MODELING PROTEIN-PEPTIDE COMPLEXES USING ROTAMER LIBRARY APPROACH: APPLICATION TO PREDICTION OF SUBSTRATES FOR MHC AND KINASES

Thesis submitted to
Jawaharlal Nehru University
for the degree of

DOCTOR OF PHILOSOPHY

April 2009

Narendra Kumar

National Institute of Immunology
New Delhi, INDIA

राष्ट्रीय प्रतिरक्षाविज्ञान संस्थान

NATIONAL INSTITUTE OF IMMUNOLOGY

CERTIFICATE

This is to certify that the thesis entitled "***MODELING PROTEIN-PEPTIDE COMPLEXES USING ROTAMER LIBRARY APPROACH: APPLICATION TO PREDICTION OF SUBSTRATES FOR MHC AND KINASES***" submitted by **Narendra Kumar** in partial fulfillment of Ph.D. degree of Jawaharlal Nehru University, comprises the work done by the scholar under my guidance at the National Institute of Immunology. The work is original and has not been submitted in part or in full for any degree or diploma of any university.

Debasisa Mohanty

Dr. Debasisa Mohanty
Thesis Supervisor
National Institute of Immunology
New Delhi

Acknowledgement

This thesis is the result of many years of research work during which I worked with many people who were more than supportive throughout my undertaking of research project. It's a pleasure to convey my sincere gratitude to them all in my acknowlegdement.

First of all, I would like to express my gratitude to my thesis supervisor Dr. Debasisa Mohanty for his supervision, advice, and guidance. When I joined his lab, I knew little about bioinformatics and computational biology. Learning and exploring this wonderful field of modern biology, under his mentorship, has been a truly memorable and enriching experience. His command over the subject, logical way of thinking, and critical analysis have always been a great source of inspiration to me. His enthusiasm and encouragement to apply the thoughts to practice have been very inspiring. I am also thankful to him for allowing me room to work in my own way. I am indebted to him more than he knows.

I gratefully thank Dr R. S. Gokhale and his Chemical Biology Lab for the stimulating discussions during our combined lab meetings. My sincere thanks and regard to all of them.

I would like to take this opportunity of thank Prof. A. Surolia, Director, NII for providing the infrastructure and facilities for conducting this research work. I also thank Mr. A. K. Aggarwal and his staff in the Academic Department for facilitating the necessary paper work. I thank Mr. P. L. Dahra and Mr. Varma for their cooperation in providing me with the extended stay in the hostel.

The financial support of NII and CSIR are greatly acknowledged. I also thank my guide Dr. Debasisa Mohanty, Dr. R. P. Roy and Dr. R. S. Gokhale for accommodating me in their research projects during my thesis writing days.

I am grateful to the computer center staff, Mr Rao, Sunita madam and Naveenji, for their assistance and support whenever I needed. The bioinformatics center at NII is incomplete without you all.

My special thanks go to the past and present colleagues and labmates. Working with you all has been a really wonderful experience. Thank you my seniors in lab, Gitanjali, Zeeshan and Pankaj, for your love and support, I have learned a lot from you all. I have been fortunate to have the company of wonderful juniors who extended their help and support in every condition. To Swadha, Garima and Nikhil, thank you for making the lab a great place to work at and also for proofreading the thesis and providing valuable comments. To Sandeep, Bhushan, Prasad, Bhumika, Neha and Preeti, I enjoyed your company a lot! Thanks. I also thank Jyoti Shehara and Jyoti Sharma for being such nice colleagues. To Charu, Richa and Deepti, you along with all others, made the environment in the lab lively, cheerful and enjoyable. The time spend with you all has been truly memorable.

I would like to specially acknowledge my batchmates for making my stay at NII a truly memorable experience. Manu, the time spent with you as a friend and roommate has been an enriching experience. To Vikrant and Shipra, thank you for standing by me in good and bad times alike. It's been only a few years since I have known you but it seems like a lifetime. The scientific and non-scientific discussions with Manu, Suvendu, Vikrant, Jitesh – the list is endless – have been thought provoking and exhilarating experience. I thank Sharmistha, Soma, Ravikant, Namrata, Satish, Naveen, Vibhu, and all my batchmates for being such a nice company all these years.

To my dearest friends, Vikas and Sudhir, long sessions of discussion with you, during early days of Ph.D., were great help in shaping my ideas and clearing my vision. Thank you.

Where would I be without my family? They deserve special mention for their unconditional love, moral support, and enduring my weekend stays at NII without complaining. Words fail me to express my appreciation to my sisters - Didi, Pinki and Rinki, and to Papa and Dadi. All of you have been my pillars of strength. I also thank Pushpendra Mamaji, Mamiji and their cute little angel Manya for their love and affection.

Last but not the least, I would like to thank everybody else who was important to the successful realization of this thesis, as well as expressing my apology that I could not mention personally one by one.

Narendra Kumar

April 2009

Narendra Kumar

Abbreviations

ABL	Abelson tyrosine kinase
ANN	Artificial neural network
aPK	Atypical protein kinase
AUC	Area under curve
BLAST	Basic local alignment search tool
BT	Betancourt and Thirumalai pair potential matrix
CaMK	Calmodulin dependent kinase
CaMKII	Calcium dependent protein kinase II
CDK	Cyclin dependent kinase
CHK	Checkpoint kinase
CK2	Casein kinase 2
CPU	Central processing unit
CTL	Cytotoxic t-lymphocytes
DAPK	Death associated protein kinase
ePK	Eukaryotic protein kinase
ER	Endoplasmic reticulum
FN	False negative
FP	False positive
FPR	False positive rate
GRK	G-protein coupled protein kinase
GSK	Glycogen synthase kinase
HLA	Human leucocyte antigen
HMM	Hidden markov model
IKK	I-kappa-B kinase
IRK	Insulin receptor kinase
MAP2K	Mitogen activated kinase kinase
MAP3K	Mitogen activated kinase kinase kinase
MAPK	Mitogen activated protein kinase
MD	Molecular dynamics
MHC	Major histocompatibility complex
MJ	Miyazawa and Jernigan pair potential matrix

MM/PBSA	Molecular mechanics / poisson-boltzman, surface area
PAK	p21 activated protein kinase
PDB	Protein data bank
PDK	3'-phosphoinositide dependent protein kinase
PHK	Phosphorylase kinase
PKA	Protein kinase A
PKB	Protein kinase B
PKC	Protein kinase C
PKG	cGMP dependent protein kinase
PLK	Polo like kinase
PP	Pair potential
PSSM	Position specific scoring matrix
PTK	Protein tyrosine kinase
QSAR	Quantitative structure activity relationship
RMSD	Root mean square deviation
ROC	Receiver operating characteristic curve
ROCK	Rho kinase
Sn	Sensitivity
Sp	Specificity
SVM	Support vector machine
TAP	Transporter associated with antigen processing
TCR	T-cell receptor
TN	True negative
TP	True positive
TPR	True positive rate
WHO	World health organization

Dedicated to my late mother

Introduction

Protein-protein interactions are of utmost importance for almost all cellular processes including replication, transcription, translation, signal transduction, immune responses and cell growth. Proteins involved in these processes usually perform their function by binding to target proteins and forming protein-protein complexes. Hence identification of the potential interacting partners of a given protein is crucial for understanding the molecular details of a variety of cellular processes. Experimental approaches for identification of such interaction partners involve yeast two-hybrid systems, c-DNA expression library screening and coimmunoprecipitation experiments. These techniques coupled with truncation and mutagenesis experiments, have been used to define the region of interaction between pairs of proteins. These experimental studies also indicate that many interactions occur over short contiguous stretches within one protein, often less than 15 amino acids in length. For example, recognition of substrate proteins by various kinases during cell signaling events is governed primarily by specific interactions between the kinase and a contiguous peptide stretch containing the phosphorylation site. Modular signal transduction domains like SH2, FHA etc. also recognize short peptide motifs on their interaction partners. A number of cellular processes are also governed by interaction of short length polypeptides with the proteins. Several receptors have peptide fragments as ligands e.g. Major histocompatibility complex (MHC), receptors of nervous system, receptors for endocrine peptide hormones etc. Thus, understanding molecular details of interactions between proteins and short peptide motifs is essential for dissecting underlying mechanism of several major cellular processes. Among the various proteins which interact specifically with short peptide motifs, MHC and kinases represent two major protein families whose substrate specificities have been extensively studied by various experimental approaches.

Major histocompatibility complex (MHC) proteins are a class of proteins of immune system which are present on the surface of antigen presenting cells. Their function is to bind processed peptides and provide a continuous update of cellular and environmental composition for the scrutiny by T-cell receptors (TCR) on the surface of cytotoxic T-lymphocytes (CTL). They bind peptides of cytosolic origin during their maturation in Endoplasmic Reticulum (ER), which are processed by proteasome and transferred by Transporter Associated with Antigen Processing (TAP). MHC class-II molecules are expressed on professional antigen presenting cells. MHC system is characterized by extensive degree of allelic polymorphism. Correct prediction of t-cell epitopes has important implications for modern epitope based vaccines design and also cancer therapy.

The phosphorylation of the Ser/Thr/Tyr residues in various target proteins by their respective protein kinases, also involves interaction between the kinase and a short peptide stretch. This phosphorylation is one of the key mechanisms which is used in signal transduction pathways to alter the functional state of various partner proteins. Protein kinases constitute one of the largest known protein families referred to as eukaryotic protein kinase (ePK). These share a common 3D fold. Since protein kinases mediate such vital cellular functions as apoptosis, growth, division, differentiation etc., they are attractive targets for the in depth analysis of signal transduction pathways. Therefore, identification of substrate proteins for various kinases is crucial for understanding signaling networks in various organisms. Availability of the complete genomes of many organisms has led to the identification of their whole kinome complements. However, deciphering the substrate specificity of these large number of kinases remains a major challenge.

Although few prediction programs employing different strategies are currently available for prediction of substrates for protein kinases and MHC binding peptides, most of them are trained on a set of known substrate peptides. As a result, they can only predict for those families for which such experimental data is available. Discovery of newer MHC alleles and identification of large number of kinases in various genomes require development of novel computational methods which can give reliable clues about their substrate specificities even in absence of extensive experimental data. Structure based substrate prediction methods can in principle address these problems. In this thesis, an attempt has been made to develop a novel structure based substrate prediction method for MHCs and kinases. This prediction method involves a multiscale approach, where at the first level putative high scoring substrate peptides are identified by threading of peptide sequences on the structural templates of kinase-peptide or MHC-peptide complexes and scoring them by residue based statistical pair potentials. High scoring peptides short listed by initial screening are modeled in the peptide binding pocket using rotamer library and detailed all atom molecular mechanics potentials, and their binding affinity is re-ranked using binding free energy values computed by MM/PBSA approach. The prediction accuracy of this approach has been extensively benchmarked using experimentally verified substrate peptide information available in Phospho.ELM and SYFPEITHI database.

Chapter one describes the review of current published literature on the protein kinases, MHC proteins and the various methodologies and computer programs currently available for the prediction of their respective substrate peptides. This chapter also discusses the crystal structures available for various kinases and MHC proteins. In addition, this chapter also gives a brief overview of the literature on novel theoretical methods like rotamer library and statistical pair potential which have been used for modeling of protein structures or protein-peptide complexes.

Chapter two describes the development of the MODPROPEP, a knowledge based program for modeling of protein peptide complexes, especially peptides in complex with the MHCs and protein kinases. The available crystal structures of protein-peptide complexes in PDB are used as templates for modeling peptides of desired sequence in the substrate-binding pocket of MHCs or protein kinases. If no crystal structures are available for a given protein kinase or MHC protein, the program can model its structure in complex with peptide of desired sequence using the crystal structure of the most homologous protein-peptide complex. The substrate peptides are modeled using the same backbone conformation as in the template and the side-chain conformations are obtained by SCWRL program which uses a rotamer library approach. This software also provides appropriate interface for identifying putative MHC-binding peptides in the sequence of an antigenic protein, and phosphorylation sites on the substrate protein of a protein kinase, by identifying and scoring inter-molecular contacts between protein and peptide, using residue based statistical pair potentials. User-friendly interfaces are provided for the detailed analysis, and visualization of structure of modeled protein-peptide complexes and analyzing the contacts made by the modeled peptide ligand in the substrate binding pocket of MHC or protein kinase.

Chapter three describes in detail the results on the prediction of phosphorylation sites in the putative substrate proteins of various protein kinases using MODPROPEP. Benchmarking of prediction accuracy of MODPROPEP on the dataset of known substrates catalogued in phospho.ELM has indicated that our structure based method can predict more than 60% of the experimentally identified substrate for 11 protein kinase families. Comparison with predictions by other available programs indicated that MODPROPEP performs significantly better than other structure based prediction tools like PREDIKIN for most of the protein kinase families, while the performance is similar or better than other widely used sequence based prediction tools such as GPS, PPSP and SCANSITE. These results also demonstrate that residue based statistical pair potential can be successfully used for scoring putative substrate peptides of kinases. Chapter three also reports development of a novel multiscale approach which involves re-ranking of the high scoring peptides shortlisted by pair potential approach using all atom MM/PBSA method. For several kinase families MM/PBSA method was able to further improve the ranks of the known substrate peptides among all possible Ser/Thr containing peptides present in the substrate proteins.

Chapter four reports the results of the analysis of solvent accessible surface areas of phosphorylation sites in the crystal structures of known substrates of protein kinases or their structural homologues. In order to understand the importance of surface accessibility of the phosphorylation site in phosphorylation event, the accessibility values for the phosphorylation sites were compared to the accessibility values of their non-phosphorylated counterparts. The average relative solvent accessible area of phosphorylation site residues was found to be significantly more than their non-phosphorylated counterparts. The difference between phospho and non-phospho residues was statistically significant as judged by Wilcoxon test p-values of 2.20×10^{-16}, 5.07×10^{-6} and 2.34×10^{-8} for serine, threonine and tyrosine containing sites. These results suggest that incorporation of solvent accessibility term along with the current scoring function based on the residue-residue statistical energy can further improve the prediction accuracy.

Chapter five describes the benchmarking of MODPROPEP for prediction of substrates for class I and class II MHC proteins. Analysis of available class I and class II MHC-peptide complexes using MODPROPEP indicate that residue based statistical potential can distinguish the MHC bound peptide with high accuracy from among all possible overlapping peptides present in the corresponding source proteins. Benchmarking of MODPROPEP using the substrate peptide data catalogued in SYFPEITHI database indicate that our structure based method can predict substrate peptides for 16 class I and class II alleles with an accuracy above 60%. This chapter also discusses results of multi-scale modeling approach for prediction of substrates for MHC proteins. It is found that, when high scoring peptides obtained by pair potential are modeled using all atom forecfield and re-ranked as per their binding energy by

MM/PBSA approach, in 8 out of the 16 alleles there is improvement in the rank of the true substrate peptides.

Chapter One

Review of Literature

1.1 PROTEIN - PROTEIN INTERACTION

Proteins perform a wide variety of functions in a diverse range of cellular processes. However most proteins in the cells do not function in isolation, but through interactions with other proteins in a pathway or interaction network. In order to understand the biological pathways responsible for execution of various cellular functions, it is necessary to identify the interacting partners for the proteins involved in these pathways. Although, the whole genome sequencing projects for a number of species have identified and annotated the proteins, their interacting partners are largely unidentified.

Identification of protein interaction network is important for the study of processes such as molecular evolution, cell perturbation and understanding the molecular basis of the diseases (Ideker and Sharan, 2008). Genome mining studies for deciphering protein interaction networks in some disease causing organism have provided novel insights for better understanding the mechanisms of pathogenesis of these organisms. Using the yeast two hybrid system, LaCount *et al.* (2005) identified the protein-protein interaction network of *P. falciparum*. They identified 2846 unique interactions from 32000 yeast two hybrid screens. Their analysis of the protein interaction network in *P. falciparum* has resulted in identification of proteins which are implicated in various important cellular processes such as chromatin modification, transcription, mRNA stability, ubiquitination, and proteins involved in invasion of host cells. Apart from experimental studies, *in silico* approaches have also been used for identification of protein interaction networks. Mawuenyega *et al.* (2005) identified, through a combination of high throughput proteomics and computational approach, globally expressed 1044 proteins and their localization in subcellular compartments. Computational analysis of metabolic pathways was used to integrate proteomics and reconstruct the response networks (Mawuenyega *et al.*, 2005). In a recent study, Yellaboina *et al.* (2007) have used SVM trained on the known interactions in EcoCyc database to predict the protein interaction networks. They observed that the proteins involved in the replication, DNA repair, transcription, translation, and cell wall synthesis are highly connected in the interaction networks.

Even though a number of experimental and theoretical studies have helped in unraveling protein interaction networks in various organisms, many of these studies do not give sufficient insight into the molecular details of the protein-protein recognition process. Interacting pairs of amino acids which control the specificity of recognition can be identified only when crystal structures are available for protein-protein complexes or interacting pairs of protein have been analyzed by specific site directed mutagenesis studies (Kube *et al.*, 1992). Protein-protein recognition is governed by non-covalent interactions. Janin *et al.* (2008) have analyzed the noncovalent interactions responsible for stabilization of protein-protein complexes. They have also deduced that the biologically significant interfaces are well packed, whereas the interfaces involving nonspecific interactions are loosely packed. In contrast to protein-protein interactions, molecular details of the recognition process are better understood for protein-peptide interactions.

Many protein-protein interactions in the cells are mediated by short contiguous peptide stretches on the proteins which specifically interact with modular domains in the interacting protein. Modular signal transduction domains like SH2, FHA etc. also recognize short peptide motifs on their interaction partners (Pawson and Nash, 2000). In addition, a number of cellular processes are also governed by interaction of short length polypeptides with the proteins. Several receptors have peptide fragments as ligands e.g. Major histocompatibility complex (MHC), receptors of nervous system, receptors for endocrine peptide hormones etc. Thus, understanding molecular details of interactions between proteins and short peptides motifs is essential for identification of interaction networks and dissecting underlying mechanisms of several major cellular processes. Among the various proteins which interact specifically with short peptide motifs, MHC and kinases represent two major protein families whose substrate specificities have been extensively studied by various experimental approaches.

1.2 PROTEIN KINASES

The protein kinases are an important class of proteins which phosphorylate a number of proteins in cells. The importance of protein kinases can be estimated by the fact that about 30% of all the proteins in the human genome are phosphorylated (Cohen, 2000; Ubersax and Ferrell, 2007), reflecting the importance of the kinases. Post translational modification of proteins by kinases has evolved as a means of transferring the information among the proteins in an orderly manner. A majority of cellular processes such as transcription, translation, replication, signal transduction, immune responses, cell growth, differentiation, apoptosis etc. are known to be regulated by means of phosphorylation and dephosphorylation of regulatory proteins by kinases and phosphatases respectively, which modify the activity of regulatory proteins. This phosphorylation and dephosphorylation event either results in the change of functional state of the protein, or produces a docking site for the modular interaction domains present in interacting partner proteins (Bauman and Scott, 2002). The change in the functional state of the protein upon phosphorylation often forms the part of positive or negative feedback mechanisms aimed to tightly control the regulation of the cellular processes (Morgan, 1997). The docking site arising as a result of phosphorylation event recruits a number of proteins containing modular docking domains such as SH2, PTB, polo-box domain, FF domain, BRCT domain, WD40 domain, MH2 domain, FHA domain, 14-3-3 domain etc. (Pawson and Nash, 2000; Pawson *et al.*, 2002). Phosphorylation event by kinases, followed by interaction through modular domain forms a network of dynamically interacting proteins which is essential for transfer of information as well as proper functioning of cellular processes (Pawson and Scott, 1997). Due to their involvement in many processes, kinases have been found to be associated in many diseases including various types of cancers, diabetes and inflammation (Pawson, 1994). As a result, they are a target for treatment of a number of diseases and the role played by kinases in various diseases is an important area of active research.

Protein kinases form one of the largest superfamily of proteins called "eukaryotic protein kinase (ePK)". All the family members of protein kinases share a common homologous catalytic domain of approximately ~250-300 amino acids (Hanks and Hunter, 1995). However, the length of protein kinases is usually more than 300 amino acids. The rest of the sequence comprises of the accessory domains. A combination of these domains imparts the unique role to protein kinases in their function. In most cases, the regulation of protein kinases is also achieved through these regulatory domains. Apart from higher vertebrates, the members of ePK have also been found in a wide range of phyla such as plants, fungi and prokaryotes. Before the identification of eukaryotic like PKs (Kennelly, 2002) in bacteria, it was widely considered that they were exclusive to higher eukaryotic species where they are involved in phosphorylation-dephosphorylation networks. However the discovery of phosphorylation events in *Escherichia coli* and *Salmonella typhimurium* provided the evidence for the existence of protein kinases in prokaryotes. A number of eukaryote like protein kinase genes have now been identified and characterized from bacterial genomes. There is another set of protein kinases called atypical protein kinases (aPK) (Kennelly, 2002) which do not share sequence homology to eukaryotic kinases, but have been shown to contain kinase activity. These kinases have been shown to have similarity to ePKs at the structural fold level (Kennelly, 2002). At functional level, the protein kinases are also broadly classified into two classes, Ser/Thr kinases and Tyr kinases. Ser/Thr kinases transfer the phosphate group specifically onto serine or threonine residue in the substrate protein while Tyr kinases phosphorylate tyrosine residues. Most of the kinases identified and characterized are Ser/Thr kinases belonging to various functional subclasses. Tyr kinases show more similarity to each other and belong to the Tyrosine kinase group (Hanks and Hunter, 1995). There is also a small number of kinases called dual specificity kinases capable of phosphorylating Ser/Thr as well as Tyr residue.

The catalytic domain of the protein kinase is about 250-300 amino acids long. The phosphotransfer reaction is catalyzed by binding and orientation of ATP/GTP-Mg^{2+}/Mn^{2+} complex, as well as the substrate protein to the kinase, and the transfer of the phosphate group from ATP/GTP to the hydroxyl group of Ser, Thr or Tyr residue (Hanks and Hunter, 1995) on the substrate protein. Sequence analysis has revealed a

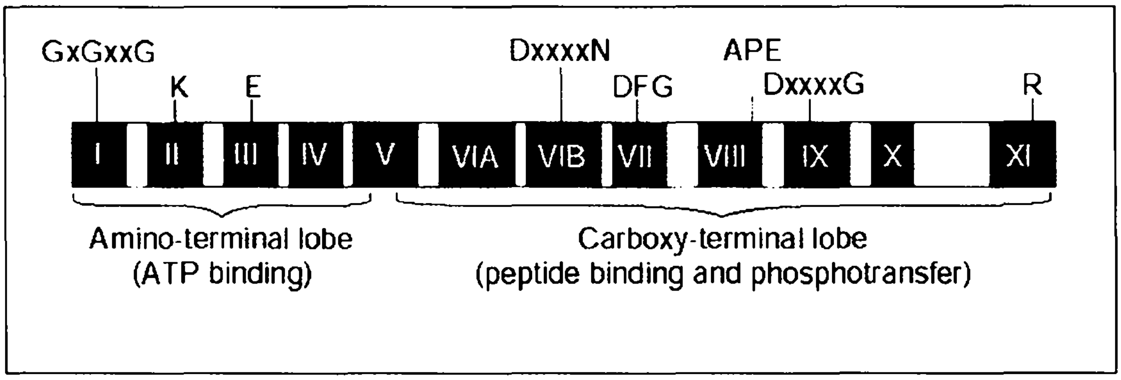


Figure 1.1: The ePK catalytic domain. The 12 conserved sequence stretches are indicated by Roman numerals. The positions of amino-acid residues and motifs highly conserved throughout the ePK superfamily are indicated above the conserved stretches using the single-letter amino-acid code with x as any amino acid. Crystal structures show that ePK domains adopt a common fold consisting of amino-terminal and carboxy-terminal lobes connected by a hinge region. Binding of Mg-ATP is largely the function of the amino-terminal lobe and hinge region, while peptide-substrate binding is mediated by the carboxy-terminal lobe. Particularly important for catalytic function are the invariant lysine in conserved stretch II and the invariant aspartate in the conserved stretch VII that function to anchor and orient ATP, and the invariant aspartate in subdomain VIB which is the likely catalytic base in the phosphotransfer reaction. (Figure is adapted from Hanks, S. K., 2003.)

number of conserved features in kinase domains. A total of 12 conserved sequence stretches have been identified in the kinases (Hanks and Hunter, 1995; Hanks, 2003). Figure 1.1 depicts the arrangement and function of these conserved sequence stretches on the kinase domain.

