograms, a measure of similarity between histograms has to be chosen. The distance of a residue from another residue is calculated with respect to the centroids of the atoms of the amino acids.

Several measures of similarity of two histograms P and Q with N bins each are available, the main used are the following:

- Euclidean Distance, (Equation 2.1)

$$d_{eucl}(P, Q) = \sum_{i=1}^N (P_i - Q_i)^2$$

- Root Mean Square Distance (RMSD), (Equation 2.2)

$$RMSD(P, Q) = \sqrt{\frac{\sum_{i=1}^N (P_i - Q_i)^2}{N}}$$

- Correlation, (Equation 2.3)

$$Corr(P, Q) = \frac{N \sum_{i=1}^N (P_i Q_i) - \sum_{i=1}^N P_i \sum_{i=1}^N Q_i}{\sqrt{(N \sum_{i=1}^N P_i^2 - (\sum_{i=1}^N P_i)^2) (N \sum_{i=1}^N Q_i^2 - (\sum_{i=1}^N Q_i)^2)}}$$

While Euclidean distance and RMSD measure an effective distance between the histograms the correlation coefficient is used to measure a common trend in the value of the histograms.

To measure the distance between two or more histograms and therefore to have a measure of similarity between two or more phosphorylation sites the average Euclidean

distance has been used, i.e. given two histograms, P and Q, representing two phosphorylation sites their similarity is defined as (Equation 2.4):

$$\text{similarity}(P, Q) = \frac{\sum_{a=1}^{20} \sum_{r=1}^R (P_{ar} - Q_{ar})^2}{20 \times R}$$

Where the first sum is made on the second dimension of the histograms, representing the number of radius, that is typical 3, and the second sum is made on the first dimension of the histograms, representing the 20 amino acid.

This concept of similarity could be misleading: a high value of the similarity just defined corresponds to a low structural similarity and vice versa. It is important to keep in mind that this is a value that indicates a distance, therefore in this thesis the concept of similarity and Euclidean (Average) distance will be equally used.

Reassuming the measures calculated in this work between two or more sites are:

1. The average Euclidean distance just discussed, dividing the spatial context in different size concentric radius (up to 5);
2. The average Euclidean distance without dividing the spatial context in different radius;
3. The average Euclidean distance considering only the flanking sequences, dividing the spatial context in different size concentric radius (up to 5); also named *Flanking distance*;
4. The average Euclidean distance considering only the flanking sequences, without dividing the spatial context in different radius;

In the following sections we'll refer to the self similarity of a kinase as the average Euclidean distance (or another of the four similarity measures above) between all pair of sites of that kinase and to the cross similarity between two kinases as the average Euclidean distance (or another of the four similarity measures above) between all pair of sites of the two kinases.

2.2.5 Prediction Method

As mentioned above, the main goal of this thesis is not the development of a new prediction algorithm, anyway a simple algorithm that scans a pdb structure file and produces a list of predicted phosphorylation sites has been realized.

The method developed simply scans a pdb file, producing a histogram for each Serine, Threonine and Tyrosine residue and compares these histograms with the histograms of experimentally annotated phosphorylation sites collected with respect of their specific protein kinase (or kinases family). The result of the algorithm, summarized in the following pseudo-code, is a list of every Serine, Threonine, Tyrosine residues accompanied with a value for each kinase, used in the algorithm, representing the average similarity, as defined in Equation 2.4, between the query site and the annotated sites of that kinase.

- Step 1. For each Serine, Threonine and Tyrosine residue in the pdb query file
- Step 2. Create the histogram of the candidate site
- Step 3. For each kinase in the available dataset
- Step 4. Calculate the average similarity between the candidate site and the Kinases' sites
- Step 5. Return the list of candidate sites sorted with respect of the calculated scores

Clearly a low score of a residue with respect to a certain kinase indicates a high structural similarity between the spatial environment of that residue and the spatial environments of the experimentally verified sites of that kinase, suggesting a possible phosphorylation of the residue by that protein kinase.

3 Software Implementation

In this chapter the software developed in this thesis is presented. After a short introduction about the program language used, the architecture of the software developed (PP3D, Protein Phosphorylation 3D), is analyzed; the main functions of the main classes are provided.

Then the main form of the software is presented and all the available functions are described and showed.

In the end some crucial implementation details are discussed as well as the main algorithm developed and implemented, as the Prediction Algorithm and the CutOff algorithm, also with the assistance of pseudo-code.

3.1 Language

The software developed, ProteinPhosphorylation 3D (PP3D), has been written in Java (JDK 6).

Java is a well-known programming language originally developed by James Gosling at Sun Microsystems (which is now a subsidiary of Oracle Corporation) and released in 1995 as a core component of Sun Microsystems Java platform. The language derives much of its syntax from C and C++ but has a simpler object model and fewer low-level

facilities (like the garbage collector). Java applications are typically compiled to bytecode (class file) that can run on any Java Virtual Machine (JVM) regardless of computer architecture. Java is general-purpose, concurrent, class-based, and object-oriented, and is specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere". Java is currently one of the most popular programming languages in use, and is widely used from application software to web applications.

In 2007 Sun released all of Java's core code available under the terms of the GNU General Public License (GPL).

Java has been chosen mainly because of his portability and for the presence of the useful bioinformatics library, BioJava [42], described in the next section.

3.1.1 External libraries: BioJava

BioJava [42] is an open-source project, licensed under LGPL 2.1, dedicated to providing a Java framework for processing biological data. It provides analytical and statistical routines, parsers for common file formats and allows the manipulation of sequences and 3D structures. The goal of the biojava project is to facilitate rapid application development for bioinformatics.

Other similar parallel projects are dedicated to processing biological data, such as BioPerl and BioPhython.

In particular, the package *org.biojava.bio.structure* has been imported and used in PP3D for the parsing of the pdb files and for the manipulation of objects such as residues and atoms.

3.2 Architecture

PP3D is a simple Java package composed by 9 classes, 1 interface and about 5,500 lines of code. In the following sections the main classes of the program will be presented. PP3D can work with a dataset of kinases and relative phosphorylation sites whose pdb files have been loaded (sections 3.3.1 and 3.4.1). It can also work with a single pdb file, representing a protein containing a possible phosphorylation site, which can be searched for a known site, compared with loaded sites or searched for possible phosphorylation sites. In addition PP3D can also be used to find annotated phosphorylation sites and construct a new dataset or extend an existing one.

The dataset of loaded sites can be analyzed for discovering similarities between phosphorylation sites and can be used to predict and rank possible phosphorylation sites of a given pdb file.

The 3 main classes representing the objects of the architecture will be now presented.

3.2.1 PSite

This class represents a phosphorylation site. The main information representing the site, implemented as class fields, is:

- The name of the phosphorylated protein containing the site;

- The Uniprot Id of the phosphorylated protein containing the site;
- The phosphorylated residue, Serine, Threonine or Tyrosine;
- The chain containing the site;
- The sequence position of the site;
- The structure position in the pdb file of the site;
- The flanking sequence around the site (14 residues + 1 for the site);
- The name of the pdb file of the phosphorylated protein containing the site;
- The histogram of the propensity of residues to be in the proximity of the site;
- The histogram of the propensity of the flanking sequence to be in the proximity of the site.

The histograms representing the spatial propensity of residues to appear in 3 different spherical regions of radius (4, 10, 16 Å) centered in the phosphorylation site, as described in section 2.2.2, are implemented as bi-dimensional array 20 x 5, the first dimension representing the residue and the second the radius (there are 5 radii instead of 3 because also radius of 22 and 28 Å are evaluated even though, as will be shown in the “Results” chapter, they don’t provide useful information).

This class is responsible for the creation of the histograms and provides a static method, that can be called without instantiating an object, useful for create the histogram of a single phosphorylation site, given its pdb file, to be compared with the loaded ones.

3.2.2 Kinase

This class represents a protein kinase. The main information representing the kinase, implemented as class fields, is:

- The name of the protein kinase;
- The Uniprot id of the protein kinase;
- The target type of the protein kinase, ST for Serine/Threonine, Y for Tyrosine, STY if the kinase is able to phosphorylate all the three residues.
- A list of instance of class PSite, representing the sites phosphorylated by the protein kinase.

This class is responsible for the addition and the deletion of instances of class PSites to and from the list of sites; it also provides the methods that calculate the measures of similarity (described in section 2.2.4) between the phosphorylated sites of the kinase.

3.2.3 PP3D

This is the main class of the package, and controls all the operations that can be executed. It contains a list of instances of the class Kinase representing the protein kinases loaded and provides numerous methods to accomplish various operations, the most relevant are:

- Load a dataset of protein kinases with their respective phosphorylation sites;
- Verify the existence in a given pdb file of a given phosphorylation sites;
- Analyze the similarity between selected sites in the main form;
- Print information selected in the main form;
- Compare a single given site with loaded sites;
- Predict possible phosphorylation sites in a given pdb file comparing them with the loaded sites and producing a ranking.

3.3 Main Form

The main form, visible in Figure 3.1, displayed when executing PP3D, contains various *group box* that enclose commands for the accomplishment of all possible operations.

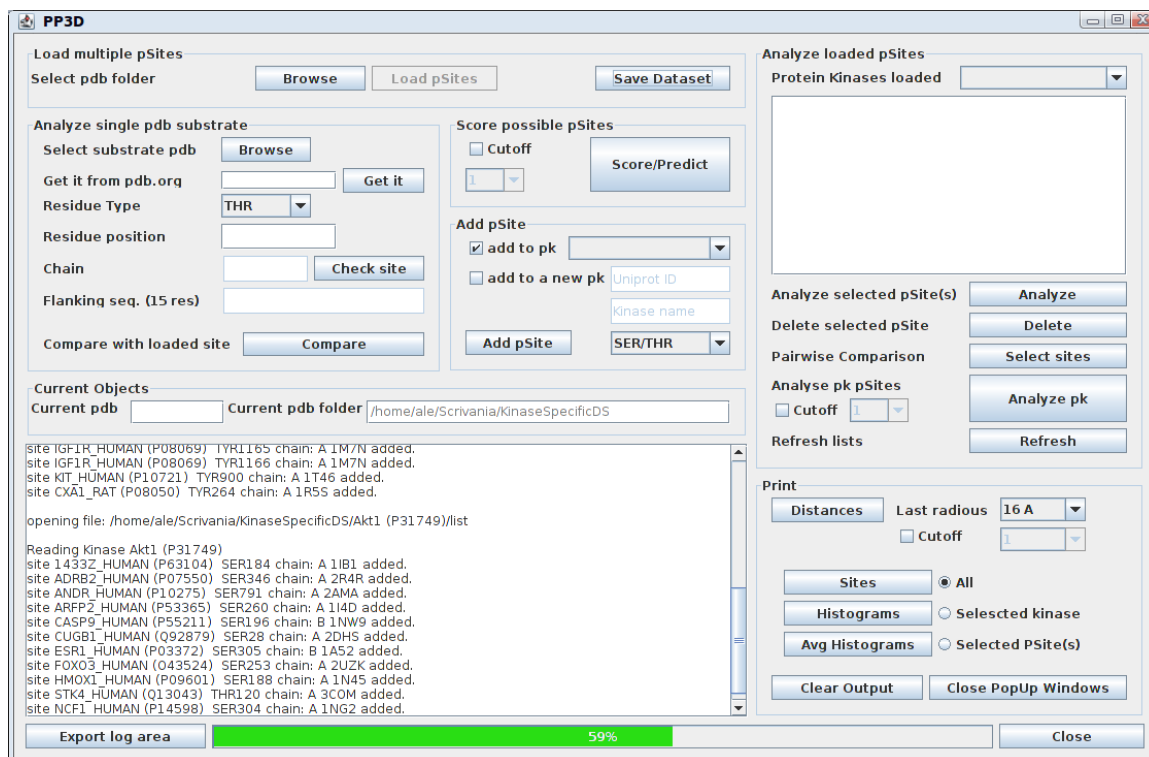


Figure 3.1 – Main Form

A large display area is used to print requested information, results of operations and to interact with the user. The display area can also be extended in a resizable pop-up window through the *Export log area* button.

The various group boxes and components will be now presented.