1.2.1 Classification of protein kinases

Protein kinases have been classified into various groups on the basis of sequence similarity in their catalytic domain, the presence of accessory regulatory domains, and their mode of regulation. For example, PKA are activated by cyclic-AMP, CDKs work in complex with cyclin proteins (Morgan, 1997) which help in recruiting their substrates, and CaM kinases are dependent on calcium binding proteins calmodulins (Hunter and Schulman, 2005). Based on the alignment of the available sequences of protein kinase catalytic domains, Hanks and Hunter in 1995 proposed a classification scheme wherein they grouped all eukaryotic protein kinases into 5 major groups (Hanks and Hunter, 1995). The first four groups were based on the phylogenetic tree obtained from the sequences of the catalytic domains. The last group contained those kinases which did not cluster, and hence could not be classified into any of the other four groups. Each one of these major kinase groups comprise of families which had similar substrate specificity and common mode of regulation and hence could be used to classify the new sequences. These major groups are AGC group (containing PKA, PKG and PKC families), CaMK group (calmodulin dependent protein kinases), CMGC group (including CDK, MAPK, GSK and CLK families), protein tyrosine kinase group (PTK), and other protein kinase group. The sequences of the protein kinases and their alignments have been catalogued and made publicly available through protein kinase resource website (Niedner *et al.*, 2006; Smith *et al.*, 1997) (http://www.nih.go.jp/mirror/Kinases/, http://pkr.genomics.purdue.edu/pkr/Welcome.do).

A high degree of conservation in the catalytic domain of protein kinases has also enabled the identification of new kinases on the basis of sequence similarity in the catalytic domains. In fact, a combination of sequence based methods such as sequence similarity searches using BLAST, PSI-BLAST, RPS-BLAST and HMM profiles of known protein kinase catalytic domains, have enabled the identification of putative protein kinases in a number of organisms whose full genomes have been sequenced. Completion of such whole genome sequencing projects for a number of

Table 1.1: The list of organisms whose protein kinase complement "kinome" has been identified. The list contains the total number of kinases identified and citation to the work reporting it.

Organism	**Number of kinases**	**Reference**
Saccharomyces cerevisiae	113	Hunter and Plowman, 1997
Ceanorhabditis elegans	411	Plowman *et al.*, 1999
Drosophila melanogaster	239	Morrison *et al.*, 2000
Homo sapiens	518	Krupa and Srinivasan, 2002; Manning *et al.*, 2002b
Mus musculus	540	Caenepeel *et al.*, 2004
Plasmodium falciparum	65	Ward *et al.*, 2004
Sterechinus neumayeri	353	Bradham *et al.*, 2006
Dictyostelium discoideum	285	Goldberg *et al.*, 2006
Entamoeba histolytica	307	Anamika *et al.*, 2008
Monosiga brevicollis	380	Manning *et al.*, 2008

species has resulted in identification, cataloguing and analysis of "kinome" complement of various genomes. Such whole genome studies have classified the putative kinase sequences into major protein kinase families based on their sequence similarities. The kinome of human (Krupa and Srinivasan, 2002; Manning *et al.*, 2002b), mouse (Caenepeel *et al.*, 2004), Caenorhabditis (Plowman *et al.*, 1999), sea urchin (Bradham *et al.*, 2006), drosophila (Morrison *et al.*, 2000), yeast (Hunter and Plowman, 1997), Dictyostelium (Goldberg *et al.*, 2006), E. histolytica (Anamika *et al.*, 2008), and plasmodium (Ward *et al.*, 2004) have been identified and classified into major kinase groups. Although many of these groups are conserved across different species in evolution, some of the kinases have been found to be unique to some organisms (Manning *et al.*, 2002a). Table 1.1 lists the species whose "kinome" complement has been identified, and the number of protein kinases in their genomes.

The initial classification of protein kinases by Hanks and Hunter (1995) comprising of five groups was further extended by addition of four new groups after the identification of yeast, worm and fly kinomes. The additional four groups are STE (comprises of MAPK cascade families Ste7/MAP2K, Ste11/MAP3K, and ste20/MAP4K), CK1 group (contains CK1, TTBK, and VRK families), TKL (contains MLK, LISK/TESK, IRAK, Raf, RIPK, and STRK families that resemble both Ser/Thr and Tyr kinases) and RGC group. This extended classification system consisted of 134 families, and 201 subfamilies belonging to 9 groups (Manning *et al.*, 2002b). This included 13 families of atypical protein kinases which showed no sequence similarity with any of the ePK families but have been shown to be having protein kinase activity. These atypical kinase families are alpha, PIKK, PHDK, RIO, BRD, ABC1, TIF1, H11, FASTK, G11, BCR, TAF, and A6 (Manning *et al.*, 2002b).

Table 1.2: Kinase distribution into major groups in human and model systems. (Table is adapted from Manning *et al.*, 2003.)

Group	Families	Sub-families	Yeast kinases	Worm kinases	Fly kinases	Human kinases	Human pseudo-genes	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	9	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

The human kinome has been analyzed and consists of 518 kinases. Out of these, 478 kinases belong to ePK and 40 belong to aPK families (Manning *et al.*, 2002b). Table 1.2 and Figure 1.2 show the distribution of human kinases and kinases from other model organisms into various families belonging to 9 major groups. The comparison of the protein kinase families in these genomes has indicated a number of interesting evolutionary relationships among each other. The 51 protein kinase families are found in all fly, worm, yeast and human genomes, which indicates that kinases belonging to these families function in cellular processes essential for existence of eukaryotic cells. In addition, there are 93 families common to fly, worm and human indicating that these families function in the processes specific to metazoan cells (Manning *et al.*, 2002b). In another study, Krupa and Srinivasan (2002) have also carried out analysis of human genome to identify the functional protein kinases. They have identified 448 distinct eukaryotic protein kinase sequences. Their analysis of the domain combinations in kinase sequences has

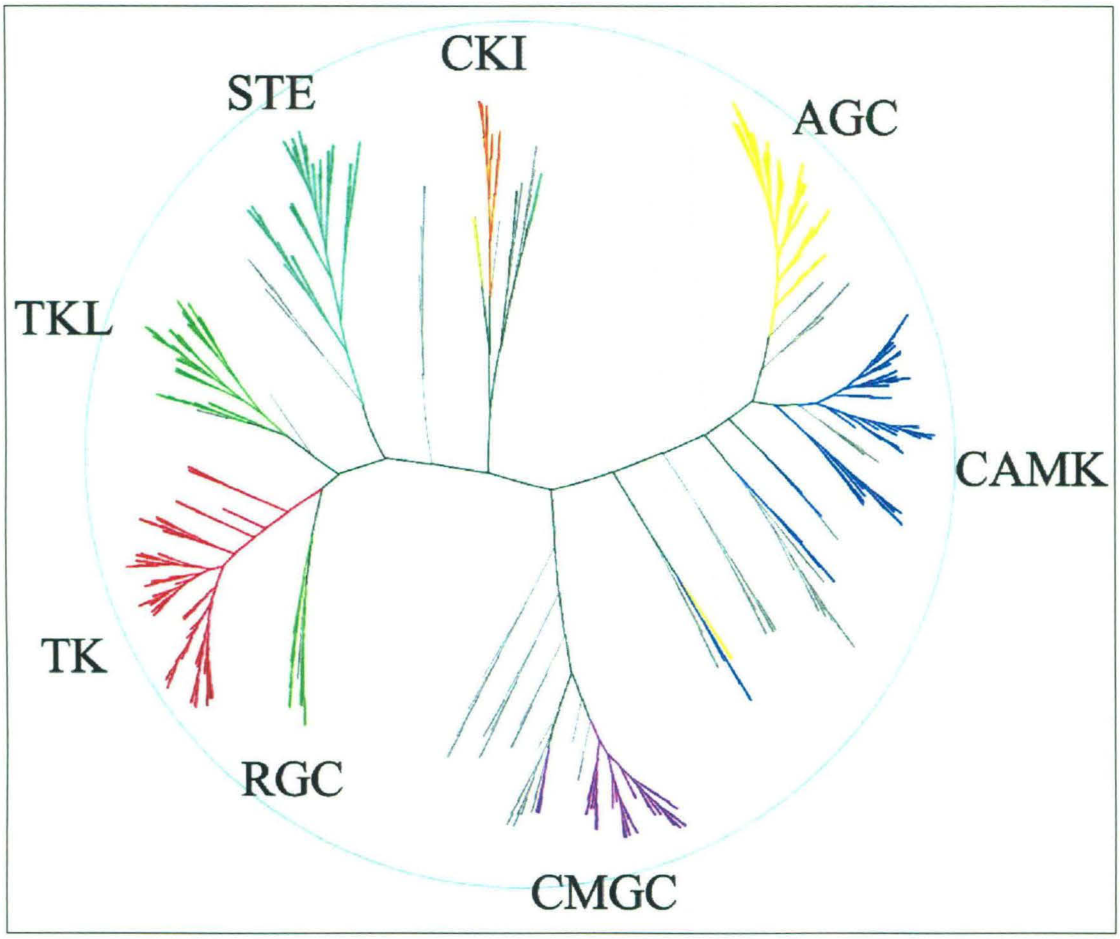


Figure 1.2: Dendrogram of 491 ePK domains from 478 genes of human genome. Major groups are labeled and colored. (Figure is adapted from Manning *et al.*, 2002b.)

suggested the roles of several kinases in a number of pathways. Their analysis has also indicated the alternative modes of regulation for kinases as suggested by interesting combination of regulatory domains (Krupa and Srinivasan, 2002).

Various genome mining studies for identification of kinase genes make use the HMM profiles derived from the experimentally characterized sequences of protein kinase domains. The identified sequences are then clustered, and classified into various families, subfamilies and groups primarily based on sequence similarity, accessory domains, and the mode of regulation. The kinase sequences identified from the genomic sequences of various species have been catalogued in the KinBase resource (http://www.kinase.com/kinbase/). The sequence entries in the KinBase are verified by experimental characterization of their cDNA sequences. Miranda-Saavedra and Barton (2007) have recently used the family specific HMM profiles for finer classification of the protein kinases of 21 genomes. This strategy was able to classify some of the previously unclassified kinases into main ePK families for a number of genomes. For example, 27 out of 28 kinases which were till now classified into "other kinases" in yeast, were classified into AGC, CAMK, CMGC and STE families (Miranda-Saavedra and Barton, 2007).

Protein kinases have also been discovered in the bacterial genomes. Krupa *et al.* (2004) have analyzed in detail the kinases identified from the prokaryotic genomes and their domain combinations, in addition to eukaryotic domains. A comprehensive database KinG (http://hodgkin.mbu.iisc.ernet.in/~king) was developed by them which contains the sequences of protein kinases along with various functional domain combinations identified using the Pfam HMM profiles (Krupa *et al.*, 2004). The latest release of the database contains 17457 protein kinases belonging to 54 eukaryotic genomes, 275 eubacterial genomes, 49 archaeal genomes and 19 viral genomes. KinG also provides a tool for the identification of catalytic domain in the putative sequences and other functional domains present alongside kinase domain (Krupa *et al.*, 2004). The comparative analysis of such domains across genomes gives interesting insights into evolutionary aspects of the functions of kinases.

1.2.2 Structure of protein kinases

Long before the first crystal structure of any kinase was solved, the high degree of conservation in the sequence of catalytic domain of the protein kinases had indicated that they are likely to take a similar fold. Protein kinase A (PKA) was the first kinase whose crystal structure was solved. It was a binary complex with a pseudosubstrate peptide (TTYADFIASGRTGRRNAIHD) having an alanine instead of serine residue at the phosphorylation site (Bossemeyer *et al.*, 1993; Knighton *et al.*, 1991). Soon after that, a ternary structure with pseudosubstrate peptide and ATP was solved (Zheng *et al.*, 1993). Subsequently a number of crystal structures of different protein kinases have been solved. The crystal structures of kinases solved till date have confirmed that all kinases adopt similar structural fold. These crystal structures have also revealed the roles of various conserved motifs found in the kinase sequences.

The kinase fold consists of two lobes, a small N-terminal lobe consisting mainly of β sheets, and a large C-terminal lobe consisting mainly of α helices. Figure 1.3 shows the structure of PKA highlighting important structural features of the kinase fold. The N-terminal lobe consists of 5 antiparallel β sheets and a small α helix. This includes conserved stretches I-IV and primarily contains the ATP binding site and hence is involved in the binding and correct orientation of ATP/GTP-Mn^{2+}/Mg^{2+} complex. The N-terminal domain also contains conserved GXGXXG (X represents any amino acid) motif between β1 and β2 strands. The binding site of the substrate peptide is located in the cleft between the two lobes (Hanks and Hunter, 1995; Zheng *et al.*, 1993). The majority of the contacts the peptide makes with the kinase is in the C-terminal lobe. A long activation loop of 20-30 residues forms a part of C-terminal lobe and occupy the region between the clefts. This activation loop adopts different conformations in active and inactive states of the kinase and contains the Asp166 which facilitates the phosphotransfer reaction by extraction of a proton from hydroxyl group of the substrate. The conformation of activation loop is an important determinant of the active or inactive form the protein kinases, and is controlled by the phosphorylation of Ser/Thr residue in activation loop (Hanks and Hunter, 1995). This phosphorylation induces many secondary changes like re-orientation of the lobes and the positioning of the catalytic residues in relation to the substrate peptide.

A number of crystal structures of protein kinases have been solved. These kinases are in complex with the ATP, GTP, kinase inhibitor compounds, substrate peptide and/or inhibitor peptide. The substrate peptide bound structures of kinases have provided valuable information for understanding the molecular basis of the substrate recognition by the protein kinases. The crystal structures in complex with substrate peptide have been solved for PKA (Knighton *et al.*, 1991; Zheng *et al.*,

Figure 1.3: Ribbon diagram of Catalytic subunit of cAMP-dependent protein kinase in ternary complex with Protein kinase inhibitor substrate, and MgATP. Conserved residues scattered throughout the core are indicated by dots (Gly50, Glu52, Lys72, Glu91, Asp166, Asn171, Asp 184, Glu208, Asp220, and Arg280). Several of the side chains are indicated. The salt bridge between Arg280 and Glu208 is indicated by dashed lines. The numbering and nomenclature of the β strands and helices is based on Knighton et al. (1991). Asp166, near the site of phosphotransfer, is positioned to serve as catalytic base. The Ala side chain at the P-site in the inhibitor is also marked with an arrow indicating where phosphotransfer would take place if a Ser was located at this site. The dotted line bridging the β and γ-phosphates indicate the position of the activating Mg^{2+} ion. (Figure is adapted from Hanks and Hunter, 1995)

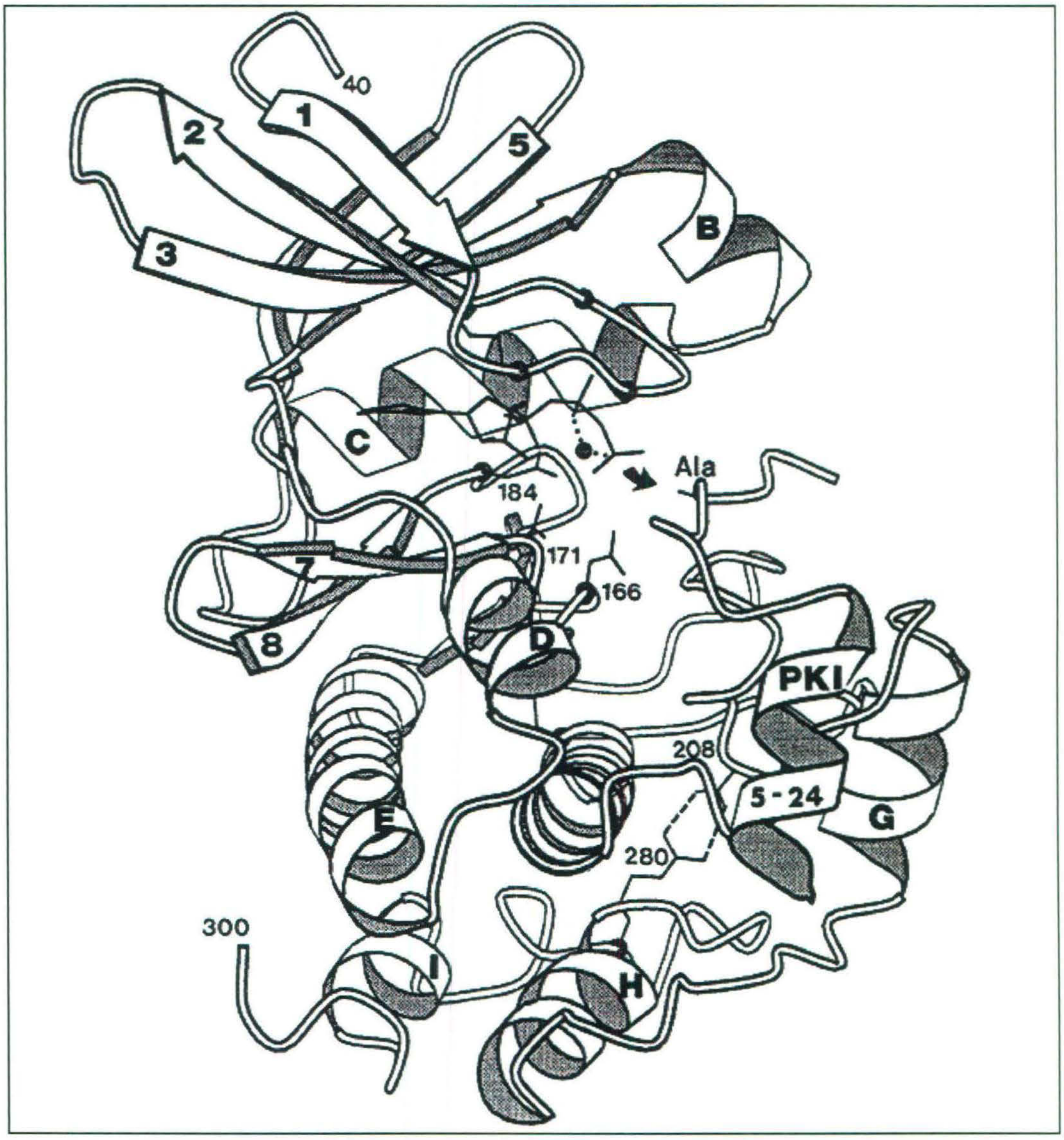


Figure 1.3

1993), PKB (Yang *et al.*, 2002), CDK2 (Brown *et al.*, 1999), and PHK (Lowe *et al.*, 1997) which are Ser/Thr kinases and ABL (Levinson *et al.*, 2006), and IRK (Parang *et al.*, 2001) which are Tyr kinases. All these crystal structures reveal that the substrate peptide binds in an extended conformation in the cleft between the two lobes of kinase. PKA structure has been solved in complex with an inhibitor peptide which has an Ala in place of Ser or Thr at the phosphorylation site. The structures of other kinases have been solved without the substrate peptides. These crystal structures have revealed that the side chains of the substrate amino acids are accommodated in the binding pockets formed between the two lobes of the kinases. The site of phosphorylation is called P0 and the residues towards N and C terminal side to P0 are referred to as P-1, P-2, P-3 and P+1, P+2, P+3 and so on. The binding pockets on the protein kinase which accommodate these residues have been designated as S-1, S-2, S-3, S+1, S+2, S+3 and so on. Figure 1.4 depicts the substrate binding in PKA and various binding pockets accommodating the side chains of the substrate peptide.

1.2.3 Substrates of protein kinases

Protein kinases play major functional role in a variety of cellular processes. As a consequence, the protein kinases phosphorylate a large number of proteins. It has been estimated that more than 30% of all proteins encoded in the human genome are phosphorylated (Cohen, 2000; Ubersax and Ferrell, 2007). Even though based on genome analysis it has been possible to identify the protein kinase complement of various organisms, the substrate proteins phosphorylated by these kinases still remain largely unknown. Even after years of targeted biochemical studies by several groups, only a small subset of kinase substrates has been identified so far. As the cascades of phosphorylation network play a major role in signal transduction (Pawson and Scott, 1997; Pawson *et al.*, 2002), the identification of substrate proteins of the kinases is crucial for understanding the information transfer during cellular processes. However, the identification of the substrates of kinase has been a slow process. The characterization of the substrate protein of a kinase based on biochemical assays requires an educated guess about the choice of substrate by the given kinase.

1.2.3.1 Experimental methods for identification of substrates of kinases

Biochemical assays used for identification of the substrates of kinases involve incubation of purified protein with the protein kinase in presence of radiolabeled ATP

Figure 1.4: Substrate binding in protein kinase A. (A) Schematic representation of the binding sites of the side-chains of the substrate peptide with the specificity-determining residues (SDRs or determinants) listed in each subsite. The sub-sites are coloured: S-3, red; S-2, yellow, S-1, green; S0, orange-red; S+1, dark blue; S+2, magenta; and S+3, light blue. The same colour scheme for the subsites is used in (B) and (C). (B) Interactions of the heptapeptide region of the substrate (grey; sequence RRASIHD) with the SDRs, coloured according to the subsite. (C) Surface representation highlighting the individual subsites, coloured as in (A), and a heptapeptide region of the substrate (black; sequence RRASIHD). (Figure is adapted from Kobe *et al.,* 2003.)

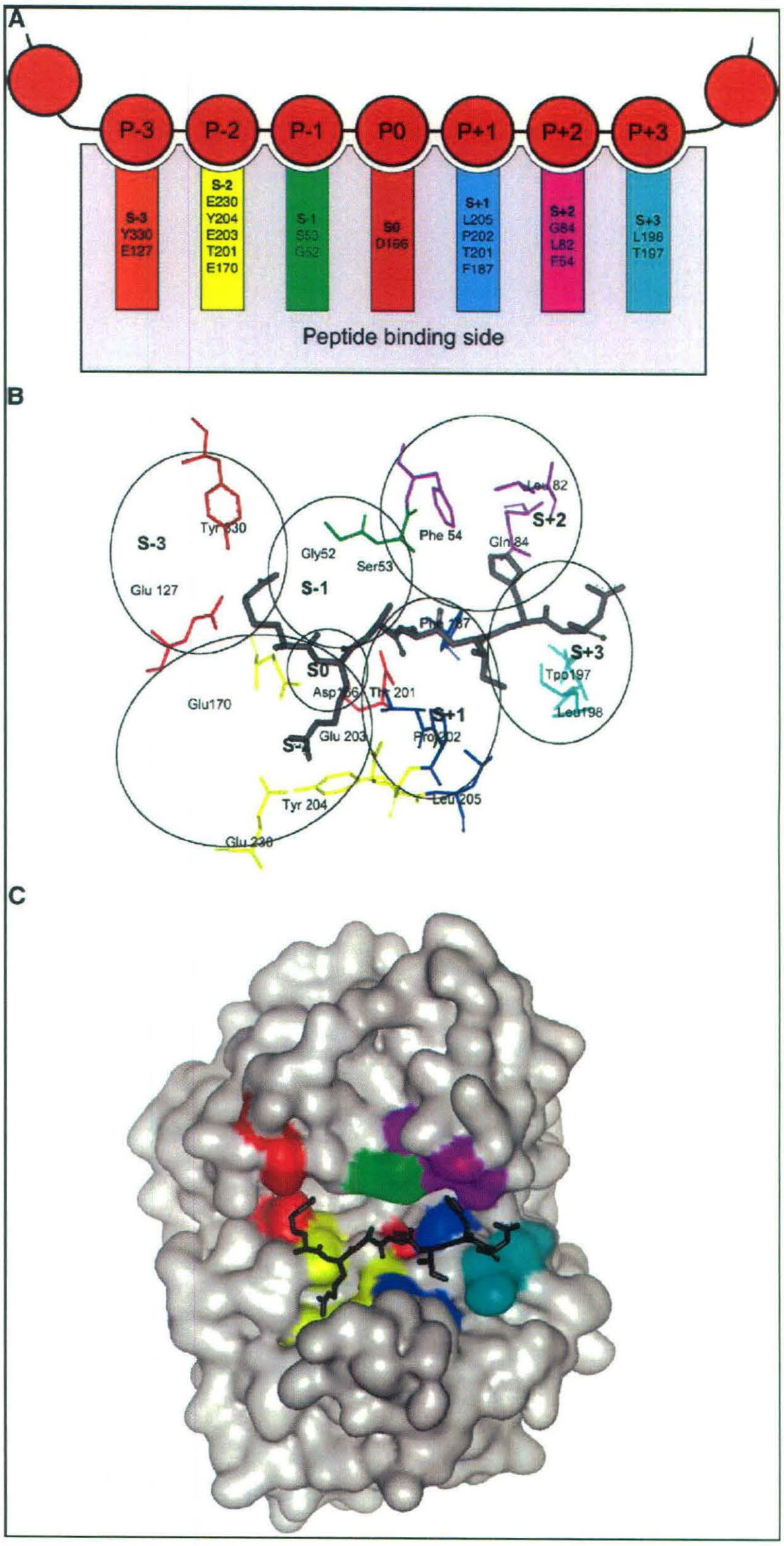

A
P-3
P-2
P-1
P0
P+1
P+2
P+3
S-3
Y330
E127
S-2
E230
Y204
E203
T201
E170
S-1
S53
G52
S0
D166
S+1
L205
P202
T201
F187
S+2
G84
L82
F54
S+3
L198
T197
Peptide binding side
B
Tyr 330
S-3
Glu 127
Gly52
Ser53
S-1
Phe 54
Leu 82
S+2
Gln 84
Phe 187
S0
Asp166
Thr 201
S+3
Tpo197
Leu198
Glu170
S+1
Glu 203
S-2
Pro 202
Tyr 204
Leu 205
Glu 230
C

Figure 1.4

under optimal conditions. This is followed by detection of phosphorylation state by autoradiography (Wooten, 2002). Although this method requires the previous experimental cues for the choice of putative substrate, this is still a widely used method. A modification of this method involves the use of whole cell lysate incubated with the kinase of choice to identify the subset of proteins which are phosphorylated by kinase. Another method involves the use of oriented peptides libraries which is based on the idea that kinases recognize primary sequence motif in the substrate proteins. In this method, a mixture of randomized soluble distinct peptides of same length, and containing only one phosphorylatable site, is incubated with the kinase of choice in the presence of radiolabeled ATP (Songyang *et al.*, 1994). Cognate peptides are phosphorylated by kinase which are separated on a ferric column, and sequenced by Edman degradation. The analysis of phosphorylated peptide sequence gives information about the relative abundance of amino acids at each position and hence the phosphorylation motif preferred by that kinase (Nishikawa *et al.*, 1997; Songyang *et al.*, 1994; Songyang *et al.*, 1996). Phage display libraries have also proved useful in identifying a set of putative substrates for a selected kinase (Cujec *et al.*, 2002). A relatively new development involves the mass spectrometric identification of phosphorylated peptides from whole cell lysates (Mann *et al.*, 2002; McLachlin and Chait, 2001). Even though this method can identify the phosphoprotein and the site of phosphorylation, this does not provide any information about the kinases responsible for the phosphorylation events.

Techniques for *in vivo* identification of protein kinase substrate make use of the phospho-specific antibodies for tracking of the phosphorylated protein. These antibodies can identify the single phosphorylated amino acid in the protein (Zhang *et al.*, 2002). The phosphorylation motif specific antibodies are also used which can recognize and bind to a certain motif specific for a kinase (Zhang *et al.*, 2002). The substrate proteins are identified by mean of immunoprecipitation or immunoblotting and analyzed by mass spectrometry. Some methods such as yeast two-hybrid screens (Yang *et al.*, 1992), which are based on the physical interaction between the kinase and substrate, are also used for identification of kinase substrates (Cujec *et al.*, 2002).

1.2.3.2 Protein kinase substrate databases

Various *in vitro* and *in vivo* experimental techniques have led to the identification of a number of substrate proteins for a large number of kinases. The amount of experimental data on kinase substrates is growing rapidly. This data is extremely valuable for understanding the mechanism of substrate recognition by protein kinases. Therefore, several groups have attempted to organize the information on substrate proteins along with their phosphorylation sites in the form of the databases. Phospho.ELM is one such widely used database which catalogues the experimentally verified substrates of the protein kinases along with the site of phosphorylation (Diella *et al.*, 2004; Diella *et al.*, 2008). The database is also interlinked to the SWISSPROT and the PUBMED. If the structure of the protein is available then the PDB ID of the substrate protein is also reported. This searchable database has been extensively used for the development of prediction rules for *in silico* identification of protein kinase substrates. PhosphoSitePlus (www.phosphosite.org) is another database containing the phosphorylation sites along with a variety of associated information such as subcellular localization, tissue types, and cell lines used in experiments. The sites identified in the MS/MS experiments are also catalogued. PHOSIDA is a database of phosphorylation sites obtained from the high throughput mass-spectrometry based proteomics analysis (Gnad *et al.*, 2007). This database also gives information about the predicted secondary structure of the phosphorylation sites along with other attributes of the protein and phosphorylation sites. PhosphoPOINT (Yang *et al.*, 2008) annotates the interaction between kinase and phosphoproteins along with the cataloguing of substrate proteins and phosphorylation sites. It also lists cSNPs that result in disease phenotypes. mtcPTM (Jimenez *et al.*, 2007) is a collection of human and mouse phosphorylation sites and patterns under different physiological conditions. It also stores the information about structural features of the phosphorylation sites (Jimenez *et al.*, 2007).

1.2.3.3 Phosphorylation site prediction programs

Although various *in vitro* and *in vivo* experimental approaches have succeeded in identifying substrate proteins and phosphorylation sites for a significant number of kinases, substrates for a large number of kinases are still unknown. Discovery of the kinome complement of genome for several organisms has necessitated the identification of kinase substrates so as to better understand the role of kinases in various cellular functions and pathways. The experimental discovery of the substrates at the genomic scale is impractical, if not impossible. Therefore, a number of computer programs have been developed based on the analysis of the phosphorylation

Table 1.3: List of the commonly used computational tools and their methodology for the prediction of the sites of phosphorylation by various kinases.

Software	World wide web link	Methodology
DIPHOS 1.3	http://www.ist.temple.edu/DISPHOS	Intrinsic disorder in phosphorylation sites
SCANSITE 2.0	http://scansite.mit.edu/	PSSM
KinasePhos 2.0	http://kinasephos2.mbc.nctu.edu.tw/	SVM
NetPhos 2.0	http://www.cbs.dtu.dk/services/NetPhos/	ANN
NetPhosK 1.0	http://www.cbs.dtu.dk/services/NetPhosK/	ANN
GPS	http://bioinformatics.lcd-ustc.org/gps_web/faq.php	Clustering
PPSP	http://bioinformatics.lcd-ustc.org/PPSP/	Bayesian model
PREDIKIN	http://florey.biosci.uq.edu.au/kinsub/home.htm	Rules based on structure and sequence analysis

sites of various families of protein kinases. These prediction programs employ a number of methodologies such as position specific scoring matrices (PSSM), clustering and bayesian analysis, artificial neural network (ANN), support vector machines (SVM), and structural analysis of kinase-peptide complexes. All of these programs are based on the rules derived from the already known phosphorylation sites of some of the protein kinases. Some of these programs predict only for a specific protein kinase family e.g. pkaPS (Neuberger *et al.*, 2007) predicts only for PKA. Similarly, Cheng *et al.* (2007) have developed a method for prediction of substrates of CDK. On the other hand, there are generalized methods predicting substrates for a number of protein kinases. Table 1.3 lists some of the widely used computer programs for prediction of phosphorylation sites in the putative substrate proteins of various kinases. The salient features of some of the major phosphorylation site prediction programs are described below.

1.2.3.3.1 SCANSITE

SCANSITE (Obenauer *et al.*, 2003) is one of the widely used programs for the prediction of protein kinase phosphorylation sites. It uses the protein kinase specific position specific scoring matrices (PSSM) derived from the peptide library experiments. PSSM for each kinase represents the favored amino acid residues flanking the phosphorylation site Ser/Thr/Tyr residues. Based on the score of each Ser/Thr/Tyr containing putative peptide, high scoring peptides are predicted as phosphorylation sites. SCANSITE also takes into account the predicted surface accessibility value of peptides for prediction of the phosphorylation sites. However, a major limitation of the SCANSITE program is that, it can predict substrate only for those kinases for which peptide library data is available.

1.2.3.3.2 KinasePhos

The first version of KinasePhos (Huang *et al.*, 2005) used the kinase specific HMM profiles generated from the known phosphorylation sites for predicting the phosphorylation sites in putative proteins. The second version, KinasePhos 2.0 (Wong *et al.*, 2007) employed the SVM model based on the coupling patterns between the amino acids at different positions in the known substrate proteins.

1.2.3.3.3 NetPhos

Netphos (Blom *et al.*, 1999) uses machine learning approach. Artificial neural network (ANN) trained on a dataset of known phosphorylation sites and non-phosphorylation sites is used to predict whether a Ser/Thr/Tyr containing peptide is phosphorylated or not.

1.2.3.3.4 NetPhosK

Instead of kinase independent ANN model used by NetPhos, NetPhosK (Blom *et al.*, 2004) uses kinase specific ANN models trained for prediction of substrate for specific kinase families. It predicts for PKA, PKC, PKG, CKII, Cdc2, CaM-II, ATM, DNA PK, Cdk5, p38 MAPK, GSK3, CKI, PKB, RSK, INSR, EGFR and Src kinase.

1.2.3.3.5 GPS

GPS: group based phosphorylation site predictor (Xue *et al.*, 2005), uses statistical model for predicting the phosphorylation sites of kinases. It groups the protein kinases similar in sequence and function together. The phosphorylation sites for each of these groups are clustered together based on the similarity in the amino acids at each position. The clustering is done based on the BLOSUM62 similarity matrix scores for the alignments of substrate peptides. If a peptide shows similarity to the phosphorylation site peptides of a kinase group, as judged by the BLOSUM62 score, that peptide is predicted as the phosphorylation site for that protein kinase. GPS predicts for about 70 protein kinase groups belonging to Ser/Thr and Tyr kinase.

1.2.3.3.6 PPSP

PPSP: prediction of PK-specific phosphorylation site (Xue *et al.*, 2006) uses statistical model based on the Bayesian decision theory for phosphorylation site prediction. Like GPS, it also groups the similar kinases into different groups. BLOSUM62 matrix is used in the Bayesian statistical model. PPSP also predicts for about 70 protein kinase groups.

GPS and PPSP both make use of statistical models based on clustering and bayesian analysis of sequences of known substrates respectively. Since for every kinase family, the number of known substrate peptides is not large, GPS and PPSP group the substrates of similar kinases together to increase their number for each kinase group. All of these programs are trained on the already available protein substrates data while employing different sequence based approaches for prediction. Although these programs perform well for those protein kinases, for which sufficient amount of substrate information is available, they cannot predict for other protein kinases for which little or no substrate information is available as yet.

1.2.3.3.7 PREDIKIN

PREDIKIN (Brinkworth et al., 2003) uses the empirical rules derived from examination of kinase sequences, kinase-peptide structures, and peptide library data. For a given kinase sequence, it identifies the residues that would interact with three peptide residues on each side of phosphorylation site in the substrate, and predicts a motif most likely to be compatible for binding with the kinase. The substrate proteins which contain the predicted motif for a kinase, are likely to be its target phosphorylation sites. In contrast to the large number of sequence based phosphorylation site prediction programs, PREDIKIN uses a structure based approach. PREDIKIN has been used to reconstruct cell cycle control pathway, and DNA damage checkpoint control pathway in yeast (Brinkworth et al., 2006).

1.2.3.3.8 DIPHOS

DIPHOS (disorder Enhanced phosphorylation site predictor) (Iakoucheva *et al.*, 2004) predicts the potential phosphorylation sites in the substrate proteins based on property of sequence stretch around the phosphorylation site Ser/Thr/Tyr residues. The predictor is based on the analysis of known phosphorylation sites of protein kinases which show that the disorder around the phosphorylation site is an important prerequisite for the phosphorylation. DIPHOS, however, does not identify the protein kinase which phosphorylates the predicted phosphorylation sites.

1.2.4 In silico identification of protein phosphorylation networks

Recently, there has been increased interest in the understanding and reconstruction of signaling networks and pathways involving phosphorylation cascades at a genomic scale. Two recent studies have investigated the reconstruction of phosphorylation networks in yeast and human (Brinkworth *et al.*, 2006; Linding *et al.*, 2007). Brinkworth *et al* (2006) have used their prediction program PREDIKIN to identify the substrates for the protein kinases identified in the yeast genome (Brinkworth *et al.*, 2006). The experimentally identified phosphorylation sites in the high throughput mass spectrometric experiments were assigned to the protein kinases based on the prediction results. Linding et al (Linding *et al.*, 2007) developed an approach called NetworKIN to identify the protein kinases responsible for the phosphorylation of the *in vivo* identified sites (Linding *et al.*, 2007; Linding *et al.*, 2008). In their method, they have combined the motif based approach of SCANSITE (Obenauer *et al.*, 2003) with the context dependent information of the localization, scaffold and expression, for associating a site with the protein kinase. Success of such projects depends largely upon the correct identification of substrate proteins of the protein kinase involved in the signaling pathways. However, specific programs tend to predict well only for specific classes of kinases, making any single program unsuitable for the prediction on a genomic scale. Hence, a systematic analysis of the prediction accuracies of different programs on the known dataset is very important in deciding which program should be used for the prediction of substrates for a particular class of kinases. As the prediction accuracy of all programs varies for different kinase classes, the choice of correct program for making predictions becomes very important. Although individual programs have reported the comparison with other programs for a specific set of kinases for defined problem sets, only one study has carried out systematic analysis of prediction accuracies of these prediction programs using an unbiased phosphorylation dataset (Wan *et al.*, 2008).

1.3 MAJOR HISTOCOMPATIBILITY COMPLEX (MHC)

Major histocompatibility complex (MHC) region of genome is found in most vertebrate species and contains a tightly linked cluster of genes of immunological importance. MHC region is found on short arm of chromosome 6 in human, and on chromosome 17 in mouse. The protein products of these genes play a key role in the self and non-self discrimination and the development of cellular immune responses. The MHC cluster is also known as HLA in human and H2 in mouse. It spans and encodes three groups of genes i.e. class I, class II and class III. Class I MHC genes encode glycoproteins which are expressed on the surface of almost all nucleated cells (Bjorkman and Parham, 1990; Klein *et al.*, 1983). Class II MHC genes also encode for the glycoproteins which are expressed on the surface of professional antigen presenting cells only. Class III genes encode for the proteins of various immunological functions. In case of human, class I MHC genes are organized in three loci A, B and C whose products are referred to as HLA-A, HLA-B and HLA-C respectively. In case of mouse, these are encoded by K and D region and are called H2-K and H2-D. Class II MHC protein are encoded by IA and IE region producing H2-IA and H2-IE proteins in mouse, and DP,DQ, and DR encoding HLA-DP, HLA-DQ and HLA-DR in human (Klein *et al.*, 1983). The MHC alleles are highly polymorphic with a large pool of alternative forms existing at each locus within the population. This variation in the MHC alleles creates a vast repertoire and variability enabling them to bind a variety of antigenic peptides of foreign origin (Klein *et al.*, 1983). HLA haplotyping studies have identified a number of alleles in case of human. These alleles have been sequenced and deposited in the IMGT/HLA database (Robinson *et al.*, 2003) after approval by WHO HLA nomenclature committee. The October 2008 release of IMGT/HLA database reported 2187 class I, and 980 class II classical MHC alleles. The high degree of polymorphism has been reported to be associated with the different levels of susceptibility to various infectious diseases. MHCs have also been shown to be associated with autoimmune diseases such as multiple sclerosis, diabetes, arthritis etc.

The function of MHC molecules is to bind the processed antigens and present them to the cytotoxic T-cells. The T-cell receptor (TCR) on the surface of T-cells recognizes the MHC bound peptide and mount an immune response if the peptide is antigenic. Class I MHC molecules present the peptide to CD8$^+$ T-cells and class II molecules present to CD4$^+$ T-cells. The presentation of peptides to the TCR on T-cell ensures a continuous update of peptides of various origins for scrutiny by the immune system (Bjorkman and Parham, 1990).

1.3.1 Antigen processing

The antigenic peptides presented by class I and class II MHC molecules are of different origins and processed by different processing pathways. Figure 1.5 depicts schematically the cytosolic and endosomal pathways for antigen processing (Cresswell, 1994; Yewdell *et al.*, 2003). Class I MHC pathway for antigen processing or cytosolic pathway provides peptides for loading to MHC class I molecules (Pamer and Cresswell, 1998). These peptides are derived from the cytosolic proteins in the healthy cells. However in the infected or mutated cells, the peptides of foreign origin are also loaded onto MHC molecules and recognized by T-cells. In this pathway, the proteins in the cytosolic milieu are degraded by multicatalytic proteasome complex into smaller peptides. Some of these peptides are transported by transporter associated with antigen processing (TAP) into endoplasmic reticulum where they are loaded onto MHC class I molecules and transported to the cell surface. Since the peptides loaded in pathway are first cleaved by proteasome and transported by TAP transporter, the cleaving patterns and the selectivity of TAP have been used for assisting in the prediction of MHC binding peptides (Pamer and Cresswell, 1998; Yewdell *et al.*, 2003).

Class II MHC molecules are loaded through the endosomal pathway (Watts, 1997). The extracellular antigenic protein, microorganism etc. are internalized by the phagocytosis by specialized cells such as macrophages, dendritic cells and B-lymphocytes. Inside the cells, the phagosomes fuse with the lysosome and the internalized proteins are digested by lysosomal aminopeptidases. The cleaved peptides are loaded onto the MHC class II molecules in the golgi apparatus and the peptide loaded MHC molecules are transported to the cell surface.

1.3.2 Structure of MHC proteins

The three dimensional structure of a number of class I and class II MHC alleles have been solved by x-ray crystallography. Class I MHC consists of a 45 KDa glycoprotein α chain which is noncovalently associated with β2 microglobulin chain. The α chain folds into three domains α1 α2 and α3 (Bjorkman *et al.*, 1987). The α1 and α2 domains are nearly identical to each other and form a platform of 8 antiparallel β sheets which are spanned by two α helices. This combination of sheets and helices

Figure 1.5: Schematic representation of the peptide presentation via MHC class I and II. (Figure is adapted from Apostolopoulos *et al.*, 2004.)

Figure 1.6: Representative structures of class I (1AKJ) and class II (1A6A) MHC molecules in complex with the bound peptide.

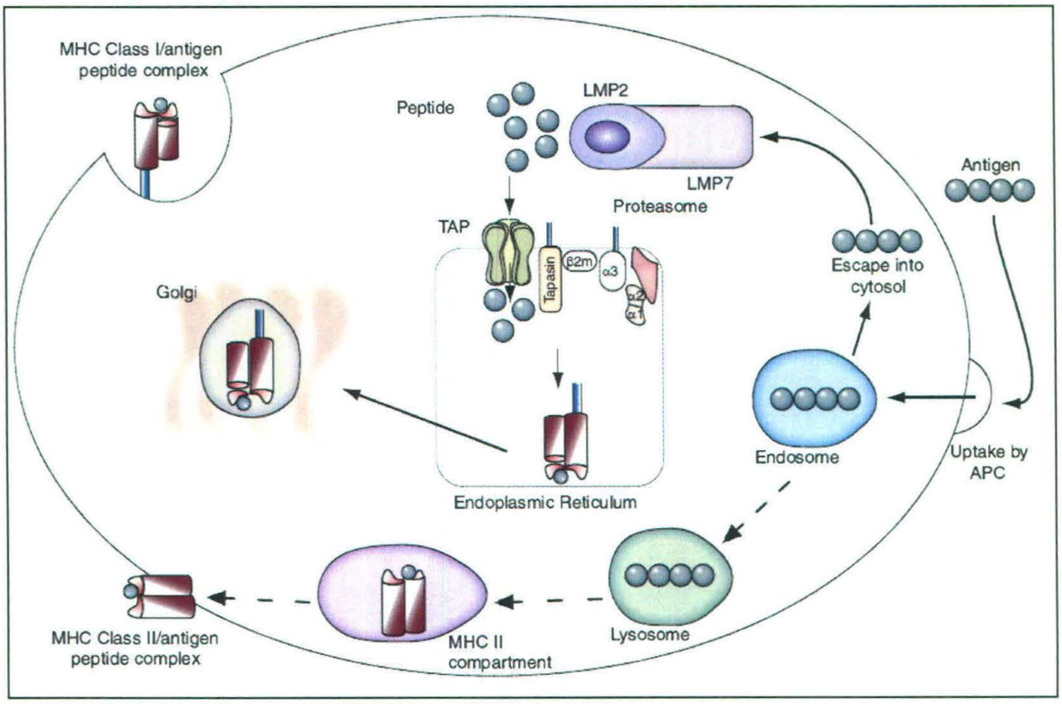


Figure 1.5

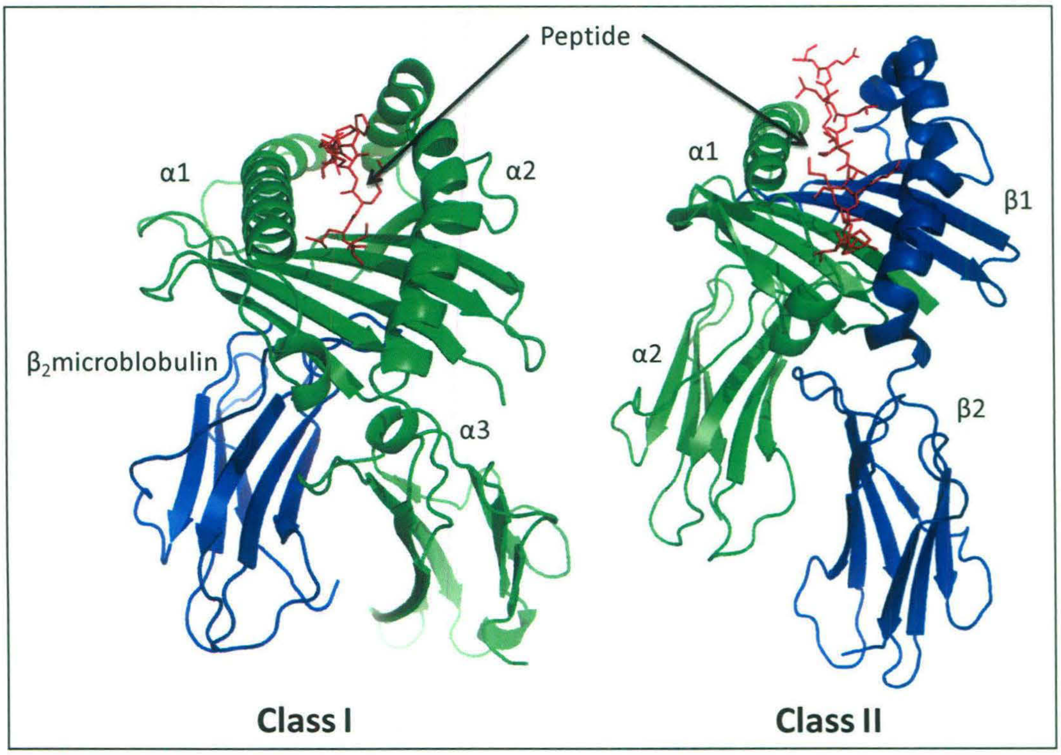


Figure 1.6

forms a well defined groove known as peptide binding cleft. The α3 and β2 microglobulin are organized into immunoglobulin fold and interact with each other (Figure 1.6) (Bjorkman and Parham, 1990).