3.3.1 Load multiple site

It is possible to select (through the *Browse* button) the pdb folder containing the dataset to be analyzed and used for the prediction, and to load (through the *Load Sites*

button) the protein kinases and their phosphorylation sites contained in the selected pdb folder. It is also possible to save the current dataset (through the *Save Dataset* button) in the format described in section 3.4.1.

3.3.2 Analyze single pdb substrate

Here a single pdb (through the *Browse* button) can be selected for the analysis. If the pdb file selected is valid its key information about the molecules and their chains is displayed in the display area (Figure 3.2).

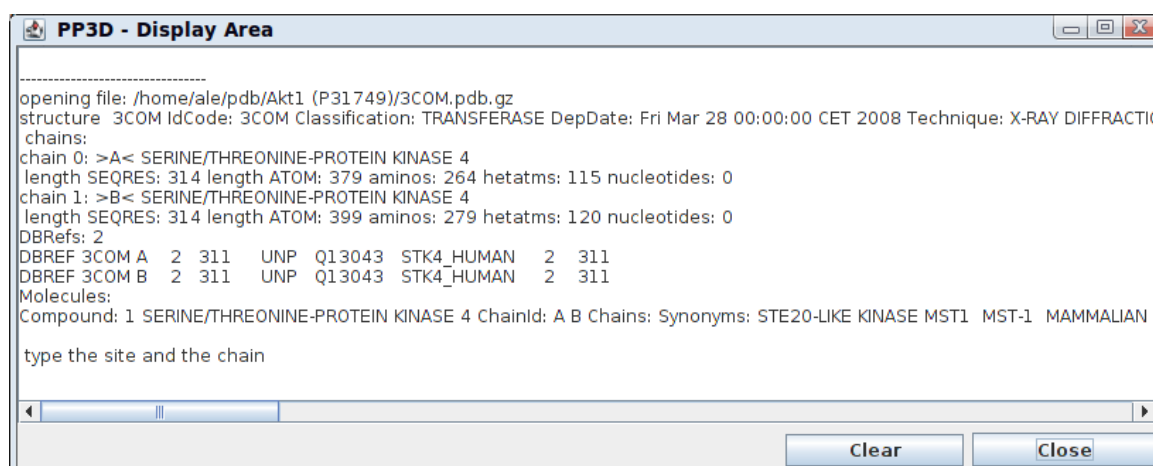
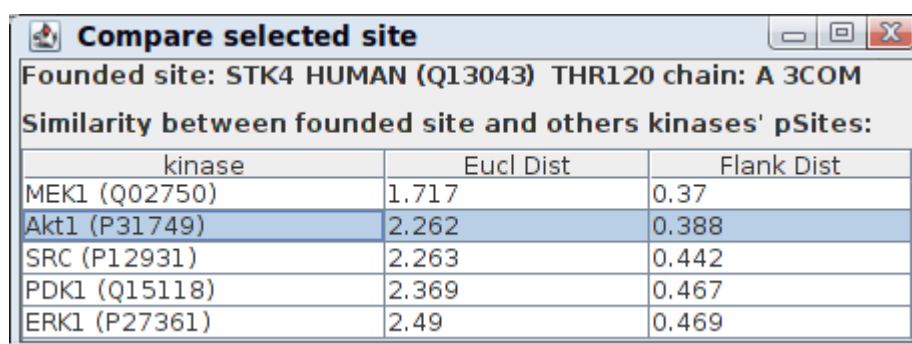


Figure 3.2 – pdb file information

In alternative it is possible to type in the text area the four letter pdb code of a pdb and try to recover it from *pdb.org* using through the *Get it* button.

To check if the loaded pdb file contains a certain phosphorylation site it is necessary to type in the proper text areas the residue position, the chain and select the residue type (THR, SER or TYR). By clicking the *Check site* button, it will be displayed if the pdb file in fact contains the given site.

If the given site has been found, using the *Compare* button, it is possible to view the similarity between the given site and the loaded sites of the dataset. In particular the average similarities between the given site and the site of each of the kinases present in the dataset will be displayed. The results of the comparison will be visible in the display area and in a pop-up window (Figure 3.3).



The screenshot shows a window titled "Compare selected site" with a table of similarity metrics. The table has three columns: "kinase", "Eucl Dist", and "Flank Dist". The rows list kinases: MEK1 (Q02750), Akt1 (P31749), SRC (P12931), PDK1 (Q15118), and ERK1 (P27361). The Akt1 row is highlighted in blue.

kinase	Eucl Dist	Flank Dist
MEK1 (Q02750)	1.717	0.37
Akt1 (P31749)	2.262	0.388
SRC (P12931)	2.263	0.442
PDK1 (Q15118)	2.369	0.467
ERK1 (P27361)	2.49	0.469

Figure 3.3 –Compare site

3.3.3 Current Objects

In this area are displayed the current pdb folder (i.e. the dataset folder), selected as described in section 3.3.1, representing the dataset of phosphorylation sites, and the current pdb file, selected as described in section 3.3.2.

3.3.4 Score possible Sites

As it will be discussed in section 3.4.2, through the *Score/Predict* button the selected chain of the current pdb file is scanned for all possible Serine, Threonine and Tyrosine phosphorylation sites.

These sites will be compared with the dataset and a ranking of predicted phosphorylation sites will be displayed. As it will be discussed in section 3.4.3 it is possible to select a cutoff value to prepare the dataset for the prediction.

3.3.5 Analyze loaded sites

In this area there is a display of the loaded kinases and, for the selected kinase, the loaded phosphorylation sites are listed. Four operations can be accomplished:

- Through the *Analyze* button will be displayed, in a pop-up window and in the display area, the similarity between the selected site(s) and the sites of all the kinases present in the loaded dataset. This operation is useful to understand if the selected site(s) is more similar to the other sites of its kinase than to the sites of other kinases.
- Using the *Delete* button the selected site will be deleted from the dataset;
- Clicking on *Select sites* a form for the selection of multiple sites in different kinases is displayed (Figure 3.4).

Pairwise Comparison

Select protein kinase 1: PDK1 (Q15118)

Select protein kinase 2: SRC (P12931)

Select pSite(s) (Left):

- AKT1_HUMAN (P31749) THR308 chain: A 3CQW
- AKT2_HUMAN (P31751) THR309 chain: A 3D0E
- ODPA_HUMAN (P08559) SER232 chain: A 1NI4
- ODPA_HUMAN (P08559) SER293 chain: A 1NI4
- ODPA_HUMAN (P08559) SER300 chain: A 1NI4
- SGK1_HUMAN (O00141) THR256 chain: A 2R5T

Select pSite(s) (Right):

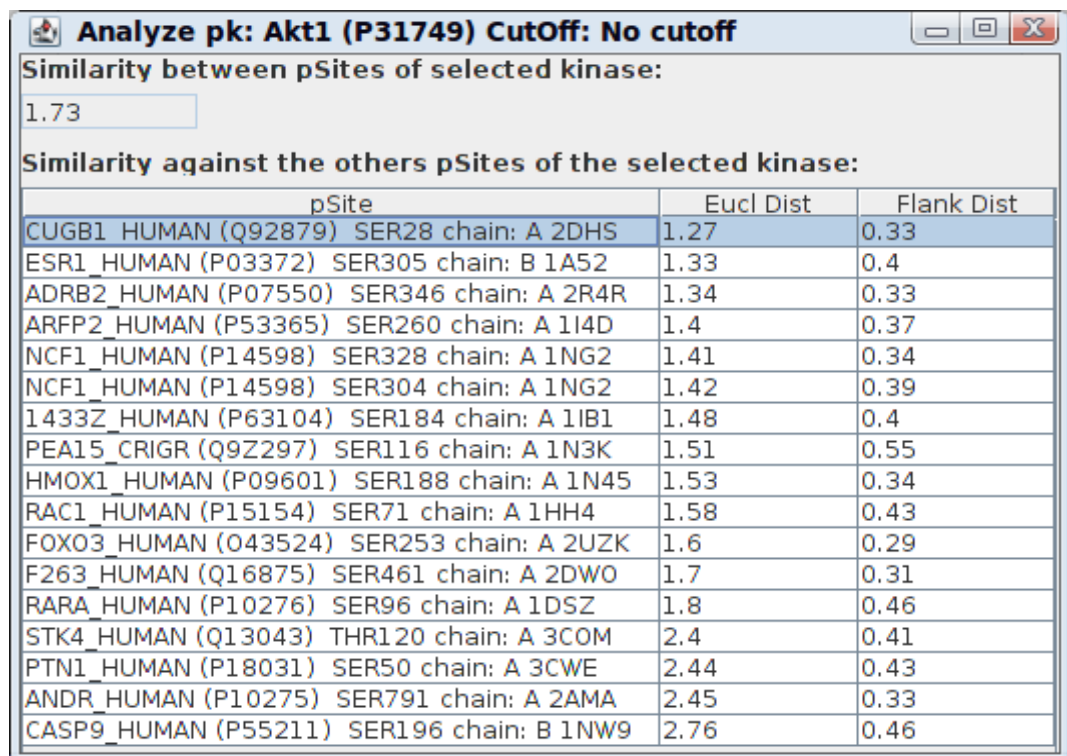
- ESR1_HUMAN (P03372) TYR537 chain: A 1GWQ
- ANX2_HUMAN (P07355) TYR23 chain: A 1W7B
- SRC_HUMAN (P12931) TYR215 chain: A 1A07
- SRC_HUMAN (P12931) TYR418 chain: A 1KSW
- SYUA_HUMAN (P37840) TYR125 chain: A 1XQ8
- IGF1R_HUMAN (P08069) TYR1161 chain: A 1M7N
- IGF1R_HUMAN (P08069) TYR1165 chain: A 1M7N
- IGF1R_HUMAN (P08069) TYR1166 chain: A 1M7N
- KIT_HUMAN (P10721) TYR900 chain: A 1T46
- CXA1_RAT (P08050) TYR264 chain: A 1R5S

Buttons: Compare, Close

Figure 3.4 - Pairwise Comparison Form

Once selected, the sites will be compared each other (through the *Compare* button) and with the kinases of the dataset, and the results of that comparison are displayed in the display area.

- With the *Analyze pk* button the similarity between the sites of selected protein kinase (self similarity of the kinase) is displayed in a pop-up table (Figure 3.5).



Analyze pk: Akt1 (P31749) CutOff: No cutoff

Similarity between pSites of selected kinase:
1.73

Similarity against the others pSites of the selected kinase:

pSite	Eucl Dist	Flank Dist
CUGB1_HUMAN (Q92879) SER28 chain: A 2DHS	1.27	0.33
ESR1_HUMAN (P03372) SER305 chain: B 1A52	1.33	0.4
ADRB2_HUMAN (P07550) SER346 chain: A 2R4R	1.34	0.33
ARFP2_HUMAN (P53365) SER260 chain: A 1I4D	1.4	0.37
NCF1_HUMAN (P14598) SER328 chain: A 1NG2	1.41	0.34
NCF1_HUMAN (P14598) SER304 chain: A 1NG2	1.42	0.39
1433Z_HUMAN (P63104) SER184 chain: A 1IB1	1.48	0.4
PEA15_CRIGR (Q9Z297) SER116 chain: A 1N3K	1.51	0.55
HMOX1_HUMAN (P09601) SER188 chain: A 1N45	1.53	0.34
RAC1_HUMAN (P15154) SER71 chain: A 1HH4	1.58	0.43
FOXO3_HUMAN (O43524) SER253 chain: A 2UZK	1.6	0.29
F263_HUMAN (Q16875) SER461 chain: A 2DW0	1.7	0.31
RARA_HUMAN (P10276) SER96 chain: A 1DSZ	1.8	0.46
STK4_HUMAN (Q13043) THR120 chain: A 3COM	2.4	0.41
PTN1_HUMAN (P18031) SER50 chain: A 3CWE	2.44	0.43
ANDR_HUMAN (P10275) SER791 chain: A 2AMA	2.45	0.33
CASP9_HUMAN (P55211) SER196 chain: B 1NW9	2.76	0.46

Figure 3.5 – Analyze pk

For each site of the selected kinase the similarity against the other sites of the kinase is also showed. This is very useful to understand if some site has a spatial environment very different from the others sites, in other words an outlier that generates noise in the computation of the average Euclidean distance of the sites of the kinases. As discussed in section 3.4.3, it is possible to choose a cutoff value to exclude from the analysis those phosphorylation

sites that have an Euclidean distance, with respect to the other sites of the kinases, higher than the selected threshold value.

3.3.6 Print

The following information can be printed in the display area or in pup up tables:

- Through the *Distances* button, for each kinase, the Average Euclidean distance between the kinase's sites and the others 3 measures discussed in section 2.2.4 (self similarity) are printed in a pop-up window (Figure 3.6); also the averages Euclidean distances, and the others 3 measures discussed in section 2.2.4, between all pair of kinases are displayed (cross similarity).

Self Similarity					Cross Similarity				
kinase	Eucl Dist	Generic ED	Flank Dist	Generic FD	kinases	Eucl Dist	Generic ED	Flank Dist	Generic FD
MEK1 (Q02750)	1.59	4.47	0.27	0.55	SRC (P12931) - Akt1 (P31749)	1.76	6.09	0.43	1.38
SRC (P12931)	1.65	5.37	0.43	1.33	SRC (P12931) - PDK1 (Q15118)	1.86	6.24	0.39	1.14
PDK1 (Q15118)	1.65	5.59	0.31	0.87	MEK1 (Q02750) - PDK1 (Q15118)	1.92	6.34	0.32	0.85
Akt1 (P31749)	1.73	6.36	0.39	1.18	Akt1 (P31749) - PDK1 (Q15118)	1.97	7.08	0.41	1.28
ERK1 (P27361)	2.48	9.2	0.45	1.37	Akt1 (P31749) - ERK1 (P27361)	2.11	7.61	0.43	1.31
					MEK1 (Q02750) - SRC (P12931)	2.13	6.29	0.4	1.18
					SRC (P12931) - ERK1 (P27361)	2.16	7.18	0.45	1.39
					MEK1 (Q02750) - Akt1 (P31749)	2.2	7.44	0.36	1.07
					ERK1 (P27361) - PDK1 (Q15118)	2.3	8.2	0.41	1.2
					MEK1 (Q02750) - ERK1 (P27361)	2.36	7.6	0.39	1.15

Figure 3.6 - Distances

In this way it can be checked if the spatial environment of the phosphorylated sites is discernable, i.e. the similarity of the sites of the same kinase (self similarity) is higher than the similarity between sites of different kinases (cross similarity). If this happens it means that sites phosphorylated by the same kinase share common structural features that are well captured by PP3D's site characterization.

By default the distances are calculated and displayed using as last radius 16 Angstroms (in this case $R=3$ in Equation 2.4), but also the values 4, 10, 16 and

22 Angstroms can be selected as last radius. This is useful to understand which dimension of the spatial environment best characterizes phosphorylation sites. As discussed in section 3.4.3, it is possible to choose a cutoff value to exclude from the calculation of the distances those phosphorylation sites that have an Euclidean distance, with respect to the others sites of the same kinase, higher than the selected threshold value; as said before (section 2.2.4), in addition of the Average Euclidean Distance (2nd column) just explained, also the same distance counting only the flanking sequence is computed (4th column). While in the 3rd and in the 5th column the same distances (Average Euclidean Distance and its version counting only the flanking sequence) are respectively computed without dividing the spatial context in different radius.

- Using the *Sites* button, information about the phosphorylation site(s) is listed with the option of listing phosphorylation sites of all kinases, or of selected kinases or of selected site.
- Clicking the *Histograms* button, histogram(s) of phosphorylation sites of all kinases, of sites of selected kinase or of a selected site are displayed according to the option selected.
- With the *Avg Histograms* button, the average histogram(s) of selected kinase or of all kinases is/are displayed. The average histogram of a protein kinase is simply the average histogram of the histograms of its sites.
- Through the *Clear Output* button, the display area is cleared.

- Using the *Close PopUp Windows* button, all the pop-up windows, including the exported display area, the *Pairwise Comparison* window and all the results windows opened, are closed.

3.3.7 Add site

In this section the current phosphorylation site (selected or found as described in section 3.3.2) can be added to the dataset loaded. The site can be added to an existent protein kinase or to a new protein kinase (obviously in this case also the new protein kinase is added to the dataset). With this option it is possible to rapidly load a pdb file and check if it contains an annotated phosphorylation sites (section 3.3.2), add it to a kinase and save the dataset (section 3.3.1); through these few steps the construction of a dataset of phosphorylation sites, to be analyzed by PP3D, can be done in a very simple and fast way.

3.4 Implementation details

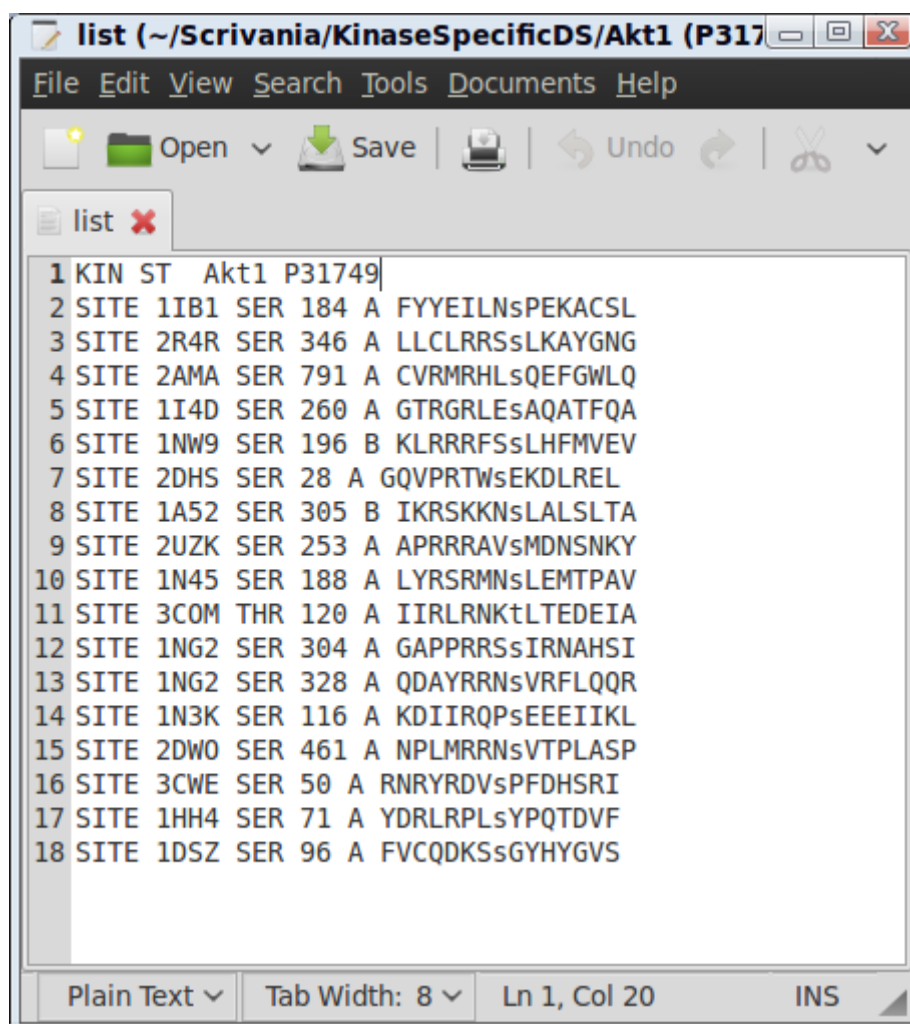
The main implementation details regarding methods adopted and implemented in PP3D will be now described.

3.4.1 Dataset Loading

A dataset of phosphorylation sites is simply a collection of folders each containing pdb files representing proteins phosphorylated by the same kinase or by the same family of kinases. The kinases folders are included in a root folder that can be selected, as described in section 3.3.1, to load the dataset.

Each folder representing a specific kinase (or kinase family) must contain a configuration text (.txt) file named *list.txt*. This file contains information regarding the kinase and its phosphorylation sites.

As can be viewed in Figure 3.7, the format of the *list.txt* file is in a pdb-like format where the information included in each row depends on the pattern at the beginning of the row.



```
1 KIN ST Akt1 P31749|
2 SITE 1IB1 SER 184 A FYYEILNsPEKACSL
3 SITE 2R4R SER 346 A LLCLRRSsLKAYGNG
4 SITE 2AMA SER 791 A CVRMRHLsQEFGLQ
5 SITE 1I4D SER 260 A GTRGRLEsAQATFQA
6 SITE 1NW9 SER 196 B KLRRRFsLHFMVEV
7 SITE 2DHS SER 28 A GQVPRTWsEKDLREL
8 SITE 1A52 SER 305 B IKRSKKNsLALSITA
9 SITE 2UZK SER 253 A APRRRAVsMDNSNKY
10 SITE 1N45 SER 188 A LYRSRMNsLEMPAV
11 SITE 3COM THR 120 A IIRLNKtLTEDEIA
12 SITE 1NG2 SER 304 A GAPRRSsIRNAHSI
13 SITE 1NG2 SER 328 A QDAYRRNsVRFLQQR
14 SITE 1N3K SER 116 A KDIIRQPsEEEEIKL
15 SITE 2DW0 SER 461 A NPLMRRNsVTPLASP
16 SITE 3CWE SER 50 A RNRYRDVsPFDSRI
17 SITE 1HH4 SER 71 A YDRLRPLsYPQTDVF
18 SITE 1DSZ SER 96 A FVCQDKSsGYHYGVS
```

Figure 3.7 - list.txt file

The fundamental rules are:

- Every line starting with the character '#' will be ignored and viewed as a comment;
- The first valid (not a comment) row must start with the pattern 'KIN', this line must contain information about the kinase;
- The valid lines following the 'KIN' line must start with the pattern 'SITE' and must contain information about a phosphorylation site of the kinase;
- Because lines are scanned by tokens, distinct information must be separated simply by one or more spaces.

The tokens, separated by one or more spaces, of the line containing the kinase information must be the following:

- 1st. The 'KIN' pattern;
- 2nd. The specificity of the protein kinase (ST for Ser/Thr, Y for Tyr and STY if the kinase is able to phosphorylate all the three residues);
- 3rd. The name of the protein kinase or of the family of protein kinases that phosphorylated the sites contained in the folder;
- 4th. The Uniprot id of the protein kinase or, if the sites contained in the folder are the sites phosphorylated by a family of protein kinases, the string 'FAMILY'.

The tokens, separated by one or more spaces, of the lines containing phosphorylation site information must be the following:

- 1st. The 'SITE' pattern;

- 2nd. The four letter code of the pdb file containing the phosphorylation site;
- 3rd. The residue type phosphorylated (THR, TYR or SER);
- 4th. The position of the phosphorylation site;
- 5th. The chain containing the phosphorylation site;
- 6th. The flanking sequence plus the site (15 residues or less) in the one letter format;
- 7th. (Optional) The pattern '3D' if the residue position refers to the numbering of the pdb file. Typically the residue position found in web database refers to the sequence position and sometimes this sequence position differs from the position contained in the pdb file of the same protein. Therefore the software automatically aligns the position with the possible different numbering of the pdb file calculating the offset. In this way, if the position refers to the pdb numbering, this must be specified so that the offset is not calculated.

The method *PP3D.loadPSites(String path)*, where the variable *path* is the path of the root file, selected as described in section 3.3.1 and containing the protein kinases folders, scans the root folder for folder containing the *list.txt* file.

When the *list.txt* file is found the method reads the information about the kinase and adds it, if it isn't already present, to the loading dataset.

For each phosphorylation file listed the method searches for the pdb file in the folder or, if not present, tries to recover it from *pdb.org*. The site's information is then read and the phosphorylation site is searched in the pdb file and, if found, it is added to the corresponding kinase.