Class II MHC structure is identical to class I MHC structure. It consists of two non-identical chains, an α-chain and a β-chain. The α-chain folds into α1 and α2 domains and β chain folds into β1 and β2 domains. The α2/ β2 pair is identical in structure to α3/β2 microglobulin pair of class I MHC. The α1/ β1 pairs form a platform to 8 antiparallel β sheets spanned by two alpha helices and form a peptide binding cleft very similar to that of MHC class I (Figure 1.6) (Brown *et al.*, 1993).

1.3.3 MHC-peptide interaction

The antigen binding groove in both class I and class II MHC molecules is nearly identical and the antigenic peptides bind in the cleft in an extended conformation running between the α helix on the floor of β sheets. MHC class I

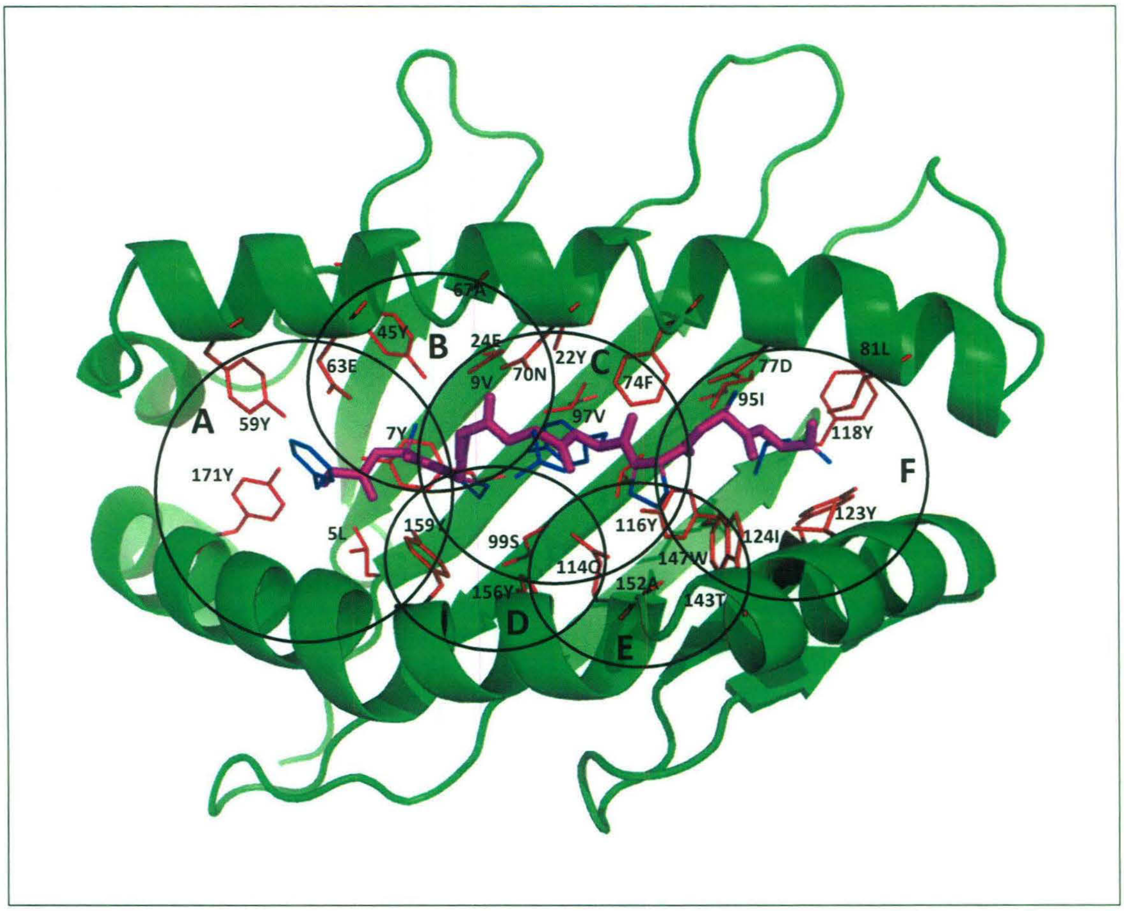


Figure 1.7: The top view of mouse MHC class I allele H2-Kd in complex with the peptide. The constituent residues of six binding pockets A, B, C, D, E, and F in the peptide binding groove are shown in red enclosed by ovals. The peptide backbone is shown in magenta and the side chains of peptide residues are depicted in blue color.

molecules bind the 8-10 mer peptides. This is explained by the fact that, peptide binding cleft in class I MHC is closed at the ends and thus, limiting the size of bound peptide (Bjorkman *et al.*, 1987; Bjorkman and Parham, 1990). MHC class I molecules can recognize vast repertoire of peptides due to the presence of so called anchor residues whose side chains are accommodated in the well defined binding pockets on the MHC (Rammensee *et al.*, 1993). These anchor residues are identified from the experimentally known binding peptides. Most of the class I binding peptides have anchor residues at the second or third positions and the carboxyl terminal. Matsumura *et al.* (1992) have reported the presence of six binding pockets designated as A, B, C, D, E, and F which accommodate the side chains of the bound peptides in class I MHC. Figure 1.7 depicts these six binding pockets and their constituent residues in H2-Kd allele.

In class II MHC molecules, the ends of the peptide binding cleft are open and the peptides of upto 18mer can be accomodated. Normally the region of peptide that interacts with the MHC is about 13 amino acids long and the remaining residues hang outside the open cleft of the class II MHC molecules. Unlike the peptides which bind class I MHC, peptides binding to class II MHCs lack well defined anchor residues, but some specificity determining residues on the peptides make key contacts with the MHC molecules (Brown *et al.*, 1993).

A large number of MHC binding peptides have been identified which specifically bind to class I and class II MHC alleles. Combinatorial peptide libraries (Wilson *et al.*, 1999) have been extensively used for the identification of the binding repertoire of various MHC alleles. Analysis of these peptides have resulted in identification of some sequence patterns which have been broadly used for predicting putative MHC binding peptides with variable degree of success (Rammensee *et al.*, 1993). In the view of the importance of the MHC epitopes in the rational design of vaccines, identification of MHC binding peptides has been an area of active research. However, polymorphism in MHC alleles (Parham *et al.*, 1995), coupled with vast variety of potential peptides has made the experimental approaches for identification of the MHC epitopes a difficult task. This has led to development of a number of computational tools for predictions of MHC epitopes.

1.3.3.1 Computational tools for prediction of MHC epitopes

A variety of computational tools have been developed for prediction of MHC binding peptides present in putative antigenic proteins. All of these methods break down the protein sequence into all possible overlapping peptides of a fixed length and calculate the MHC binding score for each of them. The peptides with high scores are classified as the potential epitopes for the MHC allele in question. These computational methods adopt various methodologies for calculating the scores for the peptides. Broadly, two types of computational methods have been developed. One group of computational methods can be broadly categorized as sequence based methods. These computational methods predict epitopes based on the sequence motifs derived from the known epitopes of various MHC alleles. The modification of this method is the matrix based methods, which use the matrices containing the frequencies of the amino acids at various positions in the peptides precalculated from the known MHC binding peptides. It is known that a significant number of peptides bind to MHC, even if they do not contain any known MHC binding motifs. They are

Table 1.4: List of commonly used computational tools and their methodology for the prediction of MHC binding peptides.

Software	**World wide web link**	**Methodology**
BIMAS	http://www-bimas.cit.nih.gov/molbio/hla_bind/	Half life of dissociation
ProPred1	http://www.imtech.res.in/raghava/propred1/	Matrices, and proteasomal cleavage
MMBPred	http://www.imtech.res.in/raghava/mmbpred/	Matrices calculated from MHCBN database
EPIPREDICT	http://www.epipredict.de/index.html	Matrices from combinatorial peptide library experiments
RANKPEP	http://bio.dfci.harvard.edu/RANKPEP/	PSSM
SYFPEITHI	http://www.syfpeithi.de/home.htm	Weighted matrix
MHCPred	http://www.jenner.ac.uk/MHCPred/	QSAR
nHLAPred	http://bic.uams.edu/mirror/nhlapred/index.html	ANN
NetMHCpan	http://www.cbs.dtu.dk/services/NetMHCpan/	ANN
NetMHC	http://www.cbs.dtu.dk/services/NetMHC/	ANN and PSSM
KISS	http://www-bs.informatik.uni-tuebingen.de/SVMHC/	SVM
SVMHC	http://cbio.ensmp.fr/kiss/	SVM
PREDEP	http://margalit.huji.ac.il/Teppred/mhc-bind/index.html	Threading type

called noncanonical MHC binding peptides and they make use of alternative binding pockets for binding to MHC (Apostolopoulos and Lazoura, 2004). Sequence based programs often fail to predict these noncanonical binders. In contrast to sequence based methods which only use the sequence information to predict putative MHC binding peptides, a second group of methods known as structure based approach, use the information from the crystal structures of MHC-peptide complex for calculating the binding scores of the peptides to various MHC alleles. The structure based programs can in principle identify noncanonical MHC binding peptides. Table 1.4 gives a list of the commonly used epitope prediction programs. A brief description of their methodology for prediction is given below.

1.3.3.1.1 BIMAS

BIMAS (Parker *et al.*, 1994) predicts binding peptides for 36 class I MHC alleles based on the predicted half life of dissociation of peptides to MHC alleles. The prediction is based on the coefficient tables containing scores for each amino acid at each position of the peptide. These coefficient tables are derived from the experimental data on the stabilization of β_2-microglobulin complex by the peptides. This method assumes that each residue of the peptide contributes to some extent towards the stability of the complex.

1.3.3.1.2 ProPred1

ProPred1 (Singh and Raghava, 2003) uses matrices derived from the BIMAS server for scoring the peptides from putative antigenic proteins. It can predict for 47 class I MHC alleles. However, instead of considering all possible overlapping peptides, it also finds the proteasome cleavage sites based on matrices derived by Toes *et al.* (2001) based on the sequence patterns in the known proteasome cleavage sites. In summary, it finds the promiscuous MHC binding peptides which have proteasomal cleavage sites at their ends.

1.3.3.1.3 ProPred

ProPred (Singh and Raghava, 2001) uses the quantitative matrices reported by Sturniolo *et al.* (1999) for prediction of class II MHC binding peptides. It can predict for 51 class II alleles belonging to HLA-DRB1 and HLA-DRB5. The quantitative matrices used by ProPred contain the pocket specificity profiles deduced from experimental peptide binding data for various MHC class II alleles.

1.3.3.1.4 MMBPred

The quantitative matrices calculated from known MHC binders in MHCBN (http://www.imtech.res.in/raghava/mhcbn/) database are used for scoring of peptides from the antigenic protein. MMBPred (Bhasin and Raghava, 2003) has been designed for identification of mutations required in a protein sequence for the creation of high affinity binders for the given MHC alleles. It is useful in identification and selection of candidates for knowledge based vaccine design.

1.3.3.1.5 EPIPREDICT

EPIPREDICT (http://www.epipredict.de/) predicts the class II MHC binding peptides for few alleles. It uses the quantitative matrices derived from the combinatorial peptide library experiments which identified the relative preference of each residue at each position of the peptide.

1.3.3.1.6 RANKPEP

RANKPEP (Reche *et al.*, 2002; Reche *et al.*, 2004) is another prediction program which uses the information on the known MHC allele specific binding peptides for predicting of class I and class II MHC binders. RANKPEP algorithm uses position specific scoring matrices calculated from known MHC binding peptides available in the public databases, for evaluating the binding scores of the peptides. It also combines the predicted scores for proteasomal cleavage site with these binding scores.

1.3.3.1.7 SYFPEITHI

SYFPEITHI (Rammensee *et al.*, 1999; Schuler *et al.*, 2007) is a widely used program for the prediction of T- cell epitopes for a number of class I and class II MHC alleles. It maintains a database of the experimentally verified MHC binding peptides and the confirmed epitopes for a wide range of alleles. An analysis of sequences of these peptides have revealed the anchor residues, unusual residues, auxiliary anchor residues, and preferred residues at certain positions in the peptides. The score for the putative peptides is calculated by assigning weights for the most frequently occurring residues at these positions. Negative scores are given to unfavorable residues at these positions. The peptides with a score above a certain cutoff value are predicted as the potential binders.

1.3.3.1.8 MHCPred

MHCPred (Guan *et al.*, 2003a; Guan *et al.*, 2003b; Hattotuwagama *et al.*, 2004) uses quantitative structure activity relationship (QSAR) models derived from experimental peptide binding data for making the predictions. The server calculates the score of the query peptides in terms of their IC_{50} values. The peptides having a score less than the threshold score are predicted as the binders. The server also predicts the high affinity heteroclitic peptides differing only at certain positions from a given peptide.

1.3.3.1.9 nHLAPred

nHLAPred (Bhasin and Raghava, 2007) is an artificial neural network based prediction server which predicts epitopes for about 30 class I MHC alleles. The binder and nonbinder peptides catalogued in MHCBN database are used for the training of allele specific ANN models. The predicted peptides are refined by filtering through the proteasomal cleavage matrices.

1.3.3.1.10 NetMHCpan

NetMHCpan (Nielsen *et al.*, 2007) also uses ANN models trained on a large set of peptides having experimental binding data for more than 80 alleles. In addition to various alleles belonging to human class I and class II MHC, NetMHCpan can also predict antigenic peptides for non human primates, mouse and pig alleles. The binding affinities of the peptides are predicted as their IC_{50} values.

1.3.3.1.11 NetMHC

Like NetMHCpan, NetMHC (Buus *et al.*, 2003; Nielsen *et al.*, 2003; Nielsen *et al.*, 2004) also uses ANN models for prediction of the MHC binding peptides. But in addition to ANN, it also uses PSSM matrices and HMM profiles generated from the MHC binding peptides catalogued in SYFPEITHI database.

1.3.3.1.12 KISS and SVMHC

KISS (Jacob and Vert, 2008) and SVMHC (Donnes and Elofsson, 2002) are webservers which predict the MHC binding peptides based on support vector machine models trained on data from MHCBN (http://www.imtech.res.in/raghava/mhcbn/), SYFPEITHI (Rammensee *et al.*, 1999; Schuler *et al.*, 2007), IEDB (http://www.immuneepitope.org/home.do), and HIV Molecular Immunology Database (www.hiv.lanl.gov). While KISS predicts only for class I MHC alleles, SVMHC also predicts for class II MHC alleles using the quantitative matrices developed by Sturniolo *et al.* (1999).

1.3.3.1.13 PREDEP

In contrast to the sequence based approach used by programs and webservers described before, PREDEP (Schueler-Furman *et al.*, 2000) uses a structure based approach for epitope prediction. It predicts the MHC binding peptides by threading them over the backbone conformation of the bound peptide in the available crystal structure of MHC-peptide complexes. The score is calculated by scoring the contacts the query peptide makes with the MHC molecules using residue based statistical pair potentials. The high scoring peptides are classified as potential MHC binding peptides. PREDEP has been successful in predicting the epitopes for several class I MHC alleles whose crystal structures are available (Altuvia *et al.*, 1995; Altuvia *et al.*, 1997).

1.3.4 Docking and molecular dynamics studies

Most of the computational methods for predicting MHC binding peptides rely heavily on the availability of experimental data on known binders, and hence can only predict for those alleles for which such data is available. Structure based methods can, in principle, predict binding peptides for any MHC allele, however their use in limited by their complexity and high computational cost. Although PREDEP uses a novel threading based approach, it has only been tested on a limited number of MHC alleles. Other structure based computational approaches such as molecular modeling, docking and molecular dynamics simulations are powerful techniques, but only a limited number of studies have been carried out because of the computational complexity. Tong *et al.* (2004) have explored the potential of the docking based methods for the identification for the binding conformations of the peptides in the MHC binding groove of the MHC alleles. Using a three step approach involving the docking of anchor residues in the binding pockets, modeling the intervening backbone structure subject to the spatial constraints of other residues and refinement of the entire backbone, they could successfully predict the structures bound peptides in case of 40 crystal structure complexes. Molecular dynamics studies also have been used to predict conformations of peptides bound to MHC and calculate their binding affinity to discriminate the binders from nonbinders (Pohlmann *et al.*, 2004). However, because of the compute intensive nature of the MD simulations, only a few molecular dynamics studies have been conducted on the MHC-peptide complexes (Pohlmann *et al.*, 2004). The nanosecond scale MD simulation of a antigenic peptide bound in complex with two MHC subtypes HLA-A*2705 and HLA-A*2709 has successfully explained why HLA-A*2705 is associated with disease, while the other is not (Pohlmann *et al.*, 2004).

1.4 ROTAMER LIBRARIES

A rotamer or rotational isomer is a unique side chain conformation represented as a set of values for each dihedral angle degree of freedom. Rotamer libraries are a collection of these rotamers for every residue collected from the solved crystal structure of proteins (Dunbrack, 2002). The rotamer libraries are usually derived from the statistical analysis of the clustering of the side chain conformations encountered in the crystal structures. The libraries also contain the frequency, mean, and standard deviation of sampled conformations. Rotamer libraries are of two types: backbone dependent, and backbone independent. Backbone-independent libraries contain the rotamers without any reference to their backbone conformation. They are calculated from all available side chain conformations for every amino acid residue. Backbone dependent libraries have conformations and frequencies for all available backbone conformations as presented by backbone ϕ and ψ angles (Dunbrack, 2002).

Rotamer libraries are used for rapid modeling of the sidechains when the backbone conformation of the protein or peptide is known. These have been used successfully in a number of studies for modeling, redesigning and improving the stability of the proteins. Shifman et al used rotamer library to redesign the protein-binding interface of calmodulin protein to improve the binding affinity towards myosin light chain kinase (Shifman and Mayo, 2002). Mooers et al have reported an automated redesign protocol for repacking the core of phage T4 lysozyme (Mooers *et al.*, 2003). For rapid modeling of side chain residue, Canutescu *et al* have developed a computer program SCWRL which can mutate and model the side chain residues based on backbone dependent rotamer library, over a given backbone conformation (Canutescu *et al.*, 2003).

The problem of epitope prediction for MHC alleles is the problem of recognition of cognate peptides by MHC receptor wherein the side chains of the peptide chain are accommodated in the compatible binding pockets in the MHC molecules. In this thesis, we have used SCWRL (Canutescu *et al.*, 2003) program to model the MHC-peptide and Kinase-peptide complexes by mutating the backbone of the peptide bound in the crystal structural templates.

1.5 RESIDUE BASED STATISTICAL POTENTIALS

Residue based statistical pair potentials are empirical energy scores for all possible residue-residue pairs, which are found to be in contact in a protein structure or protein-protein complex. These potentials are calculated from the count of residues in contact with each other in the available non-redundant crystal structures. The residues are considered as solid spheres centered at their C_β atoms. Every possible residue pairs in contact with each other are counted from the high resolution crystal structures. The contacts among residues are usually defined as those residue pairs which are within 6.5Å cutoff distance. The residue pairs which are proximal in the polypeptide sequence are explicitly left out from the counting, while long range contacts are taken into consideration. The counts are converted to energy contributions by each pair by means of quasi-chemical approximation with an appropriate treatment to account for the effect of chain connectivity. Residue pairs which are found to be in contact more often get a high energy value as compared to those which are found less often. The pair potential is in the form of a 20×20 matrix containing the energy values for all residue pairs possible. The pair potential matrix calculated by Miyazawa and Jernigan (1996) (MJ) is widely used for the threading and calculating the stability of proteins and folds. MJ matrix has been used successfully for predicting the class I MHC allele binding peptides in a threading type approach (Schueler-Furman *et al.*, 2000). MJ matrix tends to give a high emphasis on the hydrophobic amino acids and hence the polar contacts are not sufficiently scored. This problem was resolved by Betancourt and Thirumalai (BT) when they derived a new matrix with threonine as the reference solvent within the MJ scheme (Betancourt and Thirumalai, 1999). BT matrix has been shown to perform equally good for hydrophobic as well as hydrophilic contacts (Betancourt and Thirumalai, 1999). In this thesis, we have used MJ and BT matrices wherever appropriate for calculating the binding scores of peptides in complex with the MHC alleles and the protein kinases.

Chapter Two

Development of a computational tool for modeling of protein-peptide complexes

2.1 INTRODUCTION

Proteins involved in a majority of cellular processes usually perform their function by binding to some target proteins and forming protein-protein complexes. Interactions between two or more proteins often occur over short contiguous stretches of amino acids within one protein. For example, recognition of substrate proteins by various protein kinases during cell signaling events is governed primarily by specific interactions between the kinase and a contiguous peptide stretch containing the phosphorylation site. Several receptors have peptide fragments as ligands e.g. the major histocompatibility complex (MHC) (Pamer and Cresswell, 1998). Thus, understanding molecular details of interactions between proteins and short peptide motifs is essential for dissecting underlying mechanism of several major cellular processes. Among the various proteins which interact specifically with short peptide motifs, protein kinases and MHCs represent two major protein families whose substrate specificities have been extensively studied by various experimental approaches (Pawson, 1994; Songyang et al., 1994; Udaka et al., 2000).

Although a number of computational tools such as NetPhosK (Blom *et al.*, 2004), KinasePhos (Huang *et al.*, 2005), GPS (Xue *et al.*, 2005), SCANSITE (Obenauer *et al.*, 2003), SYFPEITHI (Rammensee *et al.*, 1999) and ProPred (Singh and Raghava, 2001) etc. are available for predicting the putative substrate peptides for protein kinases and MHC proteins, these methods are mostly based on available experimental binding data for a given class of protein kinase or MHC. These tools predict substrate peptides based on identification of the conserved motifs in a set of known peptide substrates and do not use information from the three dimensional structure of the protein-peptide complex. Hence, these sequence based prediction tools do not give information about key residues in kinases and MHCs which control substrate specificity. Information about specificity determining residues (SDR) can help in design of novel peptide ligands. Correct identification of SDRs of a given protein kinase or MHC can help in prediction of substrates for those protein kinases or MHCs for which no peptide binding data is available, as demonstrated successfully in structure based substrate prediction methods like PREDIKIN (Brinkworth *et al.*, 2003) and PREDEP (Schueler-Furman *et al.*, 2000). These studies have demonstrated that structural analysis of interactions in protein-peptide complexes can lead to novel insight into the mode of substrate recognition. Therefore, molecular modeling of peptide-MHC and peptide-kinase interactions have been carried out by several groups using *ab initio* docking (Tong *et al.*, 2004) or MD simulation approach (Pohlmann *et al.*, 2004). However, the compute intensive nature of these calculations have limited such studies to few protein-peptide complexes. Since knowledge based methods are less compute intensive, and have better prediction accuracy, development of suitable knowledge-based tools for modeling protein-peptide complexes would permit quick structural analysis of MHCs and protein kinases with their substrate peptides. A knowledge-based approach has been used recently for developing kinDOCK (Martin *et al.*, 2006), a powerful tool for modeling of ATP analogs into the active site pocket of protein kinases. However, no such user-friendly tool is presently available for knowledge-based modeling of peptides in the binding pockets of MHCs or protein kinases.

Therefore, we have developed MODPROPEP, a web server for structural modeling of peptides of any desired sequences in the active site pockets of kinases/MHCs having known crystal structures or homology models of kinases/MHCs. This chapter gives a brief description of the development of MODPROPEP, various assumptions made in the knowledge-based modeling protocol, various features of MODPROPEP and few examples of its use.

2.2 METHODS

2.2.1 Compilation of crystal structures

The available crystal structures of MHC and protein kinases were downloaded from PDB website at http://www.rcsb.org (Berman *et al.*, 2000). The structures were divided into two groups, i.e. structures in complex with substrate peptide ligand and structures without the bound peptide ligand. These crystal structures were manually examined and chain/residue numbering was appropriately edited if necessary. All the crystal structures were categorized into three major classes, i.e. class I MHC, class II MHC and protein kinases. Each of these three classes was further grouped into various functional families of protein kinases or MHC alleles.