3.4.2 Prediction Algorithm

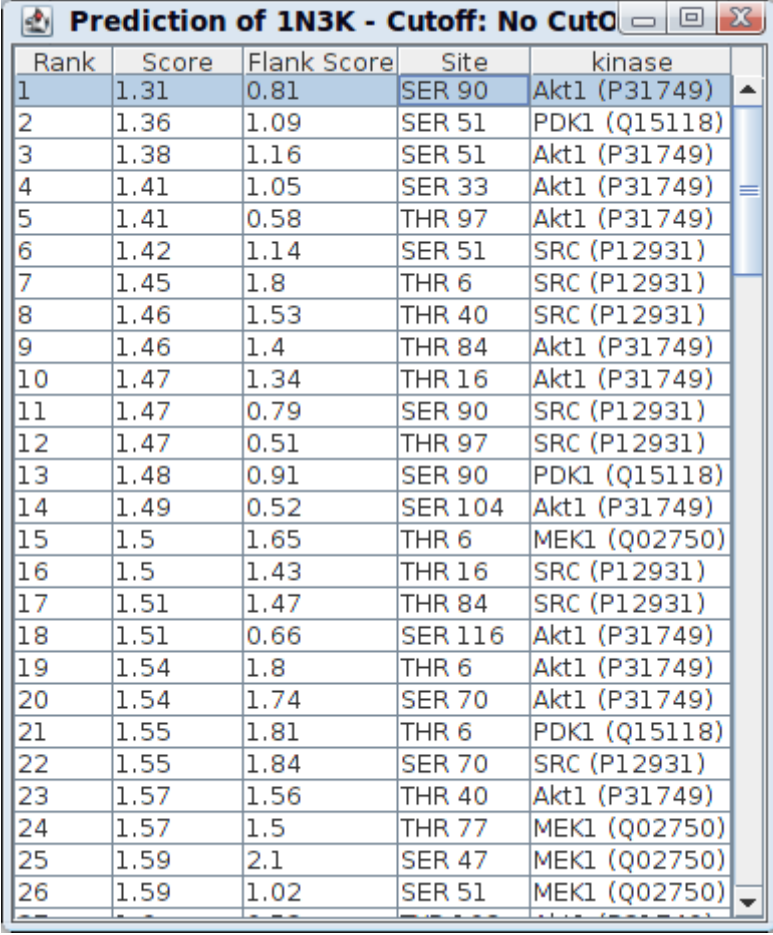
The prediction method described in section 2.2.5, has been implemented in the *PP3D.predictSites* method. The prediction algorithm takes in input a pdb file and compares all Serine, Threonine and Tyrosine residues with the dataset of phosphorylated sites, ranking the possible phosphorylation sites according to their average Euclidean distance with the sites of each kinases in the dataset.

To execute the prediction algorithm you just have to load a dataset of phosphorylated sites (as described in section 3.3.1) and a single pdb file (as described in section 3.3.2) to search for possible phosphorylation site as described in section 3.3.4.

To produce the ranking of possible phosphorylation sites of a given pdb substrate query file, similar of the example in Figure 3.8, the prediction algorithm, implemented in the *PP3D.predictSites* method, execute the following steps:

- Step 1. Create an empty Ranking List.
- Step 2. For each Serine, Threonine and Tyrosine candidate residue in the pdb query file
 - Step 3. Create the histogram of the candidate site
 - Step 4. For each Kinase in the loaded dataset
 - Step 5. Calculate the average Euclidean distance between the candidate and the Kinases' sites
 - Step 6. Add the candidate with the score to the Ranking List
- Step 7. Return the Ranking List sorted

The Ranking List returned, determined as described in section 3.3.4 is displayed, as in Figure 3.8, in a pop up table.



Rank	Score	Flank Score	Site	kinase
1	1.31	0.81	SER 90	Akt1 (P31749)
2	1.36	1.09	SER 51	PDK1 (Q15118)
3	1.38	1.16	SER 51	Akt1 (P31749)
4	1.41	1.05	SER 33	Akt1 (P31749)
5	1.41	0.58	THR 97	Akt1 (P31749)
6	1.42	1.14	SER 51	SRC (P12931)
7	1.45	1.8	THR 6	SRC (P12931)
8	1.46	1.53	THR 40	SRC (P12931)
9	1.46	1.4	THR 84	Akt1 (P31749)
10	1.47	1.34	THR 16	Akt1 (P31749)
11	1.47	0.79	SER 90	SRC (P12931)
12	1.47	0.51	THR 97	SRC (P12931)
13	1.48	0.91	SER 90	PDK1 (Q15118)
14	1.49	0.52	SER 104	Akt1 (P31749)
15	1.5	1.65	THR 6	MEK1 (Q02750)
16	1.5	1.43	THR 16	SRC (P12931)
17	1.51	1.47	THR 84	SRC (P12931)
18	1.51	0.66	SER 116	Akt1 (P31749)
19	1.54	1.8	THR 6	Akt1 (P31749)
20	1.54	1.74	SER 70	Akt1 (P31749)
21	1.55	1.81	THR 6	PDK1 (Q15118)
22	1.55	1.84	SER 70	SRC (P12931)
23	1.57	1.56	THR 40	Akt1 (P31749)
24	1.57	1.5	THR 77	MEK1 (Q02750)
25	1.59	2.1	SER 47	MEK1 (Q02750)
26	1.59	1.02	SER 51	MEK1 (Q02750)

Figure 3.8 - Output Prediction Algorithm

3.4.3 CutOff Algorithm

As mentioned in section 3.3.4 it is possible to select a cut off option in order to pre-prepare and normalize the dataset before the execution of the prediction algorithm. The same option is available when calculating overall distances (section 3.3.6) and analyzing a protein kinase's site (section 3.3.5)

The motivation of this option is to avoid during the prediction algorithm (or in general, in operations involving analysis of similarity) the "noise" generated by those phos-

phorylation sites that have a similarity value with respect to the other sites of the same kinases higher than the threshold value selected, excluding them from the scoring process.

For example, Figure 3.9 shows the result of the *Analyzepk* action (described in section 3.3.5) on the Kinase ERK1, contained in the dataset described in the next chapter; the average value of the Euclidean distance between the ERK1 sites is 2.48, but if we consider each site separately, it is clearly visible that the site of SYK_HUMAN protein has an average Euclidean distance with respect to the others sites of the kinase ERK1 much higher than the others, 4.38 with respect to the maximum of 2.45 of protein THB1_HUMAN.

pSite	Eucl Dist	Flank Dist
RAB5A_HUMAN (P20339) SER123 chain: A 1N6I	1.87	0.43
GSK3B_HUMAN (P49841) THR43 chain: A 1I09	1.93	0.49
RXRA_HUMAN (P19793) SER260 chain: A 1K74	2.01	0.47
CDK2_HUMAN (P24941) THR160 chain: A 1B38	2.3	0.48
BIEA_HUMAN (P53004) SER230 chain: A 2H63	2.43	0.41
THB1_HUMAN (P10828) SER142 chain: B 2NLL	2.45	0.39
SYK_HUMAN (Q15046) SER207 chain: A 3BJU	4.38	0.46

Figure 3.9 – Analyze pk ERK1

It is clear that when the average Euclidean distance with the ERK1's sites is calculated the SYK_HUMAN protein's site introduces a noise in this value; by introducing a cut off of 3 or less this site will be considered an outlier and ignored in the prediction algorithm (or in the calculation of overall distances).

Thus, if we select ERK1 sites excluding the SYK_HUMAN protein, as in Figure 3.10 the similarity value between the selected site obtained is 1.72, much lower than 2.45.

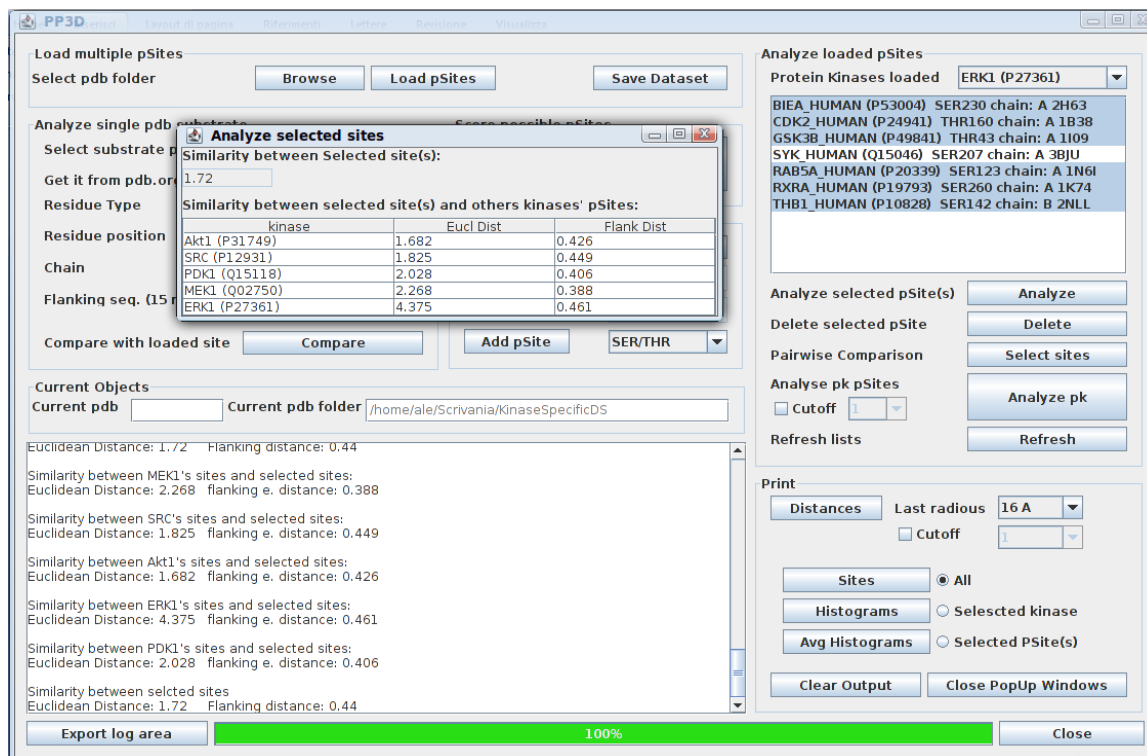


Figure 3.10 – Analyze pk ERK1's selection sites

Moreover if we delete the SYK_HUMAN protein from the dataset and analyze again the protein ERK1 (Figure 3.11) we see how the highest value of Euclidean distance with respect to the other site is now 2.09, confirming an increased similarity and homogeneity of kinase ERK1's sites.

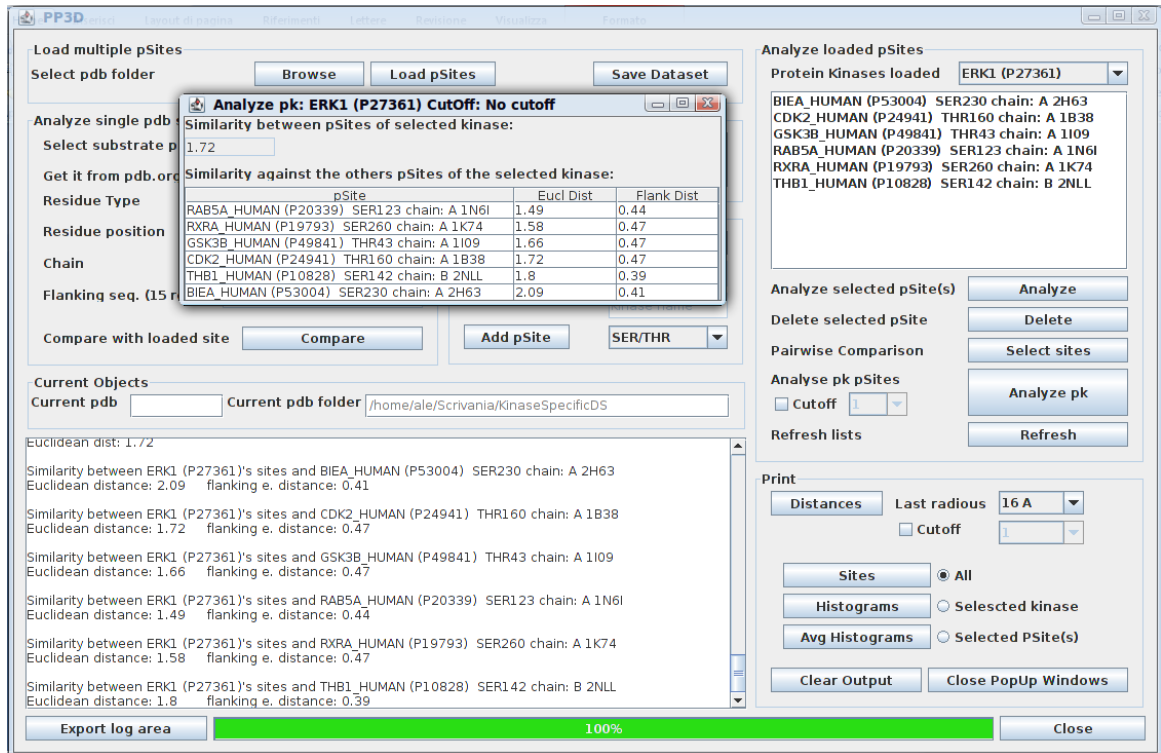


Figure 3.11 – Analyze pk ERK1 after deletion of SYK_HUMAN site

To accomplish this operation, every instance of the *PSite* class has a Boolean field, *PSite.outlier*, which assumes the value true when the site has to be ignored in the prediction algorithm (or in calculating overall distances)

The steps of the *PP3D.PreparePrediccion (double cutoff)* method called before the prediction algorithm, the *Print Distances* operation or the *Analyzepk* operation, if the cutoff option is selected, are the following:

Step 1. Set all PSite of all kinases as non outliers (*PSite.outlier = false*)

Begin loop

Step 2. For each Kinase in the loaded dataset

Step 3. For each PSite of the Kinase

Step 4. Calculate the average Euclidean distance between the PSite and the others instances of PSite of the Kinase

Step 5. If the distance calculated in Step 4 is higher than the threshold value set the PSite as an outlier (*PSite.outlier = true*)

End loop

Step 6. If no one site has been set as an outlier in the loop then terminate the procedure.