Detailed analysis of these crystal structures indicated, that all the protein kinases shared a conserved structural fold despite their sequence divergence. For example, crystal structures of IR and PHK which share a sequence identity of only 40% can be superposed with a C^{α} RMSD of 1.6Å. Similar conservation of structures were also observed both for class I and class II MHC structures which share a higher degree of sequence identity within themselves. BLAST alignment of large number of protein kinases and MHC proteins available in sequence databases with these crystal structures indicated that, homology models can be obtained for most of these sequences with reasonable accuracy. Comparison of the bound peptide structures indicated that in all these three classes of proteins, the substrate peptides bind at a structurally homologous site on the conserved fold and the bound peptides maintain a more or less similar extended conformation. This suggested the possibility that bound peptides from protein-peptide complexes can be transformed to the protein structures lacking the bound peptide based on optimum superposition of the protein structures. It may be noted that similar assumption has been used successfully in structural modeling studies of protein ligand complexes involving protein kinases (Brinkworth *et al.*, 2003), MHCs (Schueler-Furman *et al.*, 1998; Schueler-Furman *et al.*, 2000) and other enzyme families (Ansari *et al.*, 2004; Trivedi *et al.*, 2005). There are several examples where more than one crystal structure of an allele is found with bound peptides of different length. It is generally assumed that, three residues on each side of the phosphorylation site make significant contact with the protein kinase and are responsible for the specificity of a kinase (Brinkworth *et al.*, 2003). Therefore, bound peptides having more than 7 amino acids were truncated to three amino acids on either side of the phosphorylation site. All these structures were stored in the template library of MODPROPEP.

2.2.2 Modeling of protein-peptide complexes

The current template library of MODPROPEP has protein-peptide complex crystal structures for 16 alleles of class I, 12 alleles of class II MHC proteins and 6 different protein kinase families. Figure 2.1 shows a flowchart depicting various tasks which can be performed using MODPROPEP. For these MHC alleles and protein kinase families, substrate peptide of any desired sequence can be modeled. Modeling of peptide in the binding pocket of MHC or protein kinase is carried out by using the same backbone conformation as in the template complex and the side chain conformations are generated by the program SCWRL (Canutescu *et al.*, 2003), which uses a backbone dependent rotamer library approach. The template library of MODPROPEP has structures for many MHC alleles or kinase families without the bound peptide substrate. For modeling of peptide substrates in complex with any of

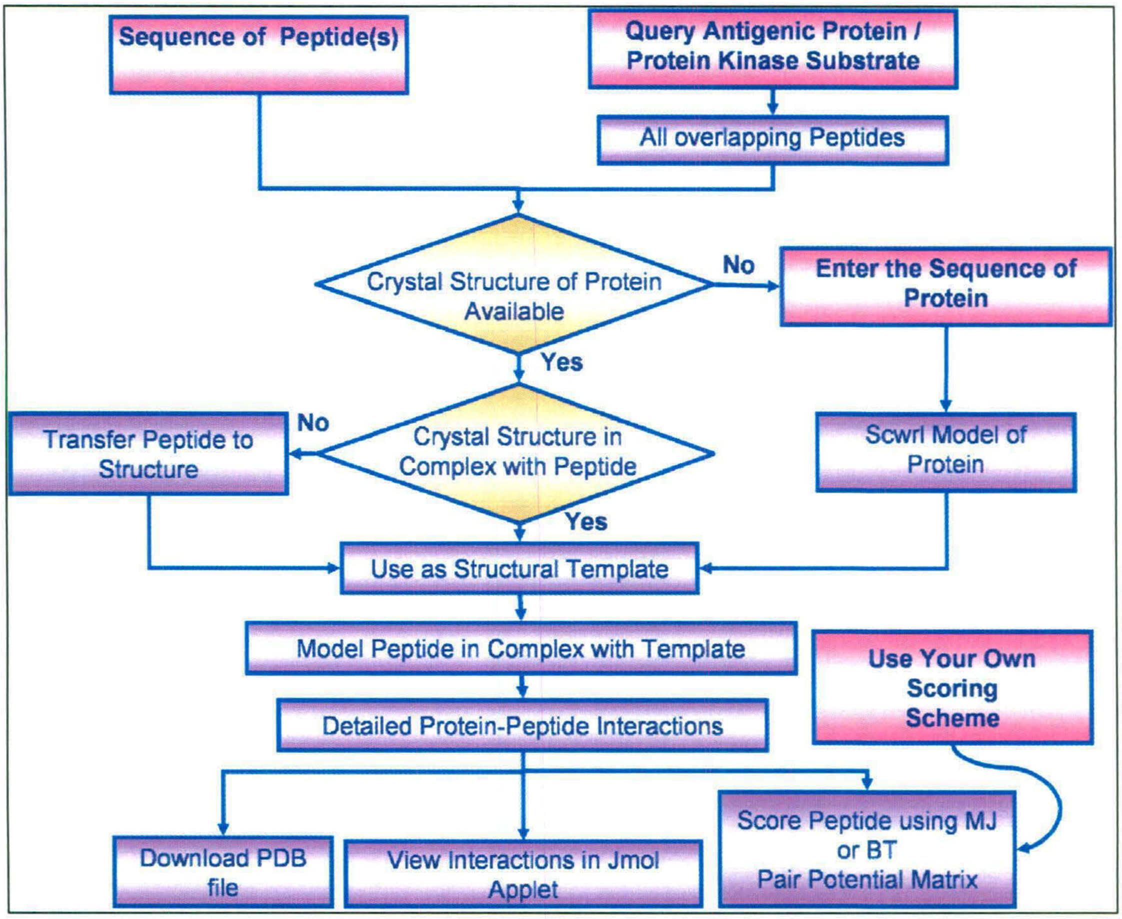

Sequence of Peptide(s)
Query Antigenic Protein / Protein Kinase Substrate
All overlapping Peptides
Crystal Structure of Protein Available
No
Enter the Sequence of Protein
Yes
Transfer Peptide to Structure
No
Crystal Structure in Complex with Peptide
Scwrl Model of Protein
Yes
Use as Structural Template
Model Peptide in Complex with Template
Use Your Own Scoring Scheme
Detailed Protein-Peptide Interactions
Score Peptide using MJ or BT Pair Potential Matrix
Download PDB file
View Interactions in Jmol Applet

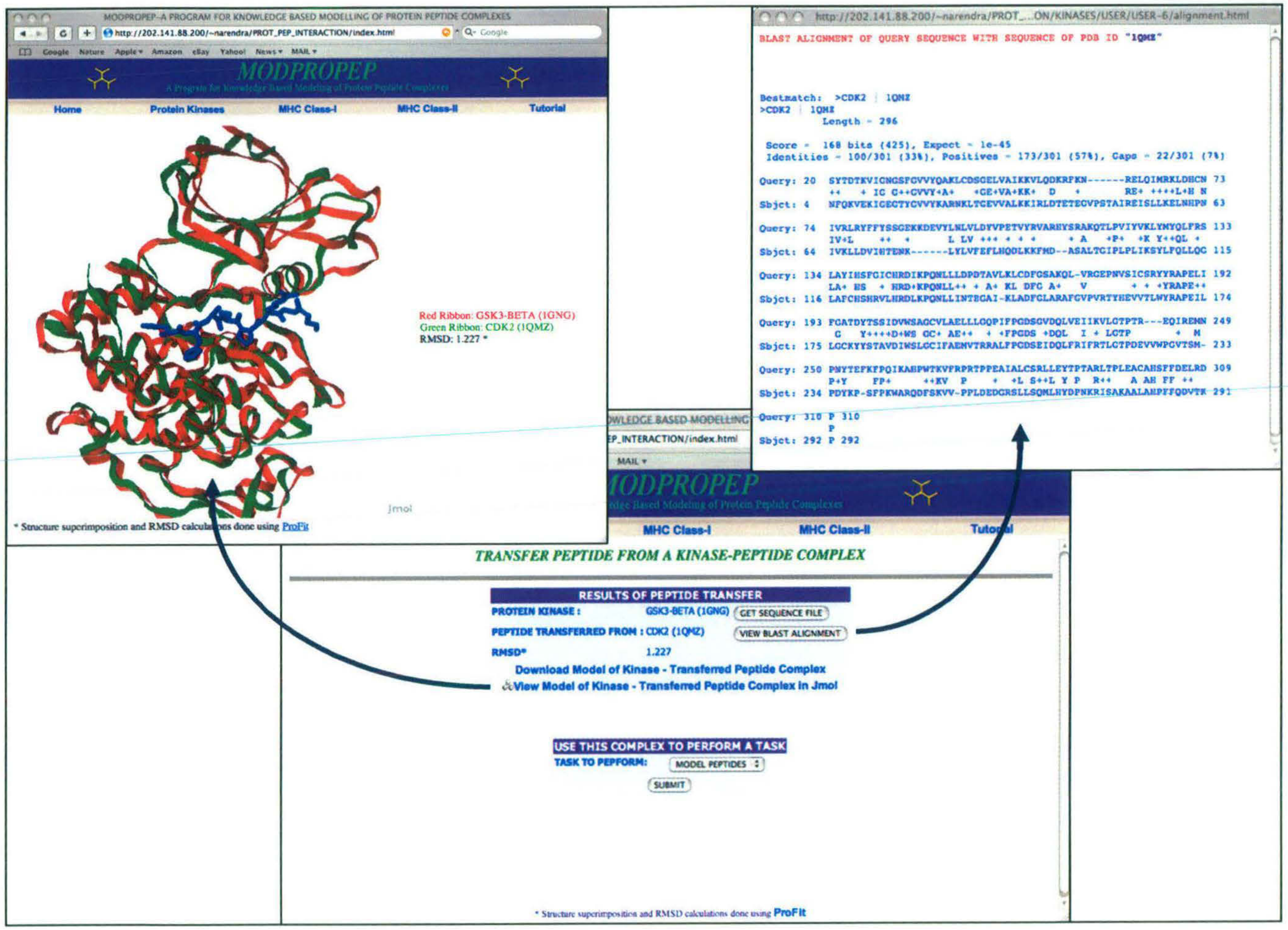
Figure 2.1: A flowchart depicting the organization and features of MODPROPEP. Pink boxes represent the information provided by the user as input.

these MHC alleles or kinase families, peptide conformations are transformed from the available crystal structures of the protein-peptide complexes after optimum superposition of the proteins. If no crystal structures are available for a given protein kinase or MHC protein, the program can model its structure in complex with peptides of desired sequence using the crystal structure of the closest homologous protein-peptide complex. Sequences of various MHC alleles have been obtained from the IMGT/HLA database (Robinson *et al.*, 2003) and stored locally so that the user can select from the list of alleles the protein to be modeled. The crystal structure having maximum sequence similarity is used as a template for modeling the structure of query allele. All sequence alignments are carried out using a local version of the program BLAST. The SCWRL program is used for mutating the residues as per the BLAST alignment and generate the desired homology model. Since only protein-peptide complexes are used to generate the homology models, the backbone of the bound peptide is appropriately mutated by SCWRL to model the substrate of desired sequence. Thus, MODPROPREP provides options for modeling peptide of any desired sequence in complex with any MHC protein or protein kinase.

In order to analyze the interactions between the peptide and the protein, the residues of the MHC or the kinase which are in contact with different side chains of the modeled peptide are identified using a distance based cut off. Based on these contact residues, putative binding pockets are defined for each of the residues in the peptide. MODPROPEP provides a user friendly Jmol java applet interface (http://jmol.sourceforge.net) for visualizing the modeled complexes and analyzing the binding pockets in detail.

Apart from structural modeling of the peptide of a given sequence in the substrate binding pocket of MHC protein or protein kinase, MODPROPEP is also capable of scanning an antigenic protein for potential MHC binding peptides. Similarly, putative substrate proteins for various protein kinases can be scanned for potential phosphorylation sites. Scanning of input sequence is done by breaking the protein sequence into all possible overlapping peptides of a given length. This length is usually the length of the bound peptide present in the template protein-peptide complex, i.e. 9 or 10 mer for class I MHC and longer peptides for class II MHC. However, for protein kinases only heptameric peptides containing Ser/Thr/Tyr as central residue are chosen. For each of these peptides, instead of building all atom side chain conformations, as a first step, contacting residue pairs between peptide and the protein are identified based on C^{β}-C^{β} distances. The binding score of these peptides with the MHC or kinase is evaluated using residue-based statistical energy function by Miyazawa and Jernigan (MJ) (Miyazawa and Jernigan, 1996). It may be noted that a similar scoring scheme has been used earlier for identifying MHC binding peptides using a threading approach (Schueler-Furman *et al.*, 2000). Apart from MJ statistical potential, the program also has options for ranking peptide binding affinities using residue-based statistical energy function by Betancourt and Thirumalai (BT) (Betancourt and Thirumalai, 1999) or other user defined residue based schemes. The peptides are sorted according to their binding score and the user can select some or all of these peptides for detailed side chain modeling by SCWRL depending on their preliminary scores.

Figure 2.2: A snapshot from MODPROPEP showing the result of transfer of bound peptide from the kinase CDK2 (template:1QMZ) to GSK3-beta (template:1GNG). Links are provided for downloading PDB coordinates of the modeled complex and viewing the superposition of the two protein structures along with the peptide in the Jmol applet. A pop-up window shows the BLAST alignment between CDK2 and GSK3-beta.


MODPROPEP–A PROGRAM FOR KNOWLEDGE BASED MODELLING OF PROTEIN PEPTIDE COMPLEXES
http://202.141.88.200/~narendra/PROT_PEP_INTERACTION/index.html
Google
Google Nature Apple ▾ Amazon eBay Yahoo! News ▾ MAIL ▾
MODPROPEP
A Program for Knowledge Based Modeling of Protein Peptide Complexes
Home
Protein Kinases
MHC Class-I
MHC Class-II
Tutorial
Jmol
http://202.141.88.200/~narendra/PROT_...ON/KINASES/USER/USER-6/alignment.html
GET SEQUENCE FILE
VIEW BLAST ALIGNMENT
MODEL PEPTIDES
SUBMIT

Red Ribbon: GSK3-BETA (1GNG)
Green Ribbon: CDK2 (1QMZ)
RMSD: 1.227 *

* Structure superimposition and RMSD calculations done using ProFit

```
BLAST ALIGNMENT OF QUERY SEQUENCE WITH SEQUENCE OF PDB ID "1QMZ"



Bestmatch:  >CDK2 | 1QMZ
>CDK2 | 1QMZ
        Length = 296

 Score =  168 bits (425), Expect = 1e-45
 Identities = 100/301 (33%), Positives = 173/301 (57%), Gaps = 22/301 (7%)

Query: 20   SYTDTKVIGNGSFGVVYQAKLCDSGELVAIKKVLQDKRFKN------RELQIMRKLDHCN 73
            ++    + IC G++GVVY+A+    +GE+VA+KK+  D   +          RE+ ++++L+H N
Sbjct: 4    NFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTETEGVPSTAIREISLLKELNHPN 63

Query: 74   IVRLRYFFYSSGEKKDEVYLNLVLDYVPETVYRVARHYSRAKQTLPVIYVKLYMYQLFRS 133
            IV+L     ++  +        L LV +++ + + +          + A   +P+  +K Y++QL +
Sbjct: 64   IVKLLDVIHTENK------LYLVFEFLHQDLKKFMD--ASALTGIPLPLIKSYLFQLLQG 115

Query: 134 LAYIHSFGICHRDIKPQNLLLDPDTAVLKLCDFGSAKQL-VRGEPNVSICSRYYRAPELI 192
            LA+ HS  + HRD+KPQNLL++ + A+ KL DFG A+    V          + + +YRAPE++
Sbjct: 116 LAFCHSHRVLHRDLKPQNLLINTEGAI-KLADFGLARAFGVPVRTYHEVVTLWYRAPEIL 174

Query: 193 FGATDYTSSIDVWSAGCVLAELLLGQPIFPGDSGVDQLVEIIKVLGTPTR---EQIREMN 249
             G   Y++++D+WS GC+ AE++  + +FPGDS +DQL  I + LGTP         +  M
Sbjct: 175 LGCKYYSTAVDIWSLGCIFAEMVTRRALFPGDSEIDQLFRIFRTLGTPDEVVWPGVTSM- 233

Query: 250 PNYTEFKFPQIKAHPWTKVFRPRTPPEAIALCSRLLEYTPTARLTPLEACAHSFFDELRD 309
            P+Y    FP+      ++KV  P     +  +L S++L Y P  R++   A AH FF ++
Sbjct: 234 PDYKP-SFPKWARQDFSKVV-PPLDEDGRSLLSQMLHYDPNKRISAKAALAHPFFQDVTK 291

Query: 310 P 310
            P
Sbjct: 292 P 292
```

TRANSFER PEPTIDE FROM A KINASE-PEPTIDE COMPLEX

RESULTS OF PEPTIDE TRANSFER

PROTEIN KINASE : GSK3-BETA (1GNG)

PEPTIDE TRANSFERRED FROM : CDK2 (1QMZ)

RMSD* 1.227

Download Model of Kinase - Transferred Peptide Complex

View Model of Kinase - Transferred Peptide Complex in Jmol

USE THIS COMPLEX TO PERFORM A TASK

TASK TO PEPFORM: MODEL PEPTIDES

* Structure superimposition and RMSD calculations done using ProFit

Figure 2.2

2.2.3 Query interface

Currently, the structural library of program contains crystal structures of class I MHC, class II MHC, and protein kinases. The modeling of protein-peptide complexes involving these three classes of protein is possible. User can access the features involving each class by clicking the links on the horizontal bar just below the header graphics.

The program requires user to select a MHC allele or protein kinase from the pull down menu. The program automatically shows the peptide length options available for modeling for that MHC allele or protein kinase. Program takes the user to available crystal structure templates for the selected protein and peptide length. From here the user can decide a task which is either modeling of peptides or scanning a protein sequence for favorable binders. The user is prompted to enter the sequence of peptides as one letter code of amino acids. The program models the peptides in complex with the selected protein which are available for download as files in PDB format. If no ligand bound structure is available for the selected protein, the peptide is modeled by transferring the ligand peptide coordinates from a homologous protein-peptide complex. Figure 2.2 shows an example where a peptide has been modeled in

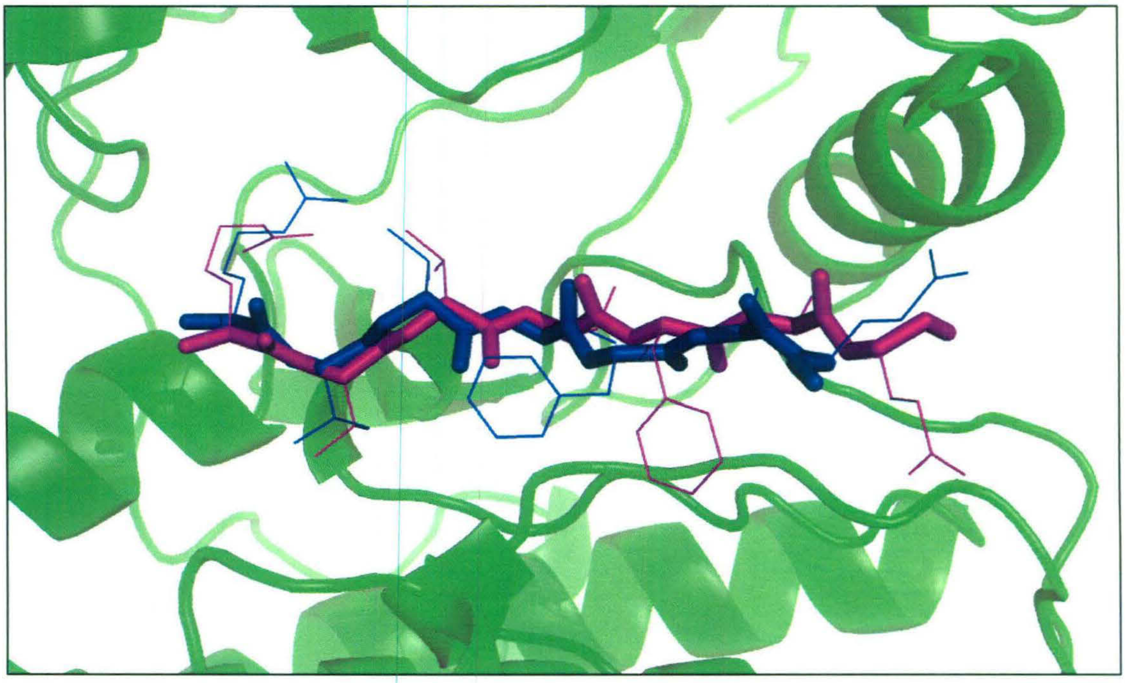


Figure 2.3: Modeling of a peptide in complex with PKB (template:1O6K) by transforming the bound peptide from the crystal structure of PKA-peptide complex (template: 1JBP). Modeled peptide is shown is magenta while original PKA bound peptide is shown in blue for comparison. PKB is shown in cartoon representation in green.

complex with GSK3-beta kinase by transferring the coordinates from CDK2. In order to test the accuracy of this ligand transformation approach, we modeled a peptide in complex with PKB by transforming the bound peptide from PKA. Figure 2.3 shows the superposition of the modeled and the experimentally determined bound peptide in the active site of PKB. As can be seen, backbone of both the peptides superpose quite well with an RMSD of 1.3 Å.

MODPROPEP provides a user-friendly interface to analyze each modeled peptide in detail for contact with the protein. Inter residue contacts can be calculated either based on the distance between C^{β} atoms or based on the distance between any two atoms in a pair of residues. A list of neighboring residues in the protein is displayed for each residue in the peptide. These amino acids on the protein define the binding subsite for each of the peptide residues. Residue pairs having steric clashes are highlighted in yellow. The program also provides interface for analyzing detailed atomic contacts between each pair of residue. Additionally, MODPROPEP uses Jmol applet for the rapid visualization of these subsites in the proteins. Mouse click on a peptide residue shows that residue and the neighbouring residues in the protein in Jmol applet on right hand side. Clicked peptide residue is depicted in ball and stick, while the neighbouring residues are shown in CPK. The protein backbone is shown in ribbon while the peptide backbone is shown in the sticks.

As mentioned earlier, the current version of MODPROPEP permits scoring various bound peptides using residue based statistical scoring matrices given by MJ and BT. Both these scoring matrices have been used in the literature for evaluating binding energy of protein-peptide complexes. It has been reported that, while MJ potential gives better results for binding of peptides involving hydrophobic interfaces, BT potential is more appropriate for binding of peptides involving polar contacts. Here, we discuss a typical example of ranking the site of phosphorylation on the beta-adducin protein (accession no: P35612) by protein kinase A (Matsuoka *et al.*, 1996). Out of a total of 118 S/T containing heptamers, RTPSFLK containing the experimentally identified phosphorylation site S713, is ranked 8 by MJ potential, while scoring by BT matrix gives it a rank of 3. Modeling of this peptide in complex with PKA shows R710 is stabilized by contacts with E127 and E170. Prediction of phosphorylation site in *Limulus* myosin III by PKA (Kempler *et al.*, 2007), and PS1 by GSK3-beta (Prager *et al.*, 2007) indicates that, the true phosphorylation sites identified in recent experiments are ranked as high scoring peptides by MODPROPEP using BT matrix. We have also tested the predictive ability of MODPROPEP on the known substrates of protein kinases cataloged in phospho.ELM database (Diella *et al.*, 2004). We have discussed the results of benchmarking for protein kinases in detail in chapter 3.

Figure 2.4 shows the ranking of a recently identified class I MHC allele HLA-A*0201 ligand by MODPROPEP (Kruger *et al.*, 2005). As can be seen, out of a total of 625 nonameric peptides present in the antigen CABL1_HUMAN (accession no: Q8TDN4), VALEFALHL has a rank of 13 and 26 by MJ and BT potentials respectively. Analysis of inter-molecular contacts indicate that, this peptide is stabilized by interactions involving K66, A150, V152, Y159 and W167. Our results indicate that, in 76% of cases the true phosphorylation site can be ranked within top 30% using BT matrix. Chapter 5 of this thesis discusses the result of benchmarking on experimentally verified MHC binding peptides catalogued in SYFPEITHI database.

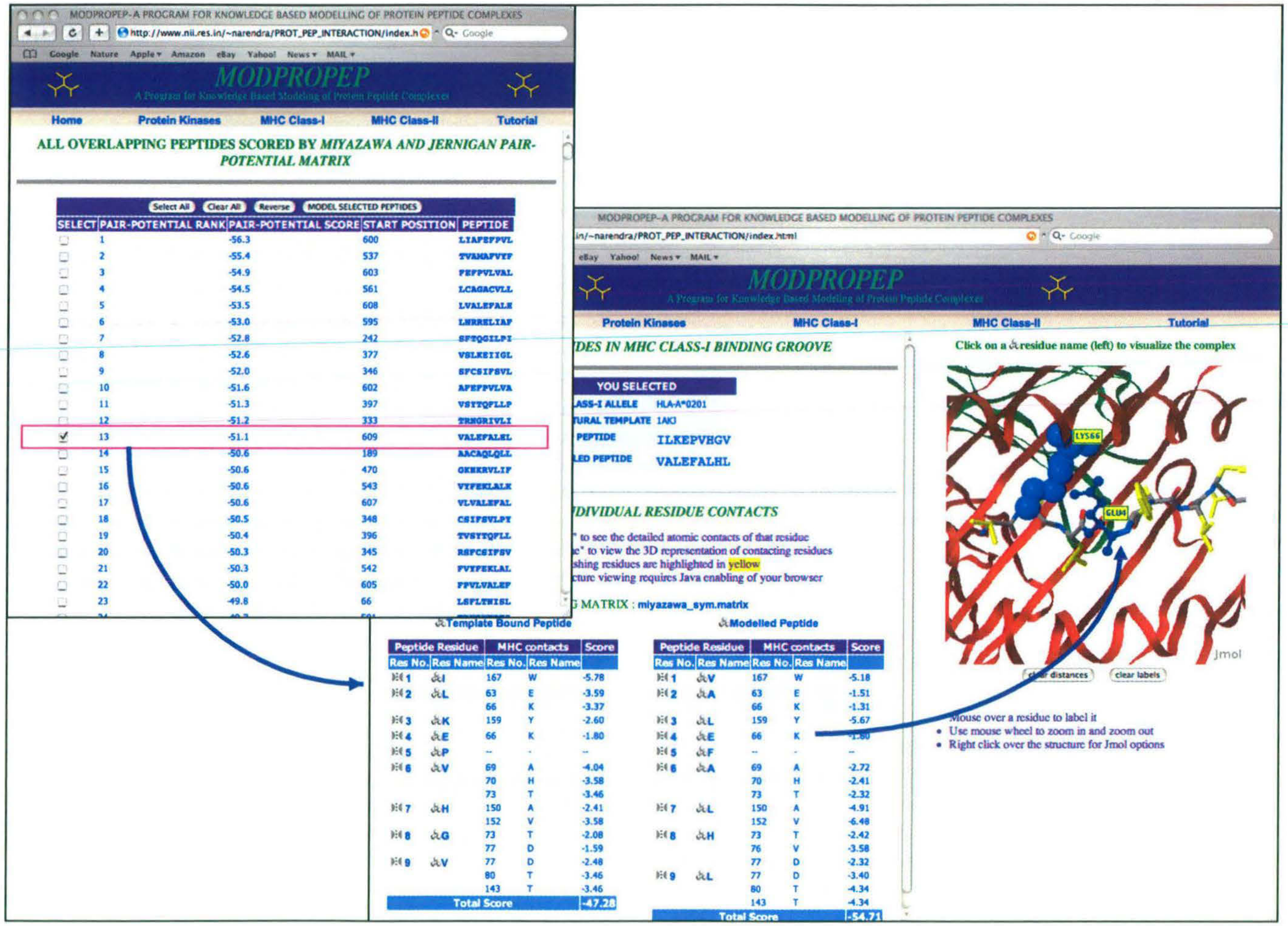
2.2.4 Implementation of the web server

MODPROPEP has been implemented using Perl, CGI scripts, java scripts, Jmol applet and apache web server. BLAST program downloaded from NCBI website is used for local alignments. SCWRL3 is used for the side chain modeling. Various structural superpositions have been carried out using the program ProFit (http://www.bioinf.org.uk/software/profit).

2.3 DISCUSSION

MODPROPEP is a web server for knowledge based modeling of peptide ligands in the active site of various MHCs and protein kinases. The software uses available crystal structures as templates and uses the program SCWRL to mutate the sequence of the protein as well as the peptide to model any peptide-MHC or peptide-kinase complex. It provides a number of user friendly interfaces for visualization and analysis of binding pockets in these protein-peptide complexes. This software has been developed based on the assumption that MHCs and protein kinases have conserved structural fold and the ligand peptides bind essentially at the same site. A major advantage of MODPROPEP over other structural modeling programs is that, it can be used to quickly model a large number of peptides in the binding pockets of MHCs and protein kinases. It can also be used for designing inhibitor peptides which

Figure 2.4: A snapshot from MODPROPEP showing the result of scanning of CABL1_HUMAN protein for HLA-A*0201 restricted antigenic peptides. The experimentally identified substrate peptide VALEFALHL, is chosen for modeling in complex with HLA-A*0201 using 1AKJ as template. The residues of HLA-A*0201 in contact with the peptide residues are depicted in tabular format. The right hand side frame shows the 3D structure of a selected peptide residue and its contacts with HLA-A*0201.



MODPROPEP-A PROGRAM FOR KNOWLEDGE BASED MODELLING OF PROTEIN PEPTIDE COMPLEXES

http://www.nii.res.in/~narendra/PROT_PEP_INTERACTION/index.h

MODPROPEP

A Program for Knowledge Based Modeling of Protein Peptide Complexes

ALL OVERLAPPING PEPTIDES SCORED BY *MIYAZAWA AND JERNIGAN PAIR-POTENTIAL MATRIX*

SELECT	PAIR-POTENTIAL RANK	PAIR-POTENTIAL SCORE	START POSITION	PEPTIDE
□	1	-56.3	600	LIAFEFPVL
□	2	-55.4	537	TVAMAFVYF
□	3	-54.9	603	FEFPVLVAL
□	4	-54.5	561	LCAGACVLL
□	5	-53.5	608	LVALEFALE
□	6	-53.0	595	LNRRELIAF
□	7	-52.8	242	SFTQGILPI
□	8	-52.6	377	VSLKEIIGL
□	9	-52.0	346	SFCSIFSVL
□	10	-51.6	602	AFEFPVLVA
□	11	-51.3	397	VSYTQFLLP
□	12	-51.2	333	TRNGRIVLI
☑	13	-51.1	609	VALEFALEL
□	14	-50.6	189	AACAQLQLL
□	15	-50.6	470	GKNKRVLIF
□	16	-50.6	543	VYFEKLALK
□	17	-50.6	607	VLVALEFAL
□	18	-50.5	348	CSIFSVLPY
□	19	-50.4	396	TVSYTQFLL
□	20	-50.3	345	RSFCSIFSV
□	21	-50.3	542	FVYFEKLAL
□	22	-50.0	605	FPVLVALEF
□	23	-49.8	66	LSFLTNISL

MODPROPEP-A PROGRAM FOR KNOWLEDGE BASED MODELLING OF PROTEIN PEPTIDE COMPLEXES

in/~narendra/PROT_PEP_INTERACTION/index.html

MODPROPEP

A Program for Knowledge Based Modeling of Protein Peptide Complexes

DES IN MHC CLASS-I BINDING GROOVE

YOU SELECTED	
ASS-I ALLELE	HLA-A*0201
URAL TEMPLATE	1AKJ
PEPTIDE	ILKEPVHGV
LED PEPTIDE	VALEFALHL

DIVIDUAL RESIDUE CONTACTS

" to see the detailed atomic contacts of that residue
e" to view the 3D representation of contacting residues
shing residues are highlighted in yellow
cture viewing requires Java enabling of your browser

G MATRIX : miyazawa_sym.matrix

Template Bound Peptide

Peptide Residue		MHC contacts		Score
Res No.	Res Name	Res No.	Res Name	
1	I	167	W	-5.78
2	L	63	E	-3.59
		66	K	-3.37
3	K	159	Y	-2.60
4	E	66	K	-1.80
5	P	--	-	--
6	V	69	A	-4.04
		70	H	-3.58
		73	T	-3.46
7	H	150	A	-2.41
		152	V	-3.58
8	G	73	T	-2.08
		77	D	-1.59
9	V	77	D	-2.48
		80	T	-3.46
		143	T	-3.46
Total Score				-47.28

Modelled Peptide

Peptide Residue		MHC contacts		Score
Res No.	Res Name	Res No.	Res Name	
1	V	167	W	-5.18
2	A	63	E	-1.51
		66	K	-1.31
3	L	159	Y	-5.67
4	E	66	K	-1.80
5	F	--	-	--
6	A	69	A	-2.72
		70	H	-2.41
		73	T	-2.32
7	L	150	A	-4.91
		152	V	-6.48
8	H	73	T	-2.42
		76	V	-3.58
		77	D	-2.32
9	L	77	D	-3.40
		80	T	-4.34
		143	T	-4.34
Total Score				-54.71

Click on a residue name (left) to visualize the complex

- Mouse over a residue to label it
- Use mouse wheel to zoom in and zoom out
- Right click over the structure for Jmol options

Figure 2.4

do not have serine, threonine or tyrosine residue at the phosphorylation site. Identification of kinase inhibitor peptides on the substrates or kinase would aid in understanding the mechanism of regulation of protein kinases. Thus MODPROPEP will complement various available sequence based programs for predicting peptide ligands for MHCs and protein kinases. Using this software the user can identify amino acids on the MHC or kinases, which are crucial for selection of a peptide ligand. Such information is important for design of novel peptide ligands or assigning specificities to new alleles of MHCs or novel families of kinases. This software also has an option for searching the MHC binding peptides in the sequence of an antigen or phosphorylation sites on the substrate protein of a protein kinase using structure based approach. Presently, the binding energy is being accessed using residue based statistical potential. This scoring function is appropriate for quick preliminary ranking of putative peptide ligands. High ranking peptides need to be modeled and detailed interactions with the proteins should be analyzed for prediction of actual binders. The results of benchmarking of MODPROPEP on the experimentally known substrates of protein kinases and MHCs have been discussed in chapter 3 and chapter 5 respectively.

Chapter Three

A structure based multiscale approach for identification of substrates of protein kinases

3.1 INTRODUCTION

Post-translational modification of proteins by phosphorylation is a major regulatory mechanism for a variety of cellular processes such as transcription, translation, replication, signal transduction, immune responses, cell growth, differentiation and apoptosis. The importance of phosphorylation can be estimated by the fact that about 30% of all proteins in the human genome are phosphorylated (Cohen, 2000; Ubersax and Ferrell, 2007). The protein kinases are an important class of enzymes which phosphorylate a number of proteins at specific amino acids. Phosphorylation of proteins by kinases, either results in the change of functional state of the protein, or produces a docking site for the modular interaction domains such as SH2, PTB etc. present in interacting partners proteins (Pawson and Scott, 1997). Phosphorylation event by kinases, followed by interaction through modular domain forms a network of dynamically interacting proteins which is essential for transfer of information as well as proper functioning of cellular processes. Due to their involvement in many essential functions, kinases have also been found to be associated with many diseases including various types of cancers (Pawson, 1994).

In view of the importance of phosphorylation event in various biochemical and cellular processes, correct identification of the substrate protein, that a kinase is likely to phosphorylate, is crucial for understanding the molecular details of various cellular and disease processes. Availability of the complete genomes of many organisms has led to the identification of their kinome or the protein kinase complements of the given organism. Prediction of the function of these kinases in various genomes has been an area of active research. Recent bioinformatics analyses using various sensitive and powerful sequence analysis tools like BLAST, PSI-BLAST and profile HMMs have successfully identified and annotated the whole kinome complements of several eukaryotic genomes. These studies have succeeded in identifying 518 putative kinases in human genome (Manning *et al.*, 2002b), 540 kinases in mouse (Caenepeel *et al.*, 2004), 411 kinase like proteins in caenorhabditis (Plowman *et al.*, 1999) and 353 kinase domains in sea urchins (Bradham *et al.*, 2006). Similarly the number of kinases found in *Dictyostelium* (Goldberg *et al.*, 2006) and *Drosophila* (Morrison *et al.*, 2000) are more than two hundred, while yeast (Hunter and Plowman, 1997) and *Plasmodium* (Ward *et al.*, 2004) seems to have 113 and 65 kinases respectively. These *in silico* studies have also classified these large numbers of kinases in various genomes into various kinase groups and families, thus giving valuable clues about putative signaling pathways in which they could possibly be involved. However, deciphering the specific substrate proteins of these large numbers of kinases still remains a major challenge. Recent studies on reconstruction of phosphorylation networks in yeast and human have demonstrated that, for identifying phosphorylation cascades and deciphering signaling networks at a genomic scale, it is essential to identify the substrate proteins of various kinases (Brinkworth *et al.*, 2006; Linding *et al.*, 2007; Linding *et al.*, 2008). Therefore, it is necessary to develop novel and powerful computational methods for predicting the substrate proteins for a given kinase and precisely identifying the sites of phosphorylation.

In fact a number of different softwares are available for predicting the substrates for kinases, even though there have been relatively fewer efforts to benchmark the prediction accuracy of these programs (Wan *et al.*, 2008). The various prediction programs can be broadly classified into two major groups, namely the sequence based methods and structure based methods. All sequence based methods derive their prediction rules from analysis of the Ser/Thr/Tyr containing sequence stretches which are known to be phosphorylated by various kinases. This information about known substrate peptides for various kinases typically comes from peptide library experiments or from experimental identification of phosphorylation sites on substrate proteins of various kinases(Songyang *et al.*, 1994). Recent studies have attempted to curate the experimental information and compiled comprehensive resources on phosphorylation sites (Amanchy *et al.*, 2007). Even though all sequence based methods use the same experimental information as their knowledge base for formulating prediction rules, different programs employ different methodologies, features and scoring functions for evaluating the phosphorylation potential of putative sites in substrate proteins. NetPhos uses artificial neural network trained on known phosphorylation sites to identify phosphorylation sites in query protein sequences. NetPhos, however, does not provide any information regarding which protein kinase phosphorylates the predicted sites (Blom *et al.*, 1999). On the other hand, the programs, such as NetPhosK (Blom *et al.*, 2004), SCANSITE (Obenauer *et al.*, 2003), GPS (Zhou *et al.*, 2004), and PPSP (Xue *et al.*, 2006) predict Ser/Thr/Tyr containing sites which are likely to be phosphorylated by specific families or groups of kinases. NetPhosK uses the artificial neural networks trained for individual kinase families, and hence can predict substrates of these kinases. SCANSITE uses position specific scoring matrices (PSSM) derived from peptide library data for various kinase families for calculating the scores of phosphorylation probability. GPS and PPSP make use of statistical models based on clustering and bayesian analysis of sequences of phosphorylation sites from known substrates of various kinases. Since for many kinase families, the number of known substrate peptides is relatively fewer, GPS and PPSP group the substrates of similar kinases together to increase the number of known substrate peptides for each kinase group. All these prediction methods are trained on sequence information derived from experimentally characterized phosphorylation sites for various kinases, thus limiting their applicability to only those kinase families for which sufficient amount of substrate information is available from experimental studies. Hence, a major limitation of the various sequence based methods is that, they cannot predict substrates for the other protein kinases for which little or no substrate information is available. Even though recently developed programs like DIPHOS (Iakoucheva *et al.*, 2004) make use of disordered regions and predicted secondary structures as features, and employ logistic regression model for the prediction of phosphorylation sites, their predictive abilities have not been tested extensively.

In contrast to sequence based methods, structure based methods attempt to predict the substrate peptides for kinases based on structural modeling of the putative Ser/Thr/Tyr containing peptides in the peptide binding pocket of the kinase and ranking various peptide ligands as per their interaction energy with the receptor kinase. The crystal structures of various kinase-peptide complexes are used as structural templates for modeling. Thus, unlike sequence based methods which use only sequence information of the putative substrate peptides for prediction, the structure based methods use information from structural features of kinase-peptide complexes apart from sequence information of substrate peptide as well as receptor kinases. Secondly, in principle the structure based methods do not require information about known substrates for a given kinase as preferred substrates are predicted based on physico-chemical interactions between kinase and the peptide. Therefore, structure based methods can in principle be applied for predicting substrates for novel kinase families for which no experimental information is available. Conservation of kinase fold despite divergence in sequence makes them ideal candidates for application of structure based prediction methods. However, the survey of available literature indicates that potential of structure based methods have not been exploited to the full extent. In contrast to large number of sequence based methods for predicting substrates for kinases, as of now, PREDIKIN (Brinkworth *et al.*, 2003) is the only structure-based program available for prediction of phosphorylation sites in proteins. Even though PREDIKIN was a major development in demonstrating for the first time the utility of structural information in successful prediction of kinase substrates, it has not been extensively benchmarked on various kinase families other than PKA, CDK and CHK. Secondly, the scoring scheme it uses for estimating the binding energy of the substrate peptides, uses information from peptide library data. In view of its reliance on experimental substrate peptide data, PREDIKIN will also share some of the disadvantages of sequence based methods so far as its applicability to new kinase families identified in the genomes.

The primary reason for the lack of generalized structure based substrate prediction method is the computational complexity associated with structural modeling of the protein-peptide complexes. Prediction methods based on the evaluation of the binding energy between a protein structure and a given peptide ligand require reliable scoring functions as well as adequate CPU time for efficient conformational search. Therefore, structure based methods like PREDIKIN use structural information to a limited extent and only consider the binding pocket environment based on 3D structure of the kinase, rather than structural modeling of the substrate in the binding site. This prompted us to explore the feasibility of developing alternative structure based approach which does not use any experimental information on known kinase substrates.

Recognition of substrate peptides by protein kinases is essentially a problem of protein–peptide interaction. An analogous problem is the recognition and binding of peptides by MHC proteins. Structure based approaches have been extensively used for predicting binder peptides for MHC proteins (Pohlmann *et al.*, 2004; Tong *et al.*, 2004). One such structure based method which permits fast automated scanning for potential MHC binding peptides used a threading type approach (Schueler-Furman *et al.*, 2000). In this approach, the peptide structure in the MHC groove is used as a template upon which query peptide sequences are threaded. The compatibility between threaded peptide sequence and MHC binding pocket is evaluated by statistical pair potential derived from the analysis of amino acid packing in protein structures. This method has been successful in identifying good binders for several

MHC alleles (Altuvia *et al.*, 1995; Schueler-Furman *et al.*, 1998; Schueler-Furman *et al.*, 2000). Therefore, we wanted to investigate whether residue based statistical potentials can be used for predicting substrate peptides for various kinases.

In this chapter, we have developed a novel structure based method for identification of peptide stretches which can be phosphorylated by various Ser/Thr kinases. We use a multi-scale approach, where putative high scoring substrate peptide candidates are identified by threading of peptides on structural templates of kinase-peptide complexes and scoring them by residue based statistical pair potentials. High scorer peptides short listed by initial screening are modeled in the peptide binding pocket of the kinase using MM/PBSA approach and ranked as per their binding free energy. Benchmarking of our prediction method on the experimental data available in phospho.ELM (Diella *et al.*, 2004) and extensive comparison with other kinase substrate prediction programs indicate that, prediction accuracy of our method is comparable to other sequence based methods even though it does not use any kinase specific experimental data.

3.2 METHODS

3.2.1 Modeling of peptides in the active site pocket of kinases and scoring the binding energy by statistical pair potentials

The crystal structures of kinase-peptide complexes available in PDB (Berman *et al.*, 2000) were used as templates for modeling various query peptides in complex with the respective kinases. In order to model different query peptides, the backbone was kept fixed in the peptide binding pocket of the kinase in the bound conformation and side chains were modeled using backbone dependent rotamer library approach of SCWRL (Canutescu *et al.*, 2003). In cases where no structures were available for kinase-peptide complex, but crystal structures were available for kinase alone, the peptide was modeled by transforming the coordinates of peptide from maximally homologous kinase-peptide complex. If structures were available neither for the kinase-peptide complex nor for kinase alone, kinase of appropriate sequence was modeled by SCWRL using the most homologous kinase structure as template and peptide was modeled subsequently using coordinate transformation approach. All the modeling tasks mentioned above were carried out using MODPROPEP (Kumar and Mohanty, 2007), a software developed in our laboratory for knowledge based modeling of kinase-peptide and MHC-peptide complexes. The contacting residue pairs between the kinase and the peptide were identified using the criteria of any two atoms of the residue pair being at a distance less than or equal to 4.5Å. Based on total number of contacts between the kinase and the peptide, the binding energy was evaluated using Betancourt-Thirumalai (BT) statistical pair potential (Betancourt and Thirumalai, 1999) and all the Ser/Thr containing heptameric peptides in a query protein were ranked as per their binding energy using appropriate interface of MODPROPEP. Since all computations for our structure based method were carried out using MODPROPEP, we will refer our structure based prediction method as MODPROPEP while comparing with other phosphorylation site identification methods.

3.2.2 Dataset for benchmarking prediction accuracy of MODPROPEP

Experimentally identified phosphorylation sites cataloged in Phospho.ELM database were used to compare the prediction accuracy of MODPROPEP with other available softwares for prediction of protein kinase substrates. Phospho.ELM (Diella *et al.*, 2004) (version 5.0, May 2006) dataset contains a total of 13603 phosphorylation instances in 4422 proteins by 263 kinase families. Out of these 263 families, 188 families belonged to Ser/Thr kinases. However, various sequence based methods have grouped many members of these 188 kinase families to single substrate specific classes. Based on the classification scheme proposed by GPS (Xue *et al.*, 2005), 110 out of these 188 Ser/Thr kinases were grouped into 38 classes with number of members in different classes varying from 1 to 12. Since, some of these kinase classes contained too few substrates; we removed those classes which contained less than 20 phosphorylation instances. Kinases which did not show significant homology with structural templates available in MODPROPEP were also excluded. Finally, out of the 38 substrate specific classes, 22 classes containing 70 kinase families were selected for benchmarking of various substrate prediction programs (Table 3.1). They contained a total of 2457 phosphorylation instances in 1180 proteins by 70 kinase families.

Table 3.1: The list of 22 kinase groups that have more than 20 substrates catalogued in phospho.ELM database. The individual kinases which constitute each group and the crystal structure template used by MODPROPEP for prediction during benchmarking are shown. The numbers in the brackets are the percentage similarity of the kinases with their structural template.

S. No	Kinase	Kinase subgroups	Structural template
1	PKA	PKA(100%)	PKA
2	PKB	PKB(100%)	PKA
3	PKG	PKG1 alpha (66%), PKG1 beta (66%), PKG2(64%)	PKA
4	PAK	PAK1(52%), PAK2(52%), PAK3(52%), PAK5(51%)	PHK
5	PDK	PDK1(58%)	PKA
6	ChK	ChK1(57%)	PKA
		ChK2(56%)	PHK
7	CK2	CK2 alpha(42%)	CDK2
8	DAPK	DAPK1(56%), DAPK2(55%), DAPK3(56%)	PHK
9	ROCK	ROCK1(57%), ROCK2(58%)	PKA
10	MAP3K	MAP3K1(49%), MAP3K5(49%), MAP3K8(43%)	PKA
		MAP3K14(45%)	PKB
		MAP3K7(48%), MAP3K11(47%)	CDK2
11	PHK	PHK(100%)	PHK
12	CaMKII	CaMKII alpha (57%)	PHK
13	CDK2	CDK2(100%)	CDK2
14	CK1	CK1 alpha(46%), CK1 delta(42%), CK1 epsilon(43%)	PKA
15	GRK	GRK1(58%), GRK2(56%), GRK3(56%), GRK4(56%), GRK5(57%), GRK6(58%)	PKB
16	GSK3	GSK3 alpha(57%), GSK3 beta(58%)	CDK2
17	IKK	IKK alpha(48%), IKK beta(47%)	PKA
18	MAPK	MAPK1(54%), MAPK3(56%), MAPK4(52%), MAPK6(54%), MAPK7(54%), MAPK8(52%), MAPK9(51%), MAPK10(51%), MAPK11(54%), MAPK12(55%), MAPK13(56%), MAPK14(54%)	CDK2
19	MAP2K	MAP2K1(52%), MAP2K2(54%), MAP2K6(48%)	CDK2
		MAP2K3(44%)	PKB
		MAP2K4(48%), MAP2K7(46%)	PHK
20	CDK1	CDK1(79%)	CDK2
21	PLK	PLK1(49%), PLK3(48%)	PKB
22	PKC	PKC alpha(68%), PKC beta(68%), PKC delta(66%), PKC epsilon(70%), PKC eta(70%), PKC gamma(66%), PKC iota(67%), PKC theta(66%), PKC zeta (66%)	PKB

3.2.3 Benchmarking of MODPROPEP, GPS, PPSP, SCANSITE, NetPhosK and PREDIKIN

MODPROPEP scans a given substrate protein and identifies all heptameric peptides containing Ser/Thr (three amino acids on either side of S/T) as the central residue. Each of these Ser/Thr containing heptamers are modeled in the substrate binding pocket of the respective kinase crystal structure or structural model. The binding energy for each of the kinase-peptide complexes are evaluated using BT statistical pair potential. All modeled peptides are sorted as per their binding energy and assigned a rank. We consider a prediction to be correct if the peptide containing the known phosphorylation site has a rank within top 30%. This procedure is repeated for all the known substrate proteins for a given kinase group and percentage of substrates for which MODPROPEP gives correct prediction is evaluated. Similarly, prediction accuracy was calculated for all the 22 kinase groups.