Step 7. If the loop has been executed for 3 times then terminate the procedure

Step 8. Repeat from Step 2

The loop Step2-Step5 scans all PSite of all Kinases verifying, for each sites, the condition to be or not to be an outlier.

As expressed in Steps 6 and 7 the procedure ends when one of the following conditions is true:

- No one site has been set as an outlier in the last loop;
- The loop has been executed for 3 times;

With the addition of the Prepare Prediction algorithm Step 5 of the Prediction algorithm became:

- Step 5. Calculate the average Euclidean distance between the candidate and the Kinases' sites **that are not been flagged as outliers**

The results of the Prepare Prediction algorithm are displayed in the displayed area every time the cut off option has been selected (and so the method executed). As an example the result of the Prepare Prediction algorithm with cut off 2.5 is showed in Figure 3.12, in this case only one site of kinases MEK1, Akt1 and ERK1 has been set as outlier.

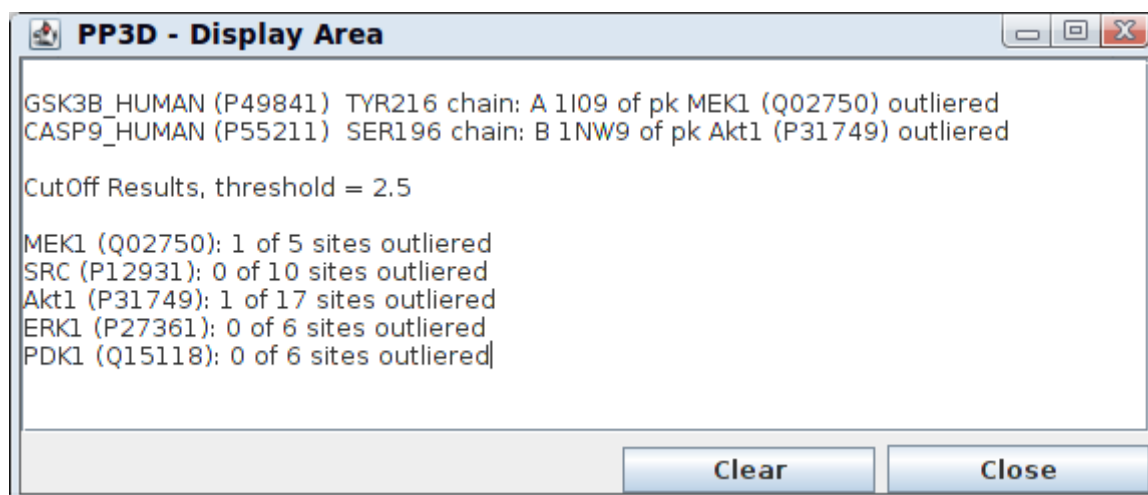


Figure 3.12 – Cut Off algorithm

4 Results

In this chapter I present the results obtained from the analysis of protein phosphorylation using the software I developed. First the procedures to collect the datasets of phosphorylation sites are described. Then the dataset collected and studied is presented; finally the analysis of the datasets and the results obtained are discussed.

4.1 The Datasets

A dataset is a collection of known phosphorylation sites for which a pdb file is available. As discussed in section 1.7, working with structural information entails problems in availability and homogeneity of data.

For example the well-established phosphorylation sites database Phospho.ELM [24] does not contain links to structural information; thus external resources, such as Seq2Struct [39] or manual searches on pdb.org have to be used to find possible structural information. However still a limited number of protein structures are available.

Besides, even when the pdb file containing the protein phosphorylated is found, often the chain or the segment of the chain containing the phosphorylated residue is missing and so the pdb file, and consequently the phosphorylation site, has to be rejected.

To manually construct a dataset is therefore necessary to search for annotated phosphorylation site of the kinase of interest in a web database, for example Phospho.ELM [24], PhosphoSitePlus [41], Phospho3D [38], and verify if the relative pdb file is available on pdb.org and if the site is present on that file.

To construct the datasets used for testing the developed software three web databases have been used: Phospho.ELM [24], PhosphoSitePlus [41] and Phospho 3D [38]. PhosphoSitePlus [41] has been widely used because for each annotated phosphorylation site it provides useful links to available pdb files containing it.

Two datasets has been realized, they will be briefly described in the next two sections.

4.2 Kinase Specific Dataset (KS Dataset)

The Kinase Specific Dataset (KS Dataset) has been manually curated using the methods described in the previous section. The particularity of this dataset is that each phosphorylation site is catalogued with respect of its specific protein kinase and not with respect of the family of the protein kinase.

The dataset, consisting of 45 phosphorylation sites of 5 different protein kinases, is displayed in Table 4.1. All of kinases reported, except kinase SRC, are involved in insulin pathways.

Kinase (UniprotID)	Substrate (Uniprot ID)	Psite's information	Pdb file
Akt1 (P31749)	1433Z_HUMAN (P63104)	SER184	1B1
	ADRB2_HUMAN (P07550)	SER346	2R4R
	ANDR_HUMAN (P10275)	SER791	2AMA
	ARFP2_HUMAN (P53365)	SER260	1I4D
	CASP9_HUMAN (P55211)	SER196	1NW9

	CUGB1_HUMAN (Q92879)	SER28	2DHS
	ESR1_HUMAN (P03372)	SER305	1A52
	FOXO3_HUMAN (O43524)	SER253	2UZK
	HMOX1_HUMAN (P09601)	SER188	1N45
	STK4_HUMAN (Q13043)	THR120	3COM
	NCF1_HUMAN (P14598)	SER304	1NG2
	NCF1_HUMAN (P14598)	SER328	1NG2
	PEA15_CRIGR (Q9Z297)	SER116	1N3K
	F263_HUMAN (Q16875)	SER461	2DWO
	PTN1_HUMAN (P18031)	SER50	3CWE
	RAC1_HUMAN (P15154)	SER71	1HH4
	RARA_HUMAN (P10276)	SER96	1DSZ
ERK1 (P27361)	BIEA_HUMAN (P53004)	SER230	2H63
	CDK2_HUMAN (P24941)	THR160	1B38
	GSK3B_HUMAN (P49841)	THR43	1I09
	SYK_HUMAN (Q15046)	SER207	3BJU
	RAB5A_HUMAN (P20339)	SER123	1N6I
	RXRA_HUMAN (P19793)	SER260	1K74
	THB1_HUMAN (P10828)	SER142	2NLL
MEK1 (Q02750)	MK03_HUMAN (P27361)	THR202	2ZOQ
	MK03_HUMAN (P27361)	TYR204	2ZOQ
	MK01_HUMAN (P28482)	THR185	1PME
	MK01_HUMAN (P28482)	TYR187	1PME
	GSK3B_HUMAN (P49841)	TYR216	1I09
PDK1 (Q15118)	AKT1_HUMAN (P31749)	THR308	3CQW
	AKT2_HUMAN (P31751)	THR309	3D0E
	ODPA_HUMAN (P08559)	SER232	1NI4
	ODPA_HUMAN (P08559)	SER293	1NI4
	ODPA_HUMAN (P08559)	SER300	1NI4
	SGK1_HUMAN (O00141)	THR256	2R5T
SRC (P12931)	ESR1_HUMAN (P03372)	TYR537	1GWQ
	ANX2_HUMAN (P07355)	TYR23	1W7B
	SRC_HUMAN (P12931)	TYR215	1A07
	SRC_HUMAN (P12931)	TYR418	1KSW
	SYUA_HUMAN (P37840)	TYR125	1XQ8
	IGF1R_HUMAN (P08069)	TYR1161	1M7N
	IGF1R_HUMAN (P08069)	TYR1165	1M7N
	IGF1R_HUMAN (P08069)	TYR1166	1M7N
	KIT_HUMAN (P10721)	TYR900	1T46
	CXA1_RAT (P08050)	TYR264	1R5S

Table 4.1 – Kinase Specific Dataset

4.3 Kinase Family Specific Dataset (KFS Dataset)

The Kinase Family Specific dataset is a dataset curated from the annotation of the work of Durek et al. [33] available as external file of the published article. These phos-

phorylation files have been collected by Durek and coworkers according to their kinase family; in this thesis only sites whose kinase family was not defined as 'Unknown' have been used.

The dataset, consisting of 149 phosphorylation sites of 14 kinases families, is displayed in Table 4.2. Among the kinase families only the INSR protein kinases are involved in insulin pathways.

Kinase Family	Substrate (Uniprot ID)	Psite's information	Pdb file
AURORA	STK6_HUMAN (O14965)	SER266	1OL5
	STK6_HUMAN (O14965)	SER342	1OL5
	H32_BOVIN (P16105)	SER10	1KX5
	RGAP1_HUMAN (Q9H0H5)	SER410	2OVJ
	STK6_HUMAN (O14965)	SER226	1OL5
	P53_HUMAN (P04637)	SER215	1GZH
	RGAP1_HUMAN (Q9H0H5)	SER387	2OVJ
	STK6_HUMAN (O14965)	THR287	1OL5
	IAP4_HUMAN (O15392)	THR117	1E31
CAMK	IMA2_MOUSE (P52293)	SER105	1EJL
	KPBG_RABIT (P00518)	SER81	1QL6
	EGFR_HUMAN (P00533)	SER768	1M14
	PHS2_RABIT (P00489)	SER14	1C8L
	KPBG_RABIT (P00518)	SER30	1QL6
	P17676 (P17676)	SER325	1GTW
	DLG4_RAT (P31016)	SER73	1IU0
	SPRE_HUMAN (P35270)	SER213	1Z6Z
CDK	FEN1_HUMAN (P39748)	SER187	1UL1
	UBE2B_HUMAN (P63146)	SER120	1JAS
	RAB5B_HUMAN (P61020)	SER123	2HEI
	ITPR1_MOUSE (P11881)	SER421	1N4K
	STXB1_RAT (P61765)	SER158	3C98
	DDX3X_HUMAN (O00571)	THR322	2I4I
	EZRI_HUMAN (P15311)	THR234	1NI2
	STXB1_RAT (P61765)	THR574	3C98
	IAP4_HUMAN (O15392)	THR34	1E31
DDX3X_HUMAN (O00571)	THR203	2I4I	
CK2	CALM_HUMAN (P62158)	SER101	1CDL
	IF5_HUMAN (P55010)	SER390	2IU1
	IRS1_HUMAN (P35568)	SER99	1QQG
	EGR1_MOUSE (P08046)	SER351	1A1F
	IF4E_YEAST (P07260)	SER15	1RF8
	MEF2A_HUMAN (Q02078)	SER59	1C7U
	CALM_HUMAN (P62158)	SER81	1CDL
	G6PI_HUMAN (P06744)	SER185	1IAT