GPS, PPSP and SCANSITE take the protein sequence of the substrate protein as input and report the potential phosphorylation sites as output. GPS predict the phosphorylation sites by finding the similarity between the Ser/Thr containing peptides in the query sequence and the peptides in training sets consisting of known phosphorylation sites. PPSP uses Bayesian statistics and SCANSITE predicts using the scoring matrices derived from the alignment of phosphorylated peptides identified by peptide library experiments. Each of these programs predicts a set of potential phosphorylation sites for each substrate protein and the actual number of sites predicted by a given program depends on the threshold values of various parameters or the choice of the stringency level. For GPS, the default threshold values were selected. PPSP provides three prediction options: high sensitivity, balance and high specificity. We selected "balance" for PPSP. SCANSITE has provision for three stringency levels i.e. high, medium and low. For SCANSITE, low stringency level was used in this study. NetPhosK does not permit a selection of stringency level, hence all NetPhosK predictions were carried out using default values. If the known phosphorylation site is not in the list of potential phosphorylation sites listed by the program, then the prediction is considered incorrect. However, if the correct phosphorylation site is listed, but the output list contains more than 30% of all possible Ser/Thr containing peptides, then also the prediction was considered as incorrect. This additional criterion makes the stringency level of all these three programs comparable to that of MODPROPEP.

PREDIKIN uses the amino acid sequence of the kinase as input and predicts a heptameric amino acid motif as putative phosphorylation site. However, it was found that, most often the correct phosphorylation site deviated from the predicted motif at many positions. Since, a relaxed criterion was used for estimating prediction accuracy of other programs, similar flexibilities were also allowed in evaluating predictions by PREDIKIN. The number of positions at which the correct phosphorylation site matched with the motif predicted by PREDIKIN was calculated and a score was assigned to the correct phosphorylation site based on a simple scoring scheme of +1 for match and -1 for mismatch. All other Ser/Thr containing peptides in the substrate protein were also scored in a similar manner. All Ser/Thr containing peptides having scores higher than the correct phosphorylation sites were considered as predictions by PREDIKIN. The prediction for the substrate protein was considered correct, if the number of peptides having the score greater or equal to the score of the phosphorylation site, are within 30% of all Ser/Thr containing peptides in the substrate protein.

3.2.4 Human Kinome Analysis

A total 518 kinase sequences have been identified in human genome (Manning *et al.*, 2002b). The sequences of these kinases were downloaded from the KinBase website (http://www.kinase.com/kinbase/). All the tyrosine kinases were removed from this data set and this resulted in 424 Ser/Thr kinases belonging to 104 kinase families. Each of these 424 kinases were aligned with representative kinase sequences belonging to each of the 22 substrate specific classes considered in this study. A local version of BLAST program from NCBI was used for pairwise alignments. A total of 324 kinases showed statistically significant alignments with length more than 250 amino acids. Out of these 324 kinases, a set of 160 kinases had percentage identity of more than 40% with representative members of our 22 substrate specific groups. Hence, substrate specific classes could be assigned with high confidence to each of these 160 human kinases based on the best matching representative sequence.

3.2.5 Re-ranking of high scoring substrate peptides using MM/PBSA

The binding free energy of substrate peptides in complex with the protein kinase was evaluated using MM/PBSA approach. MM/PBSA involves supplementing the conventional molecular mechanics (MM) energy terms with solvation energy terms calculated using poisson-boltzman electrostatic calculations and accessible surface areas of polar and nonpolar atoms. For each substrate protein sequence, the high ranking 30% Ser/Thr containing peptides scored using BT statistical potential by MODPROPEP, were selected and re-ranked as per their MM/PBSA binding free energy. The kinase-peptide complexes were first energy minimized. The MM/PBSA module of AMBER9 molecular dynamics package (Case *et al.*, 2006) was used to calculate the binding free energy of these energy minimized kinase-peptide complexes.

3.2.6 Receiver operating characteristic curves (ROC) calculations

The discriminatory power of our prediction method was calculated using ROC method. ROC is generally favored over the percent accuracy because it does not depend on the cutoff value chosen for the classification of prediction outcomes. The outcome of prediction event for a peptide may be classified into four categories: true positive (TP), the binding peptides which are predicted to be binders; true negative (TN), nonbinding peptides predicted not to bind; false positive (FP), nonbinding peptide predicted to bind; false negative (FN), binding peptide predicted not to bind. ROC is a plot of true positive rate (TPR) vs. false positive rate (FPR) as the cutoff value is varied from minimum to maximum. True positive rate (TPR) is calculated as the fraction of true positive among all positives (TPR = TP / (TP + FN)). TPR, also known as sensitivity, is the fraction of binding peptides among predicted binding peptides. FPR is calculated as the fraction of false positives among all negatives (FPR = FP / (FP + TN)). Specificity is another quantity calculated as the fraction of true negative among all negatives (specificity = TN / (TN+FP)). The area under the ROC curve in the plot is indicative of the predictive power of the prediction method. For each of the prediction exercise during the benchmarking, ROC function of R statistical language environment (http://www.r-project.org) was used for calculation of ROC, specificity and sensitivity values.

3.3 RESULTS

3.3.1 Identification of phosphorylation sites by modeling kinase-peptide complexes

The crystal structures of various Ser/Thr kinases in complex with substrate peptides were obtained from PDB. Structural alignment of the crystal structures of serine/threonine kinases PKA (Madhusudan *et al.*, 1994), PKB (Yang *et al.*, 2002), CDK2 (Brown *et al.*, 1999), and PHK (Lowe *et al.*, 1997) revealed that they share a conserved fold in spite of the sequence divergence, with RMSD values ranging from 1.0 to 1.5 Å. This implies that the homology models of the kinase sequences can be built with high confidence level. Also, the substrate peptide is bound in the catalytic site in more or less similar conformation with a maximum C^{α} RMSD of 1.3 Å among PKA, PKB, and PHK. The only exception was CDK2, which had the substrate peptide in a different conformation. A careful analysis of these crystal structures indicated that, protein kinases homologous to these crystal structures are likely to adopt similar structural folds and conserve their peptide binding pockets.

Since the substrate peptide is bound in the similar extended conformation and relative orientation in most of the crystal structures of kinase-peptide complexes, the peptide binding pocket on the kinase fold has been described in literature in terms of subsites (Kobe *et al.*, 2005) corresponding to binding pockets for each residue of the substrate peptide (Figure 3.1). The site of phosphorylation on the substrate peptide is referred as P0, while the three residues flanking the phosphorylation site on the N- and C-terminus are referred as P-3, P-2, P-1 and P+1, P+2 and P+3 respectively (Kobe *et al.*, 2005). The subsites in the protein kinase which accommodate these seven residues of the substrate peptide are referred as S-3, S-2, S-1, S0, S+1 etc. (Kobe *et al.*, 2005). The specificity of the substrate peptide for a given protein kinase is determined by the complementarities between the peptide residues and the residues lining these subsites. Therefore, theoretical estimation of the binding energy between the peptide and the kinase can be obtained by appropriately scoring the interactions between the amino acids of the kinase and the substrate peptide.

We have previously developed a computational protocol, named MODPROPEP, which can model all possible Ser/Thr containing peptides from a putative substrate protein in the binding pocket of the corresponding kinase. If the crystal structure is not available for a protein kinase, a homology model is built using the crystal structure showing the highest similarity as structural template. For each amino acid of the substrate peptide, the list of contacting residues in each subsite is obtained based on a distance based cutoff. If any atom of a residue in the peptide is within 4.5 Å of any atom belonging to a residue in the kinase structure or structural model, the corresponding residue pairs are defined as contacting residue pairs by MODPROPEP. Binding energy for a query peptide is calculated by scoring the

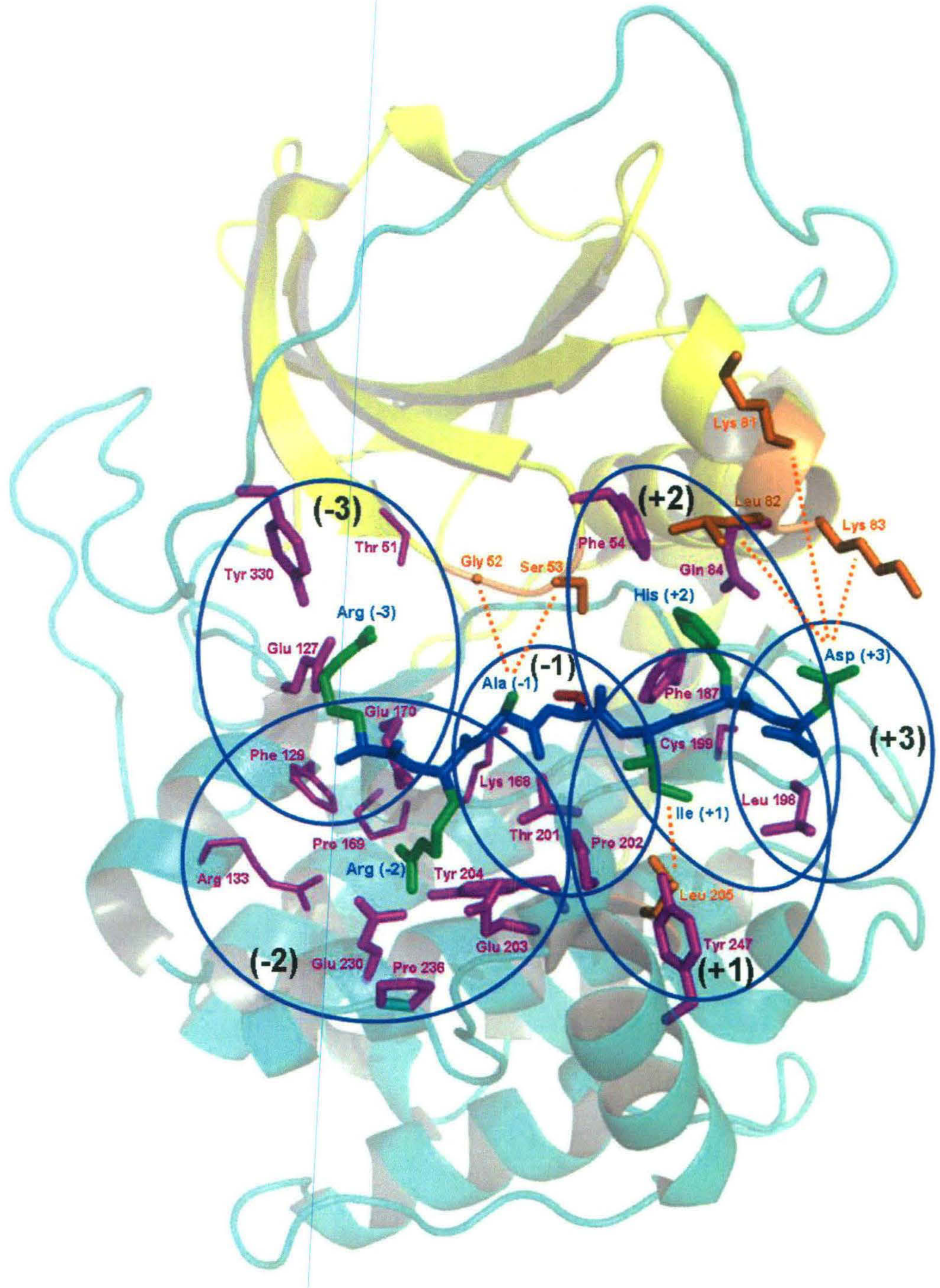


Figure 3.1: Crystal structure of PKA (1JBP) in complex with bound substrate peptide. The small N-terminal lobe and large C-terminal lobes are shown in yellow and cyan ribbons respectively. Binding subsites accommodating the side chain of the peptide residues are shown within the ovals marked -3 to +3. Kinase residues within each subsite which are within a cutoff distance of 4.5Å from the peptide residues they accommodate, are shown in magenta color. Residues which are used in the modified version of the algorithm are shown in the stick representation. Residues represented in orange stick are those which do not come within the distance cut off, but have been reported in the literature to be involved in the determination of substrate specificity. Peptide backbone is shown in blue and the side chains are shown in green.

interactions of each of its amino acids with the residues in the corresponding subsites of the protein kinase, and summing the scores of all amino acids in the peptide. Given a putative substrate protein of a protein kinase, all the Ser/Thr containing peptides are scored for their binding energy values and the peptides are ranked as per their binding affinity.

3.3.2 Benchmarking predictive power of statistical pair potentials

We have used residue-residue statistical pair potentials for scoring the interactions between the residues of the substrate peptide and the residues lining the various subsites of the protein kinase. These knowledge based potentials are derived from the analysis of the packing preference of the various amino acids in the crystal structures present in PDB. Since these pair potentials are derived from analysis of a set of non-redundant structures belonging to various different fold families present in PDB, they are suitable for scoring protein-peptide interactions in general. Therefore, unlike the scoring functions used in sequence based substrate prediction methods for kinases or hybrid methods like PREDIKIN, the scoring functions are independent of the protein kinase families. Therefore, knowledge based potentials can also be used for protein kinase families for which no experimental substrate peptide data is available. We wanted to investigate the predictive ability of the two widely used pair potential matrices viz. Miyazawa and Jernigan (MJ) (Miyazawa and Jernigan, 1996), and Betancourt and Thirumalai (BT) (Betancourt and Thirumalai, 1999). Earlier studies on application of statistical pair potentials for estimating the binding energies of protein peptide complexes have suggested that, MJ matrix is appropriate for interactions involving primarily hydrophobic interfaces, because it has been derived using the solvent as the reference state for the estimation of the favorability of interactions between different amino acid pairs. However, it does not score correctly the interactions involving hydrophilic amino acids. On the other hand, BT matrix overcomes this by changing the reference state to a solvent like molecule threonine. Our analysis on a data set of known kinase substrates also demonstrated that, BT matrix is more suitable for ranking the known kinase substrates with high score. Since the substrate peptides for various kinases often contain charged and polar amino acids in addition to hydrophobic contacts, BT matrix gives better results compared to MJ pair potential. It may be noted that, similar observations have also been made in the context of identification of the MHC binding peptides using statistical pair potential (Altuvia *et al.*, 1995; Schueler-Furman *et al.*, 1998; Schueler-Furman *et al.*, 2000). We discuss below the prediction results obtained for various kinase families using BT matrix.

All Ser/Thr kinase families for which at least 20 different substrate proteins were cataloged in Phospho.ELM database were used to benchmark the predictive power of our structure based approach. MODPROPEP along with BT scoring matrix was used to predict substrates for 22 different kinase groups. As discussed before, all Ser/Thr containing heptameric peptides were modeled in the active site pocket of the respective kinases and were ranked as per their binding energy score. A prediction for a given substrate protein was considered correct, if the actual experimentally identified phosphorylation site was ranked among the top 30% of all the Ser/Thr containing peptides present in the substrate protein. Figure 3.2 shows the results of our structure based prediction method for 10 different kinase families, namely PKA, PKB, PKG, PAK, PDK, CHK, CK2, DAPK, ROCK and MAP3K. As can be seen from Figure 3.2, prediction accuracies of our structure based method for PKA, PKB and PDK are above 70%, while for PKG the prediction accuracy exceeds 80%. For the kinase families PAK, CHK, CK2, DAPK, ROCK and MAP3K, our structure based method could also predict with accuracy higher than 65%. For the purpose of comparison, Figure 3.2 also shows the results from other commonly used programs for prediction of phosphorylation sites using the same dataset, which was used for benchmarking the prediction accuracy of MODPROPEP. Our structure based prediction method outperformed all other programs for PKG, PDK, ChK, CK2, DAPK, ROCK, and MAP3K. For the protein kinases PKA, PKB, and PAK, although MODPROPEP had an accuracy of more than 65%, PPSP did better than MODPROPEP for PAK, while for PKA and PKB, GPS performed better than MODPROPEP. These results were *per se* extremely encouraging, because our structure based approach was comparable in performance to best performing sequence based methods, even though it does not use any experimental data for training.

3.3.3 Improvement of prediction accuracy by alteration of scoring scheme

Figures 3.3 and 3.4 show the prediction results for the other 12 protein kinase families in our data set. As can be seen, for PKC, IKK, PLK, CaMKII, PHK and

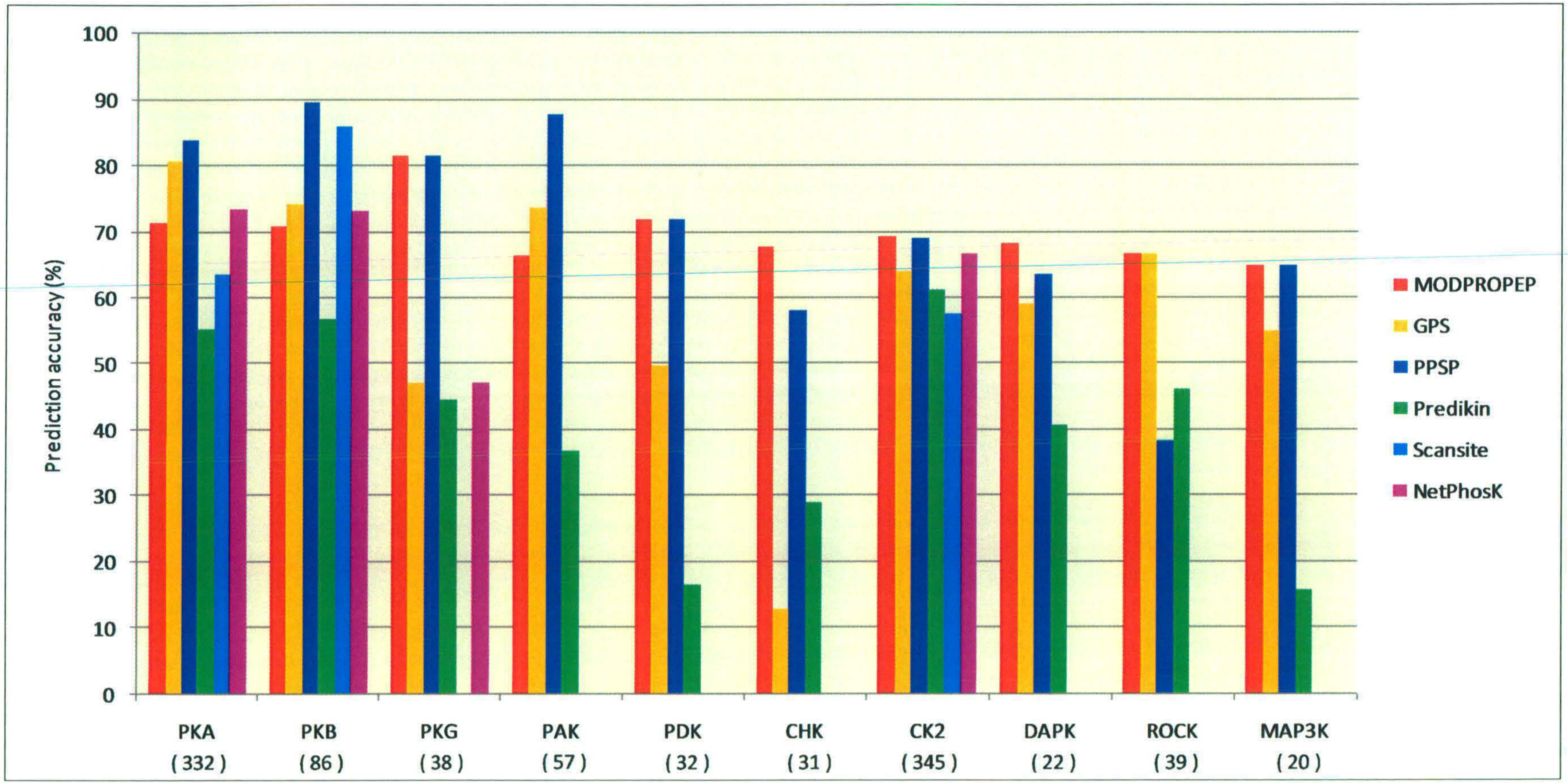


Figure 3.2: Comparison of the prediction accuracies of MODPROPEP for ten different kinase families with other phosphorylation site prediction programs, namely, GPS, PPSP, PREDIKIN, SCANSITE, and NetPhosK. MODPROPEP has a prediction accuracy of more than 60% for these kinases. The total number of known substrate peptides used in prediction is mentioned below the name of respective kinase family.

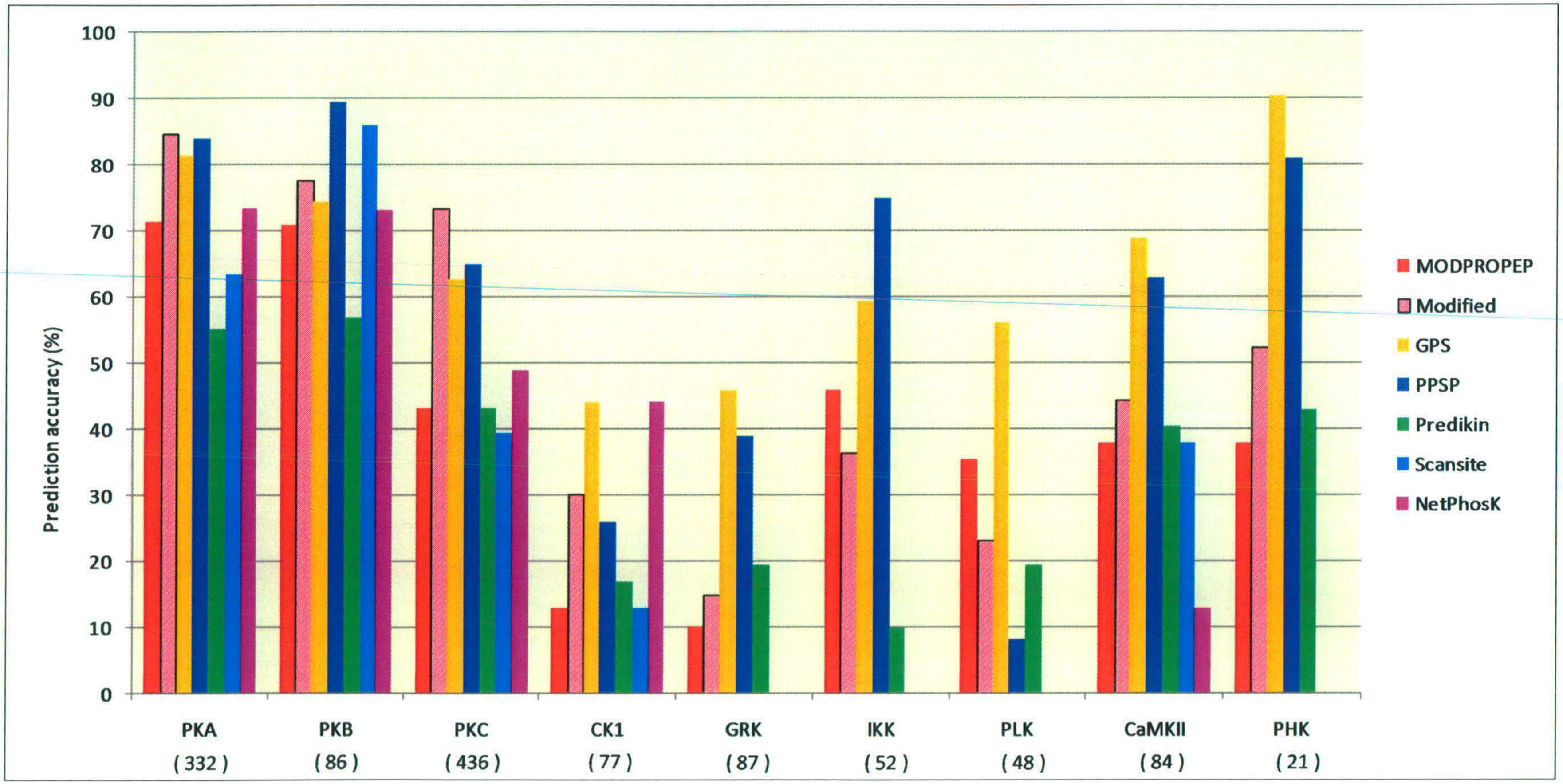


Figure 3.3: Comparison of prediction accuracies by modified version of MODPROPEP for those kinase families where prediction accuracy by original MODPROPEP was less than 60%. The total number of known substrate peptides used in prediction is mentioned below the name of respective kinase family.