	CDN1B_HUMAN (P46527)	SER83	1JSU
	SAT1_HUMAN (P21673)	SER149	2B3U
	NCF1_HUMAN (P14598)	SER208	1UEC
	NCF1_HUMAN (P14598)	SER283	1UEC
	WASP_HUMAN (P42768)	SER484	1EJ5
	RAD_HUMAN (P55042)	SER214	2DPX
	APEX1_HUMAN (P27695)	SER289	1DE8
	SRPK1_HUMAN (Q96SB4)	SER555	1WAK
	CALM_HUMAN (P62158)	THR79	1CDL
	SAT1_HUMAN (P21673)	THR10	2B3U
	BID_MOUSE (P70444)	THR58	1DDB
	CTNB1_HUMAN (P35222)	THR393	1QZ7
CSK	CATA_HUMAN (P04040)	TYR231	1DGB
	GPX1_HUMAN (P07203)	TYR96	2F8A
	RA52_HUMAN (P43351)	TYR104	1H2I
	RAD51_HUMAN (Q06609)	TYR315	1N0W
	SRC_HUMAN (P12931)	TYR529	1Y57
	RAD51_HUMAN (Q06609)	TYR54	1B22
	CD45_HUMAN (P08575)	TYR1216	1YGR
DAPK	STX1A_RAT (P32851)	SER188	3C98
	DAPK3_HUMAN (O43293)	THR265	1YRP
	DAPK3_HUMAN (O43293)	THR180	1YRP
	DAPK3_HUMAN (O43293)	THR225	1YRP
FGFR	FGFR1_HUMAN (P11362)	TYR653	1AGW
	FGFR1_HUMAN (P11362)	TYR605	1AGW
	FGFR1_HUMAN (P11362)	TYR280	1EVT
	RET_HUMAN (P07949)	TYR809	2IVS
	FGFR2_HUMAN (P21802)	TYR466	3B2T
	FGFR1_HUMAN (P11362)	TYR730	1AGW
	HNRPQ_HUMAN (O60506)	TYR373	2DGU
GRK	RK_BOVIN (P28327)	SER21	3C4W
	TBB_PIG (P02554)	SER430	1FFX
	RK_BOVIN (P28327)	SER488	3C4W
	RK_BOVIN (P28327)	THR489	3C4W
	TBB_PIG (P02554)	THR419	1FFX
INSR	PTN1_HUMAN (P18031)	TYR153	1BZC
	IGF1R_HUMAN (P08069)	TYR1280	2OJ9
	BIEA_HUMAN (P53004)	TYR228	2H63
	ADRB2_HUMAN (P07550)	TYR141	2RH1
	P85A_HUMAN (P27986)	TYR580	2V1Y
	FGFR1_HUMAN (P11362)	TYR154	1EVT
	GTF2I_HUMAN (P78347)	TYR398	2DN4
	P85A_RAT (Q63787)	TYR368	1FU5
PAK	GDIR_HUMAN (P52565)	SER101	1CC0
	GDIR_HUMAN (P52565)	SER174	1CC0
	PAK7_HUMAN (Q9P286)	SER573	2F57
	MYC_HUMAN (P01106)	SER373	1NKP
	PGAM1_HUMAN (P18669)	SER117	1YFK
	PGMU_RABBIT (P00949)	THR467	1C47
	PP14A_PIG (O18734)	THR38	1J2M
PKA	CGHB_HUMAN (P01233)	SER116	1HCN

	Q4VXV4_HUMAN (Q4VXV4)	SER221	2JP9
	ESR1_HUMAN (P03372)	SER236	1HCQ
	LYSC_CHICK (P00698)	SER42	1B0D
	PLM_HUMAN (O00168)	SER83	2JO1
	NR4A1_RAT (P22829)	SER344	1CIT
	CAN2_HUMAN (P17655)	SER369	1KFU
	PIN1_HUMAN (Q13526)	SER16	1F8A
	HDAC8_HUMAN (Q9BY41)	SER39	1T64
	GLYG_RABIT (P13280)	SER44	1LLO
	REL_CHICK (P16236)	SER266	1GJI
	TF65_MOUSE (Q04207)	SER276	1K3Z
	GRIK2_RAT (P42260)	SER715	1S50
	GRIK2_RAT (P42260)	SER697	1S50
	P17676 (P17676)	SER288	1GTW
	LYSC_CHICK (P00698)	SER68	1B0D
	CGHB_HUMAN (P01233)	SER86	1HCN
	NOS1_RAT (P29476)	SER374	1K2R
	Q4VXV4_HUMAN (Q4VXV4)	SER249	2JP9
	MEPD_HUMAN (P52888)	SER642	1S4B
	NFKB1_MOUSE (P25799)	SER335	1IKN
	RARA_HUMAN (P10276)	SER369	1DKF
	CGHB_HUMAN (P01233)	THR117	1HCN
	MEF2A_HUMAN (Q02078)	THR20	1C7U
	GB13_MOUSE (P27601)	THR203	1ZCB
	CAN2_RAT (Q07009)	THR370	1DF0
	FGF2_HUMAN (P09038)	THR121	1CVS
PKB	BAXA_HUMAN (Q07812)	SER184	1F16
	CDC42_HUMAN (P60953)	SER71	1A4R
	BIRC4_HUMAN (P98170)	SER87	2POI
	ANDR_HUMAN (P10275)	SER791	1E3G
PKC	VINC_HUMAN (P18206)	SER1033	1RKE
	DPOB_HUMAN (P06746)	SER43	1BPX
	PIPNA_RAT (P16446)	SER165	1T27
	PIPNA_MOUSE (P53810)	SER165	1KCM
	1433B_HUMAN (P31946)	SER131	2BQ0
	IREB1_HUMAN (P21399)	SER711	2B3X
	IREB1_HUMAN (P21399)	SER138	2B3X
	PEBP_HUMAN (P30086)	SER152	1BD9
	GRIA2_RAT (P19491)	SER683	1FTJ
	TNNI3_HUMAN (P19429)	SER41	1J1D
	RGS7_HUMAN (P49802)	SER434	2A72
	MK10_HUMAN (P53779)	SER167	1PMN
	NCF4_HUMAN (Q15080)	SER315	1OEY
	KIT_HUMAN (P10721)	SER817	1PKG
	CXA1_RAT (P08050)	SER367	1R5S
	IF4E_MOUSE (P63073)	SER209	1EJ1
	NHERF_HUMAN (O14745)	SER162	2OZF
	PIPNB_RAT (P53812)	SER261	2A1L
	INSR_HUMAN (P06213)	SER1064	1GAG
	TF65_MOUSE (Q04207)	SER311	1K3Z
	FSCN1_HUMAN (Q16658)	SER39	1DFC

	MYOD_MOUSE (P10085)	THR115	1MDY
	LAMA_HUMAN (P02545)	THR480	1IFR
	LMNA_MOUSE (P48678)	THR480	1UFG
	1433B_HUMAN (P31946)	THR142	2BQ0
	EGFR_HUMAN (P00533)	THR416	1UFG
	EF1A_YEAST (P02994)	THR430	1F60
SYK	TBA_PIG (P02550)	TYR432	1FFX
	ZAP70_HUMAN (P43403)	TYR126	2OZO
	MK14_HUMAN (Q16539)	TYR323	1BL6
	KSYK_HUMAN (P43405)	TYR525	1XBA
	DUS3_HUMAN (P51452)	TYR138	1J4X

Table 4.2 – Kinase Family Specific Dataset

The only, but important and interesting, difference with the KS dataset is that these phosphorylation sites are annotated with respect to the family of the kinase irrespective of the particular kinase.

4.4 Importance of considering kinase specific phosphorylation sites

As pointed out in section 2.1, the fact that, in most of the cases, only one kinase of a certain family phosphorylated a particular phosphorylation site, suggests that a characterization of the spatial environment of substrates of a single kinases rather than of families of kinases could be more accurate and useful.

To confirm that, here I compare the results of the *Distances* routine, described in section 3.3.6, that display the Average Euclidean distance between sites of each single kinase (or families), i.e. the self similarity of the kinase, and the Cross Average Euclidean distance, i.e. the cross similarity, of each pair of sites of kinases (or families), separately on the KS dataset and the KSF dataset.

In Figure 3.6 the result of the *Distances* routine on the Kinase Specific dataset without cut off option is displayed. As can be viewed, excluding kinase ERK1, the self similarity (left table) range from 1.59 and 1.73 while the cross similarity (right table) between kinases are all over the 1.76 value. In other words the spatial environment characterized as in PP3D seems to captures a structural similarity of sites phosphorylated by the same kinase.

To take in account also kinase ERK1 the cut off option with threshold 3 would have to be used to exclude outliers sites; the results of *Distances* routine with the cutoff option are displayed in Figure 4.1.

Self Similarity					Cross Similarity				
kinase	Eucl Dist	Generic ED	Flank Dist	Generic FD	kinases	Eucl Dist	Generic ED	Flank Dist	Generic FD
MEK1 (Q02750)	1.59	4.47	0.27	0.55	Akt1 (P31749) - ERK1 (P27361)	1.68	6.02	0.43	1.31
SRC (P12931)	1.65	5.37	0.43	1.33	SRC (P12931) - Akt1 (P31749)	1.76	6.09	0.43	1.38
PDK1 (Q15118)	1.65	5.59	0.31	0.87	SRC (P12931) - ERK1 (P27361)	1.82	5.99	0.45	1.36
ERK1 (P27361)	1.72	6.3	0.45	1.29	SRC (P12931) - PDK1 (Q15118)	1.86	6.24	0.39	1.14
Akt1 (P31749)	1.73	6.36	0.39	1.18	MEK1 (Q02750) - PDK1 (Q15118)	1.92	6.34	0.32	0.85
					Akt1 (P31749) - PDK1 (Q15118)	1.97	7.08	0.41	1.28
					ERK1 (P27361) - PDK1 (Q15118)	2.03	7.11	0.41	1.13
					MEK1 (Q02750) - SRC (P12931)	2.13	6.29	0.4	1.18
					MEK1 (Q02750) - Akt1 (P31749)	2.2	7.44	0.36	1.07
					MEK1 (Q02750) - ERK1 (P27361)	2.27	7.42	0.39	1.13

Figure 4.1 – Distances KS Dataset CutOff 3

Now also self similarity of ERK1 is under 1.73 and considering the cross similarities, only that between AKT1 and ERK1 is under this threshold value; thus this results confirm the goodness of the spatial environment characterization of PP3D.

Repeating the same experiments using the Kinase Family Specific Dataset, where sites are collected with respect to the kinases family, it can be confirmed that the specific kinase characterization of sites is more accurate.

Figure 4.2 and Figure 4.3 display the results of the *Distance* routine on the KFS dataset; the self similarities (left table) range from 1.35 and 3.53 while cross similarities (right table) from 1.57 (Figure 4.2) to 3.47 (Figure 4.3) showing that results are not discernable.

Self Similarity					Cross Similarity				
kinase	Eucl Dist	Generic ED	Flank Dist	Generic FD	kinases	Eucl Dist	Generic ...	Flank Dist	Generic ...
GRK (FAMILY)	1.35	4.79	0.46	1.68	CAMK (FAMILY) - GRK (FAMILY)	1.57	5.4	0.48	1.57
CAMK (FAMILY)	1.53	5.3	0.42	1.19	CAMK (FAMILY) - FGFR (FAMILY)	1.59	5.34	0.41	1.11
FGFR (FAMILY)	1.57	5.24	0.43	1.08	CSK (FAMILY) - FGFR (FAMILY)	1.61	5.77	0.38	1.04
CSK (FAMILY)	1.74	6.3	0.38	1.04	CSK (FAMILY) - GRK (FAMILY)	1.65	6.39	0.43	1.37
SYK (FAMILY)	1.8	6.05	0.39	1.05	GRK (FAMILY) - FGFR (FAMILY)	1.65	6.29	0.43	1.41
AURORA (FAMILY)	1.85	6.36	0.45	1.49	FGFR (FAMILY) - SYK (FAMILY)	1.65	5.66	0.39	1.03
PKB (FAMILY)	1.96	7.13	0.42	1.3	GRK (FAMILY) - AURORA (FAMILY)	1.67	5.85	0.49	1.8
CDK (FAMILY)	2.25	8.16	0.45	1.29	CAMK (FAMILY) - CSK (FAMILY)	1.71	5.94	0.41	1.13
CK2 (FAMILY)	2.41	9.01	0.51	1.6	CAMK (FAMILY) - AURORA (FAMI...	1.74	6.04	0.43	1.31
PAK (FAMILY)	2.45	7.47	0.43	1.3	PKB (FAMILY) - FGFR (FAMILY)	1.81	6.32	0.42	1.15
PKA (FAMILY)	2.82	9.9	0.56	1.76	AURORA (FAMILY) - FGFR (FAMI...	1.81	6.35	0.42	1.26
PKC (FAMILY)	2.95	9.99	0.56	1.84	CAMK (FAMILY) - PKB (FAMILY)	1.82	6.62	0.39	1.12
DAPK (FAMILY)	2.97	7.58	0.44	1.12	PKB (FAMILY) - GRK (FAMILY)	1.82	6.96	0.47	1.66
INSR (FAMILY)	3.53	12.31	0.51	1.57	CDK (FAMILY) - FGFR (FAMILY)	1.86	6.4	0.42	1.14
					GRK (FAMILY) - SYK (FAMILY)	1.87	7.14	0.41	1.37
					GRK (FAMILY) - CK2 (FAMILY)	1.88	6.96	0.5	1.69
					CAMK (FAMILY) - SYK (FAMILY)	1.89	6.42	0.41	1.13
					PKB (FAMILY) - SYK (FAMILY)	1.91	6.56	0.41	1.18
					PAK (FAMILY) - FGFR (FAMILY)	1.91	6.0	0.4	1.13
					CSK (FAMILY) - SYK (FAMILY)	1.92	6.97	0.41	1.18
					CSK (FAMILY) - AURORA (FAMILY)	1.93	6.93	0.43	1.28
					CSK (FAMILY) - PKB (FAMILY)	1.98	7.39	0.41	1.19
					CAMK (FAMILY) - CDK (FAMILY)	2.01	7.17	0.44	1.24
					CAMK (FAMILY) - CK2 (FAMILY)	2.01	7.34	0.48	1.49
					CAMK (FAMILY) - PAK (FAMILY)	2.02	6.38	0.41	1.19
					CDK (FAMILY) - SYK (FAMILY)	2.06	6.88	0.42	1.17

Figure 4.2 – Distances KFS Dataset ½

Results

Self Similarity					Cross Similarity				
kinase	Eucl Dist	Generic ED	Flank Dist	Generic FD	kinases	Eucl Dist	Generic ...	Flank Dist	Generic ...
GRK (FAMILY)	1.35	4.79	0.46	1.68	PKC (FAMILY) - PAK (FAMILY)	2.67	8.72	0.48	1.5
CAMK (FAMILY)	1.53	5.3	0.42	1.19	PKA (FAMILY) - CK2 (FAMILY)	2.68	9.73	0.59	1.91
FGFR (FAMILY)	1.57	5.24	0.43	1.08	CAMK (FAMILY) - DAPK (FAMILY)	2.69	8.06	0.45	1.16
CSK (FAMILY)	1.74	6.3	0.38	1.04	INSR (FAMILY) - PKB (FAMILY)	2.69	9.74	0.49	1.45
SYK (FAMILY)	1.8	6.05	0.39	1.05	DAPK (FAMILY) - GRK (FAMILY)	2.72	9.33	0.46	1.46
AURORA (FAMILY)	1.85	6.36	0.45	1.49	PKC (FAMILY) - CDK (FAMILY)	2.73	9.66	0.52	1.64
PKB (FAMILY)	1.96	7.13	0.42	1.3	CSK (FAMILY) - DAPK (FAMILY)	2.74	8.39	0.42	1.03
CDK (FAMILY)	2.25	8.16	0.45	1.29	CDK (FAMILY) - DAPK (FAMILY)	2.74	8.76	0.44	1.16
CK2 (FAMILY)	2.41	9.01	0.51	1.6	PKC (FAMILY) - CK2 (FAMILY)	2.76	9.69	0.58	1.9
PAK (FAMILY)	2.45	7.47	0.43	1.3	PKA (FAMILY) - PAK (FAMILY)	2.76	9.23	0.5	1.53
PKA (FAMILY)	2.82	9.9	0.56	1.76	INSR (FAMILY) - AURORA (FAMILY)	2.81	10.21	0.49	1.5
PKC (FAMILY)	2.95	9.99	0.56	1.84	DAPK (FAMILY) - SYK (FAMILY)	2.81	8.59	0.42	1.14
DAPK (FAMILY)	2.97	7.58	0.44	1.12	CDK (FAMILY) - PKA (FAMILY)	2.84	10.4	0.56	1.7
INSR (FAMILY)	3.53	12.31	0.51	1.57	DAPK (FAMILY) - PAK (FAMILY)	2.86	8.35	0.46	1.28
					PKC (FAMILY) - PKA (FAMILY)	2.87	9.84	0.57	1.82
					INSR (FAMILY) - PAK (FAMILY)	2.88	9.39	0.46	1.47
					PKB (FAMILY) - DAPK (FAMILY)	2.88	9.64	0.45	1.34
					INSR (FAMILY) - CDK (FAMILY)	2.93	10.52	0.49	1.43
					DAPK (FAMILY) - AURORA (FAMI...	3.04	9.87	0.47	1.37
					INSR (FAMILY) - CK2 (FAMILY)	3.11	11.45	0.54	1.74
					DAPK (FAMILY) - CK2 (FAMILY)	3.14	10.21	0.49	1.37
					PKC (FAMILY) - INSR (FAMILY)	3.24	11.1	0.56	1.78
					INSR (FAMILY) - PKA (FAMILY)	3.29	11.68	0.56	1.76
					PKC (FAMILY) - DAPK (FAMILY)	3.33	10.32	0.52	1.61
					INSR (FAMILY) - DAPK (FAMILY)	3.34	10.29	0.49	1.35
					DAPK (FAMILY) - PKA (FAMILY)	3.47	11.16	0.57	1.66

Figure 4.3 – Distances KFS Dataset 2/2

Using a cutoff value of 2.5 (Figure 4.4) we obtain a good normalization of self distances, and the biggest value is now 1.98. However, even in this case, cross similarities do not present values greater than self similarities, thus preventing a clear characterization.

Self Similarity					Cross Similarity				
kinase	Eucl Dist	Generic ED	Flank Dist	Generic FD	kinases	Eucl Dist	Generic ...	Flank Dist	Generic ...
PKA (FAMILY)	1.09	3.82	0.47	0.97	GRK (FAMILY) - PKA (FAMILY)	1.34	4.62	0.49	1.72
GRK (FAMILY)	1.35	4.79	0.46	1.68	PKA (FAMILY) - AURORA (FAMILY)	1.36	4.62	0.43	1.3
AURORA (FAMILY)	1.45	4.76	0.45	1.48	PKC (FAMILY) - PKA (FAMILY)	1.37	4.79	0.42	1.28
PKC (FAMILY)	1.52	5.18	0.48	1.46	PKC (FAMILY) - GRK (FAMILY)	1.4	4.69	0.5	1.78
CAMK (FAMILY)	1.53	5.3	0.42	1.19	GRK (FAMILY) - AURORA (FAMILY)	1.44	5.02	0.49	1.84
FGFR (FAMILY)	1.57	5.24	0.43	1.08	PKA (FAMILY) - CK2 (FAMILY)	1.48	5.26	0.48	1.45
DAPK (FAMILY)	1.58	6.35	0.31	0.85	PKC (FAMILY) - AURORA (FAMILY)	1.5	5.09	0.45	1.44
CK2 (FAMILY)	1.71	6.06	0.52	1.42	GRK (FAMILY) - CK2 (FAMILY)	1.5	5.32	0.48	1.6
CSK (FAMILY)	1.74	6.3	0.38	1.04	CAMK (FAMILY) - AURORA (FAMI...	1.56	5.45	0.43	1.33
PAK (FAMILY)	1.77	4.84	0.43	1.32	CAMK (FAMILY) - GRK (FAMILY)	1.57	5.4	0.48	1.57
SYK (FAMILY)	1.8	6.05	0.39	1.05	AURORA (FAMILY) - CK2 (FAMILY)	1.58	5.55	0.48	1.59
CDK (FAMILY)	1.81	6.06	0.47	1.28	CAMK (FAMILY) - FGFR (FAMILY)	1.59	5.34	0.41	1.11
PKB (FAMILY)	1.96	7.13	0.42	1.3	INSR (FAMILY) - GRK (FAMILY)	1.59	5.55	0.46	1.73
INSR (FAMILY)	1.98	6.59	0.37	1.23	PKC (FAMILY) - CAMK (FAMILY)	1.6	5.63	0.44	1.39
					CAMK (FAMILY) - PKA (FAMILY)	1.6	6.12	0.4	1.17
					CSK (FAMILY) - FGFR (FAMILY)	1.61	5.77	0.38	1.04
					CDK (FAMILY) - FGFR (FAMILY)	1.63	5.43	0.43	1.18
					PKC (FAMILY) - CK2 (FAMILY)	1.64	5.66	0.51	1.66
					INSR (FAMILY) - FGFR (FAMILY)	1.65	5.57	0.39	1.12
					CSK (FAMILY) - GRK (FAMILY)	1.65	6.39	0.43	1.37
					CDK (FAMILY) - GRK (FAMILY)	1.65	6.2	0.46	1.59
					GRK (FAMILY) - PAK (FAMILY)	1.65	5.53	0.46	1.64
					GRK (FAMILY) - FGFR (FAMILY)	1.65	6.29	0.43	1.41
					FGFR (FAMILY) - SYK (FAMILY)	1.65	5.66	0.39	1.03
					CAMK (FAMILY) - INSR (FAMILY)	1.66	5.47	0.41	1.23
					CAMK (FAMILY) - CDK (FAMILY)	1.66	5.59	0.46	1.21

Figure 4.4 - Distances KFS Dataset CutOff 2.5

It is necessary to stress at this point that the size of the Kinase Specific dataset is too small for an accurate analysis. However, these data suggest that the not excellent results of structural approaches in prediction methods are probably due to a characterization of phosphorylation sites with respect to their kinase family rather than with the specific protein kinase.

4.5 Comparative analysis of the distance metrics used

In the previous subsection, the comparison among phosphorylation sites have been carried out using the Average Euclidean distance. Here the results for the other three distance measures (the Euclidean distance of the flanking sequence and the same measures without dividing the spatial context in different radii) are discussed.

4.5.1 Not relevance of Flanking Distance

The Flanking distance between two or more sites, is the average Euclidean distance of their histograms considering only the flanking sequences of the phosphorylation sites (section 2.2.4).

The relevance of the spatial propensity of the flanking sequence to be in the proximity of phosphorylation sites is studied by comparing the classification of phosphorylation sites based on the flanking sequence and the average Euclidean distance computed on the amino acids irrespective of their sequence position (see Figure 4.1).

The value of the flanking distance between sites of the same kinases range from 0.27 to 0.45, while between sites of different kinases it ranges from 0.32 up to 0.45,

showing no evident difference in the distribution. Considering the KSF dataset, with a cutoff of 2.5 (see Figure 4.4), leads us to the same conclusion.

Therefore the conclusion is that phosphorylation sites phosphorylated by the same kinase or kinase family do not assume a particular spatial propensity around the site of their flanking sequence. On the other hand, it has been revealed a particular spatial propensity of their amino acids irrespective of their sequence position.

4.5.2 Importance of using different radii in phosphorylation site characterization

To best characterize the spatial environment, this has been divided into concentric regions with a different radius. In the experiments shown in the previous section the last (most external) radius was set to 16 Å. Here we discuss a comparison between the measures obtained with this characterization and the ones produced by a single region, namely the Generic Euclidean distance.

For the KS dataset, looking at Figure 4.1, the values of the Generic Euclidean distance between sites of the same kinases range from 4.47 to 6.36, while values from different kinases from 5.99 to 7.44. Moreover, six out of ten values are below the threshold value of 6.36 showing no clear separation between the two classes.

For the KSF dataset, similar results are shown in Figure 4.4, showing an imprecise phosphorylation sites characterization using a single radius instead of more radius.

The choice to set the (last) radius at 16 Å is justified by the results reported in Figure 4.5 and Figure 4.6. These two figures show the measures of the self and cross similarities in the KS dataset considering, respectively, a spatial context of 10 Å and 22 Å.

It is evident that under these settings the self and cross similarities do not show a more discriminative behavior, confirming that considering a spatial context of 16 Å gives the best characterization of the residues propensity around the phosphorylation sites.

Distances - CutOff: No cutoff - Last radius: 10 A									
Self Similarity					Cross Similarity				
kinase	Eucl Dist	Generic ED	Flank Dist	Generic FD	kinases	Eucl Dist	Generic ED	Flank Dist	Generic FD
PDK1 (Q15118)	0.46	0.92	0.26	0.52	MEK1 (Q02750) - PDK1 (Q15118)	0.53	1.07	0.25	0.5
MEK1 (Q02750)	0.51	1.02	0.21	0.43	MEK1 (Q02750) - SRC (P12931)	0.63	1.25	0.35	0.7
SRC (P12931)	0.7	1.37	0.45	0.88	MEK1 (Q02750) - Akt1 (P31749)	0.63	1.26	0.32	0.65
Akt1 (P31749)	0.7	1.41	0.4	0.8	SRC (P12931) - PDK1 (Q15118)	0.65	1.29	0.37	0.73
ERK1 (P27361)	1.15	2.39	0.41	0.81	Akt1 (P31749) - PDK1 (Q15118)	0.65	1.31	0.39	0.79
					SRC (P12931) - Akt1 (P31749)	0.71	1.41	0.45	0.89
					MEK1 (Q02750) - ERK1 (P27361)	0.86	1.79	0.33	0.66
					ERK1 (P27361) - PDK1 (Q15118)	0.89	1.86	0.38	0.75
					SRC (P12931) - ERK1 (P27361)	0.95	1.93	0.46	0.88
					Akt1 (P31749) - ERK1 (P27361)	0.97	1.99	0.42	0.83

Figure 4.5 – Distances KS Last Radius 10Å

Distances - CutOff: No cutoff - Last radius: 22 A									
Self Similarity					Cross Similarity				
kinase	Eucl Dist	Generic ED	Flank Dist	Generic FD	kinases	Eucl Dist	Generic ED	Flank Dist	Generic FD
Akt1 (P31749)	2.59	15.93	0.3	1.21	SRC (P12931) - Akt1 (P31749)	3.28	19.18	0.34	1.4
MEK1 (Q02750)	3.31	15.35	0.23	0.58	Akt1 (P31749) - ERK1 (P27361)	3.33	20.49	0.35	1.41
SRC (P12931)	3.47	19.51	0.34	1.33	SRC (P12931) - ERK1 (P27361)	3.67	20.72	0.37	1.5
PDK1 (Q15118)	3.49	18.24	0.26	0.79	ERK1 (P27361) - PDK1 (Q15118)	3.67	20.6	0.35	1.35
ERK1 (P27361)	3.72	22.72	0.38	1.52	Akt1 (P31749) - PDK1 (Q15118)	3.68	21.3	0.33	1.31
					SRC (P12931) - PDK1 (Q15118)	3.78	20.34	0.32	1.09
					MEK1 (Q02750) - PDK1 (Q15118)	4.25	23.88	0.27	0.83
					MEK1 (Q02750) - ERK1 (P27361)	4.46	24.25	0.33	1.32
					MEK1 (Q02750) - SRC (P12931)	4.75	26.25	0.32	1.22
					MEK1 (Q02750) - Akt1 (P31749)	4.88	29.95	0.29	1.14

Figure 4.6 – Distances KS Last Radius 22 Å

5 Conclusion

In this chapter a short “executive-summary” of the activities carried out during this thesis is provided.

Then the remaining open question and the future developments are discussed.

5.1 Work performed

As confirmed by two decades of research activities, protein phosphorylation is a central and important research topic not only in molecular biology but also in bioinformatics.

PTMs involve about one third of human proteins with a crucial role especially in signals transduction and in cellular pathways. The huge costs in time and money to study PTMs and the capacity of informatics tools to rapidly analyze large amounts of data makes bioinformatics efforts in this field not only useful but crucial.

The constant growth of available data and the development of several software tools in the recent year suggest that protein phosphorylation will be the subject of study by bioinformatics community for several years.

As widely discussed, most of analysis and predictor tools rely only on sequence information, and only recently structural information has been introduced in this field but always in addition of sequence information.

Starting from the recognized and fundamental importance of structural conformation of proteins for their activity, the aim of this thesis was to use computer vision methods in an attempt to characterize the spatial context of phosphorylation sites using only 3D information.

A tool, PP3D, which describes phosphorylation sites with the propensity of amino acids to appear in fixed radii regions centered in the sites has been developed. PP3D allows to study many aspects of the spatial context of a phosphorylation site as well as to study similarity between phosphorylation sites, and to find common structural features in sites phosphorylated by the same kinase.

In addition of the bioinformatics tool, two databases of phosphorylation sites have been collected and some interesting results have been obtained by analyzing them with PP3D. As expected, describing the spatial context with the propensity of the amino acids to appear in regions corresponding to 3 different radii of 4, 10 and 16 Å centered in the phosphorylation site, sites of the same kinase appear to show greater structural similarity than sites of different kinases, confirming the relevance of the methods developed.

5.2 Future developments

As mentioned before, the manually collected dataset (4.2) is too small to draw general conclusion, thus the results would have to be verified and confirmed with an extended dataset including future available structural data.

Again, the aim of this thesis was not to construct a structural based predictor but to understand the role of structural information in the recognition of such site.

Therefore, considering the infancy status of structural predictors, the work done in this thesis can be viewed as a starting point to help understanding the importance of protein structure in protein phosphorylation.

The results obtained, mainly due to the particular spatial characterization and the distance functions used, suggest that the use of the PP3D's *PredictionAlgorithm* in conjunction to a well established sequence based approach could lead to improved prediction results.

References

- [1] P Cohen, "Dissection of the protein phosphorylation cascades involved in insulin and growth factor action," *Biochemical Society Transaction*, 1993.
- [2] P Cohen, "The regulation of protein function by multisite phosphorylation - a 25 year update," *Trends in Biochemical Sciences*, 2000.
- [3] PV Attwood, MJ Piggot, Zu. XL, and PG Besant, "Focus on phosphohistidine," *Amino Acids*, 2007.
- [4] P Cohen, "The origin of protein phosphorylation," *Nature Cell Biology*, 2002.
- [5] A Kreegipuu, N Blom, and S Brunak, "PhosphoBase, a database of phosphorylation sites: release 2.0," *Nucleic Acids Research*, 1999.
- [6] RA Lindberg, AM Quinn, and T Hunter, "Dual-specificity protein kinases: will any hydroxyl do?," *Trends in Biochemical Sciences*, 1992.
- [7] RM Stroud, "Mechanism of biological control by phosphorylation," *Current Opinion in Structural Biology*, 1991.
- [8] B Kobe, T Kampmann, JK Forwood, P Listwan, and RI Brinkworth, "Substrate specificity of protein kinases and computational prediction of substrates," *Biochimica et Biophysica Acta*, 2005.
- [9] G Manning, D.B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The

- Protein Kinase Complement of the Human Genome," *Science*, 2002.
- [10] SK Hanks and T Hunter, "The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification," *FASEB*, 1995.
- [11] SK Hanks, AM Quinn, and T Hunter, "The protein kinase family: conserved features and deduced phylogeny of the catalytic domains," *Science*, 1988.
- [12] SK Hanks, "Eukaryotic protein kinases," *Current Opinion in Structural Biology*, 1991.
- [13] SK Hanks and AM Quinn, "Protein kinases catalytic domain sequence database: identification of conserved features of primary structure and classification of family members," *Methods Enzymol*, 1991.
- [14] The UniProt Consortium, "The Universal Protein Resource (UniProt) in 2010," *Nucleic Acids Research*, 2010.
- [15] Robert Finn et al., "The Pfam protein families database," *Nucleic Acids Research*, 2010.
- [16] Lilia Iakoucheva et al., "The importance of intrinsic disorder for protein phosphorylation," *Nucleic Acids research*, 2004.
- [17] H.M. Berman et al., "The Protein Data Bank," *Nucleic Acids Research*, 2000.
- [18] JL Jimenez, B. Hegemann, JR Hutchins, JM Peters, and R Durbin, "A systematic comparative and structural analysis of protein phosphorylation

sites based on the mtcPTM database," *Genome Biology*, 2007.

- [19] S.C. Fan and X.G. Zhang, "Characterizing the microenvironment surrounding phosphorylation protein sites," *Genomics, Proteomics & Bioinformatics*, 2005.
- [20] John Obenauer, Lewis Cantley, and Micheal Yaffe, "Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs," *Nucleic Acids Research*, 2003.
- [21] Yu Xue et al., "GPS 2.0, a Tool to Predict Kinase-specific Phosphorylation Sites in Hierarchy," *Molecular & cellular proteomics : MCP*, 2008.
- [22] Z Songyang et al., "Use an oriented peptide library to determine the optimal substrates of protein kinases," *Current Biology*, 1994.
- [23] Z Songyang et al., "SH2 domains recognize specific phosphopeptide sequence," *Cell*, 1993.
- [24] Francesca Diella et al., "Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins," *BMC Bioinformatics*, 2004.
- [25] R.M. Biondi and A.R. Nebreda, "Signaling specificity of Ser/Thr protein kinases through docking-site-mediated interactions," *Biochemistry*, 2003.
- [26] P.M. Holland and J.A. Cooper, "Protein modification: docking sites for kinases," *Current Biology*, 1999.

- [27] L. Salwinski et al., "The database of interacting proteins," *Nucleic Acids Research*, 2004.
- [28] C. Stark et al., "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, 2006.
- [29] A. Zanzoni et al., "MINT: a Molecular INTeraction database," *Nucleic Acids Research*, 2002.
- [30] LJ Jensen et al., "STRING 8: a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Research*, 2008.
- [31] Hsien-Da Huang, Tzong-Yi Lee, Shih-Wei Tzeng, and Jorng-Tzong Horng, "KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites," *Nucleic Acids Research*, 2005.
- [32] Yung-Hao Wong et al., "KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns," *Nucleic Acids Research*, 2007.
- [33] Pawel Durek, Christian Schudoma, Wolfram Weckwerth, Joachim Selbig, and Dirk Walther, "Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins," *BMC Bioinformatics*, 2009.
- [34] Alexandros Kaatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis, "kernlab - An S4 Package for Kernel Methods in R," *Journal of Statistical Software*,

2004.

- [35] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," *Nucleic Acids Research*, 2000.
- [36] Y. Park and V. Helms, "Assembly of transmembrane helices of simple polytopic membrane proteins from sequence conservation patterns," *Proteins*, 2006.
- [37] Blom Nikolaj, Gammeltoft Steen, and Brunak Soren, "Sequence and Structure-based Prediction of Eukaryotic Protein Phosphorylation Sites," *Journal of molecular biology*, 1999.
- [38] Andreas Zanoni, Gabriele Ausiello, Allegra Via, Pier Federico Gherardini, and Manuela Helmer-Citterich, "Phospho3D: a database of three-dimensional structures of protein phosphorylation sites," *Nucleic Acids Research*, 2007.
- [39] Allegra Via, Andreas Zanzoni, and Manuela Helmer-Citterich, "Seq2Struct: a resource for establishing sequence-structure links," *Bioinformatics*, 2004.
- [40] Gabriele Ausiello, Allegra Via, and Manuela Helmer-Citterich, "Query3d: a new method for high-throughput analysis of functional residues in protein structures," *BMC Bioinformatics*, 2005.
- [41] P.V. Hornbeck, I. Chabra, J.M. Kornhauser, E. Skrzypek, and B. Zhang, "Phosphosite: a bioinformatics resource dedicated to physiological protein

phosphorylation," *Proteomics*, 2004.

- [42] R.C.G. Holland et al., "BioJava: an Open-Source Framework for Bioinformatics," *Bioinformatics*, 2008.