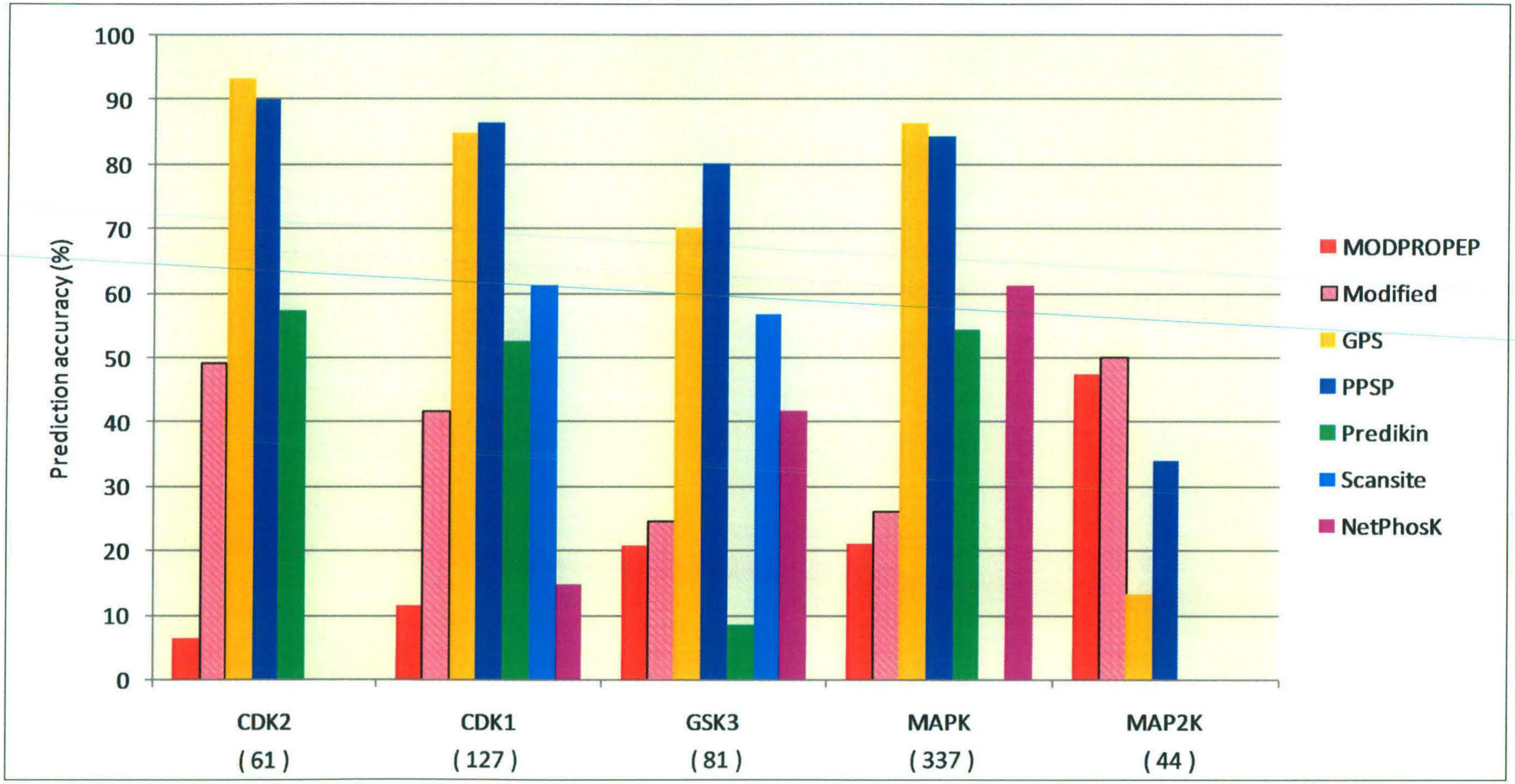


Figure 3.4: Comparison of prediction accuracies by modified version of MODPROPEP for the kinases which were modeled using CDK2 as structural template. The total number of known substrate peptides used in prediction is mentioned below the name of respective kinase family.

MAP2K the prediction accuracy of MODPROPEP was between 35% to 40%, while for the remaining six kinase families MODPROPEP had a prediction accuracy of 20% or lower. It must be noted that, for many of these 12 kinase classes other sequence based prediction tools also had a prediction accuracy of lower than 50%, thus indicating that, they might be genuinely difficult cases for prediction. Therefore, we proceeded to analyze the possible reasons for the failure of MODPROPEP to rank the known phosphorylation sites with high score in case of these kinases. Our modeling protocol had used the crystal structures of PKA and PKB as structural templates in case of five of these kinases, the crystal structures of CDK2 had been used as templates for five other kinase families, while the crystal structure of PHK was used as template in the remaining two cases. The fact that the prediction accuracy for PKA and PKB was good, led us to examine the reasons for the poor prediction accuracy for other kinases families which were homologous to PKA and PKB. We analyzed the crystal structure of PKA for the composition of each of the subsites which accommodate the peptide residues. Analysis of the residues lining each of the subsites indicated that, because of our simplistic distance based cutoff, some residues are included in list of subsites even though they are occluded by other residues and do not make direct contact with the residues of the substrate peptide. Similarly, many residues were included as putative substrate binding pocket residues, even though the interactions were mediated primarily by backbone atoms. Such interactions are unlikely to be determinants of specificity of recognition, but their inclusion as binding pocket residues was resulting in poor scores for actual substrate peptides. For example, in subsite S-1, kinase residues K168, T201, and P202 are included as pocket residues, even though their side chains do not make contact with the side chain of Ala at P-1 (Figure 3.1). On the other hand, residues G52 and S53 have been reported to be the specificity determining residues for the amino acid at P-1 position of the peptide, even though in the crystal structure of PKA they do not have direct contact with the Ala at P-1 position in peptide. Therefore, by careful examination of the crystal structure and inclusion of information from literature about additional specificity determining residues (Nishikawa *et al.*, 1997; Obata *et al.*, 2000), we modified the list of binding pocket residues for each of these 12 kinase families. Predictions were carried out again by MODPROPEP for these 12 kinase families after these modifications. Figures 3.3 and 3.4 also show the results from modified version of MODPROPEP. As can be seen from Figure 3.3, the inclusion of selected residues resulted in further improvement of prediction accuracy for PKA and PKB. Prediction accuracy of PKA improved from 71.4% to 84.8%, while for PKB the improvement in prediction accuracy was from 70.9% to 77.6%. Most dramatic improvement was observed for PKC with prediction accuracy reaching to 73.4% from 43.3%. The accuracy improved only slightly for CK1 and GRK, but for these kinases other programs also did not predict well. IKK and PLK showed a decrease in accuracy. CaMKII and PHK whose template was PHK did not show a significant improvement. GSP and PPSP clearly performed better in case of these kinases (Figure 3.3).

The predictions by MODPROPEP for the kinases CDK1, CDK2, GSK3 beta ar APK had accuracy lower than 20%, while GPS, PPSP, PREDIKIN and S(ITE performed significantly better in these cases (Figure 3.4). Our analysis of th sible reasons for this poor performance of MODPROPEP indicated that, most of ubstrates of all these kinases had Pro at P+1 position. Since, these sequence ba nethods were trained to use Pro as a signature motif, they could predict the substrates of these kinase families with higher accuracy. On the contrary, for MAP2K family which lacked any conserved motif in the substrate peptides, MODPROPEP had a prediction accuracy of 47%, while GPS and PPSP showed accuracy of 12% and 34% respectively (Figure 3.4). These five kinase families were modeled using peptide bound CDK2 as template. We analyzed various modeled peptides in complex with CDK2 based templates to understand, why our method failed to rank known binders with high score. Figure 3.5 shows the interacting residues in various subsites in the crystal structure of CDK2-peptide complex. As can be seen, Arg at P+2 position on the substrate peptide residue is in fact exposed to the solvent and does not make direct contact with any of the kinase residues. Therefore, one would *a priori* expect that substrates having polar or charged residues will be preferred at P+2 position. However, our algorithm does not include any penalty for the hydrophobic residues being exposed to the solvent. Hence, it fails to discriminate peptides having hydrophobic residues at P+2 position. Similarly, the Pro at P+1 position on the substrate peptide has Glu162 and Arg169 as potential interaction partners in S+1 subsite based on distance based cut off. However, careful examination of the orientations of the side chains in S+1 pocket indicates that, Glu162 is oriented away from Pro and Val164 occludes direct contact between Arg169 and Pro. However, our distance based criteria includes these two charged residues as interaction partners for Pro, thus resulting in poor score for peptides containing Pro at P+2. Similarly, the

Lys at P+3 is stabilized by interactions with the phospho-Thr160 residue in the kinase (Figure 3.5). However, our current scoring potential does not contain any score for the interaction involving non standard amino acids. Therefore, while investigating effect of modifications to the scoring scheme and pocket residues, we used the potentials of Asp in place of phospho-Thr for scoring the peptides and excluded the subsites P+1 and P+2 from scoring. As can be seen from Figure 3.4, this resulted in a dramatic improvement in the prediction accuracy for CDK1 and CDK2. The prediction accuracy for CDK2 changed from 7% to 49%, while that of CDK1 the improvement was from 11% to 41%. However, in case of other kinases the improvement was only marginal. This may be because of the involvement of other regulatory mechanisms in determination of the substrate specificity of these kinases. Thus, our detailed analysis of the predictions for these 12 kinase families by MODPROPEP clearly highlighted the possible reasons for the poor performance of the structure based method and it gave valuable clues for improvement in the prediction accuracy by suitable alteration to the computational protocol.

3.3.4 Receiver operating characteristic curves (ROC) analysis

Since our structure based method performed comparable to or better than the best available sequence based methods for 11 kinase families and kinomes of several organisms are known to have many members belonging to these families, we decided to further benchmark the robustness of our predictions for these kinase families by a rigorous analysis involving ROC curves. As discussed earlier, we classified a prediction of phosphorylation site for a given substrate as correct, if the known phosphorylation site was ranked within top 30% of all Ser/Thr containing peptides in terms of the calculated binding energy score. Altering the cut off for this rank percentile will not only alter the overall prediction accuracy, but it will also change the number of false positive predictions. Therefore, detailed analysis of sensitivity and specificity of prediction through ROC curve is necessary to judge the true significance of the predictions by MODPROPEP. We computed the ROC curve for each of the 10 kinase families which showed a prediction accuracy of higher than 65% in our analysis by unmodified MODPROPEP. Figures 3.6A and 3.6B show the ROC curves for the representative case of CK2 and also when all these 10 groups were merged into a single class. As can be seen, the values for area under the curve (AUC) are 0.792 and 0.764 respectively. Table 3.2 gives the values for AUC, sensitivity (Sn)

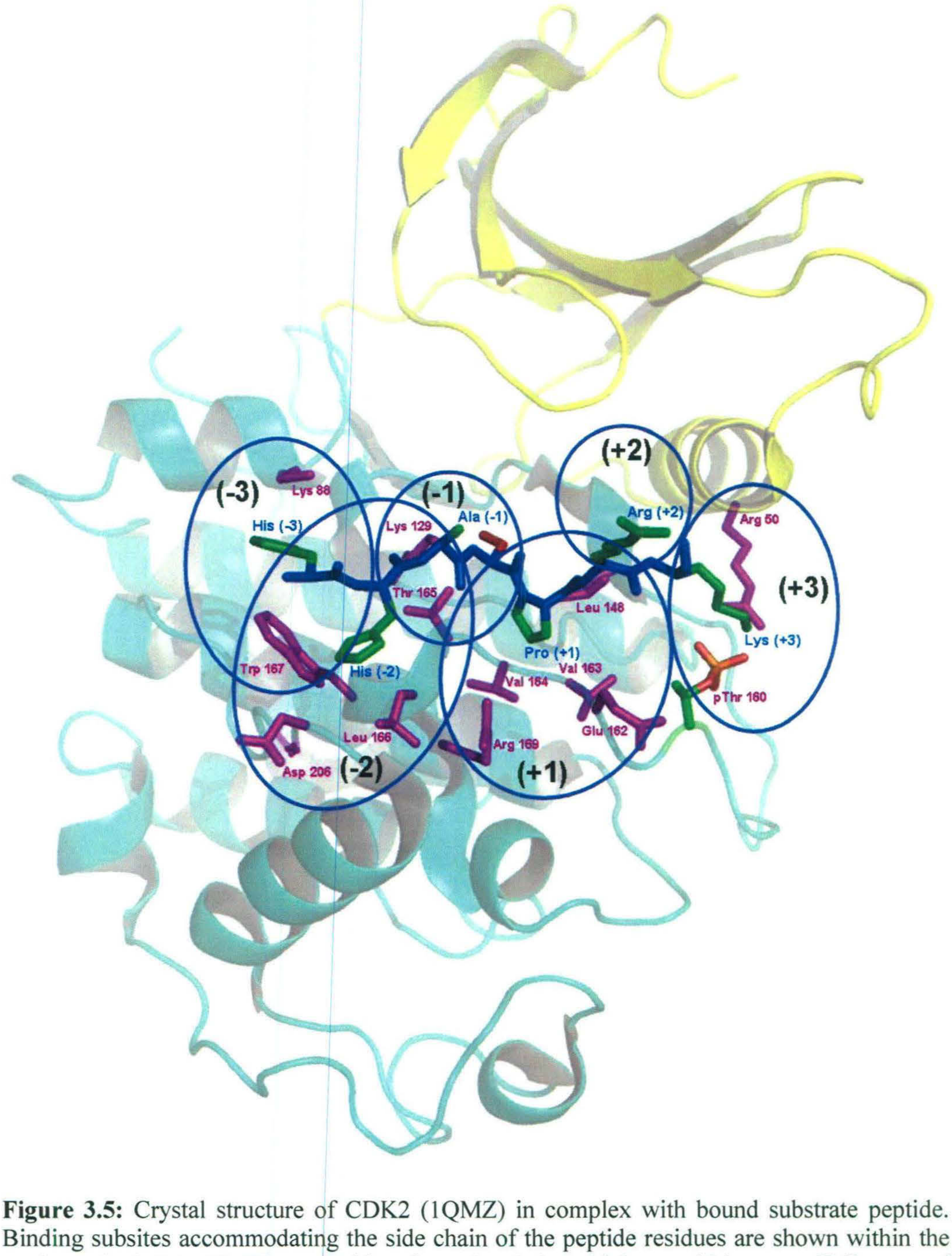


Figure 3.5: Crystal structure of CDK2 (1QMZ) in complex with bound substrate peptide. Binding subsites accommodating the side chain of the peptide residues are shown within the ovals marked -3 to +3. Kinase residues in each subsite which are within a cutoff distance of 4.5Å from the peptide residues they accommodate are shown in magenta color. Peptide backbone has been shown in blue and the side chains of peptide residues are shown in green. The small N-terminal lobe and large C-terminal lobes are shown in yellow and cyan ribbons respectively.

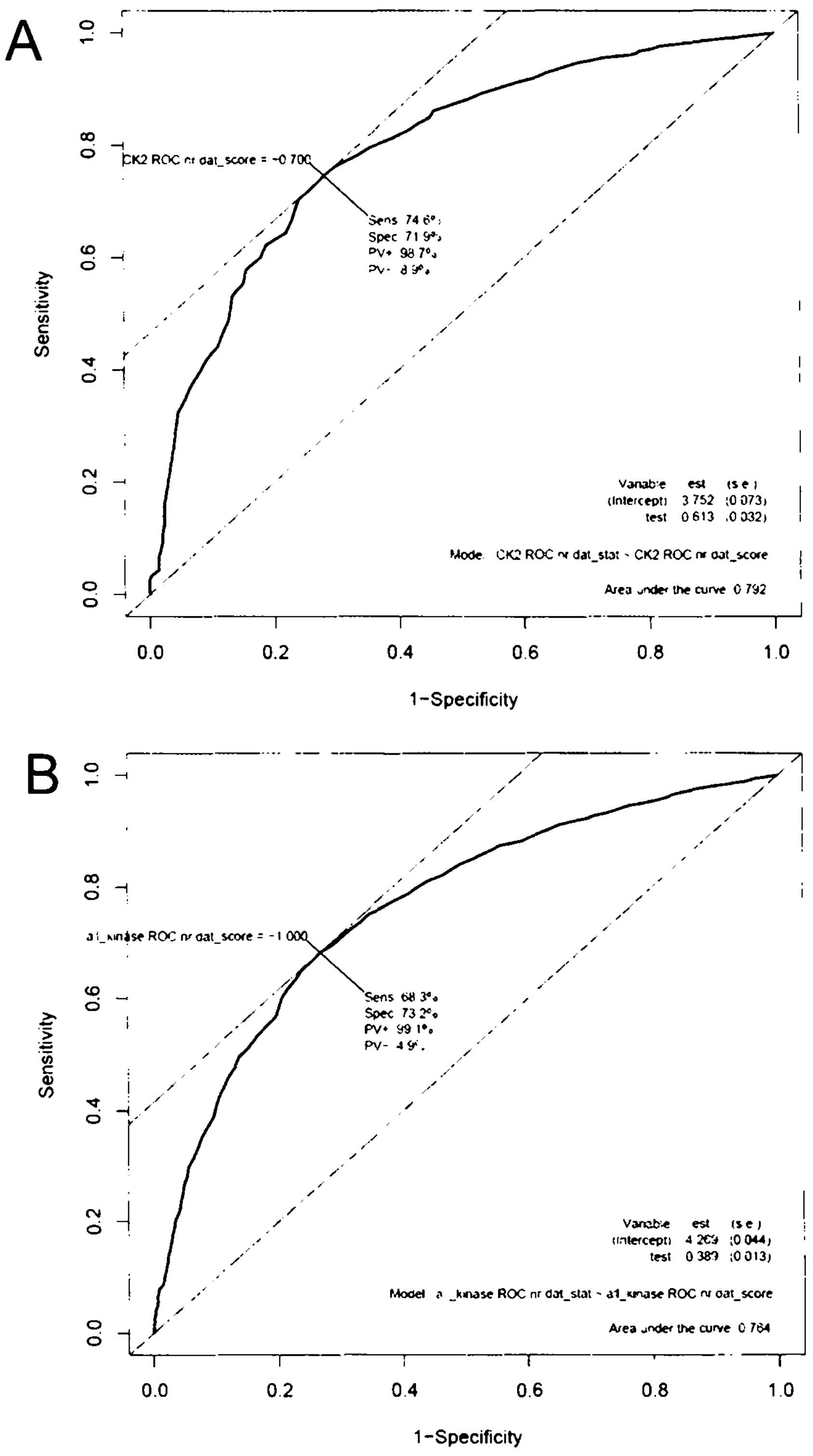


Figure 3.6: Representative ROC curves for **(A)** CK2 and **(B)** all kinases merged together in a single group. The AUC values, sensitivity and specificity values have been shown in the graphs.

Table 3.2: Receiver operating characteristic curve (ROC) analysis for 10 protein kinase groups whose prediction accuracy by MODPROPEP was more than 60%.

Kinase	Area under curve	Sensitivity (%)	Specificity (%)
PKA	0.8	72.6	74.4
PKB	0.794	74	70.6
PKG	0.838	80.7	81.6
PAK	0.736	64.2	75.4
PDK	0.729	76.9	68.8
ChK	0.738	83.6	61.3
CK2	0.792	74.6	71.9
DAPK	0.776	61.1	86.4
ROCK	0.76	88	61.5
MAP3K	0.681	94.3	42.1
All kinases	0.764	68.3	73.2

and specificity (Sp) for the predictions by MODPROPEP for each of these 10 kinase families. These results further establish the statistical significance of our predictions.

3.3.5 Analysis of human kinome

Recently, there has been increased interest in the reconstruction of signaling pathways involving protein phosphorylation networks (Brinkworth *et al.*, 2006; Linding *et al.*, 2007; Linding *et al.*, 2008). These studies require the identification of the substrate protein for the known and putative protein kinases. As different prediction programs predict with different accuracy, correct choice of prediction program for each family of kinase becomes very important. However, it is often observed that, for *in silico* analysis of substrates of kinases only one or two prediction approaches are used irrespective of the kinase family and no attention is paid to the prediction accuracy of those particular programs for the kinase family being analyzed. A major bottleneck in the choice of most appropriate phosphorylation prediction tool for different kinase families is the lack of systematic benchmarking studies on large data sets. Apart from the current study, the only other systematic comparative analysis of the predictive ability of phosphorylation site prediction tools is a recent study by Wan *et al.* (2008). However, analysis by Wan *et al.* (2008) only compared the available sequence based prediction methods. Since our analysis of the comparative prediction accuracy of sequence and structure based prediction methods clearly demonstrated the superior performance of different prediction methods for different kinase families, we decided to identify best prediction tools for various different kinase groups present in human genome. The human kinome consists of 518 functional genes for protein kinases which have been classified into 10 major groups and 133 families. Sequence homology searches were carried out for these kinases, against a library of kinase sequences belonging to the 22 kinase families, which have been analyzed in the current work for prediction accuracy of different programs. These homology searches showed that 160 kinases belonging to 33 families produced an alignment of length more than 250 and sequence identity more than 40% with members of the 22 substrate specific kinase families. These genomic protein kinases are homologous to at least one of the kinases in this library, so the closest homologous kinase in the library can be used to predict substrates of each one of these kinases. Table 3.3 shows the number of different human kinases belonging to each of the 22 kinase groups present in our library. It also lists for each group the prediction methods which perform with an accuracy higher than 60% listed in the decreasing order of their prediction accuracy. As can be seen from Table 3.3, our structure based approach MODPROPEP, predicts best for the 15 out of 33 kinase families. For the remaining families, PPSP and GPS were best predictors. Therefore, during *in silico* exploration of the phosphorylation network, different prediction programs can be selected based on our results in Table 3.3, and the structure based program developed in the current work could be a powerful tool in addition to available sequence based methods. One of the puzzling observations from this analysis is the absence of SCANSITE (Obenauer *et al.*, 2003) in Table 3.3, even though it is a widely used kinase substrate prediction tools. It must be noted that, SCANSITE is primarily trained on peptide library data, while our benchmarking was carried out on data from in vivo phosphorylations of proteins. It appears that, there are distinct differences in the phosphorylation patterns seen in peptide library data verses in *vivo* phosphorylation data.

3.3.6 Re-ranking of kinase-peptide complexes using MM/PBSA

Our analysis demonstrated that, MODPROPEP is a powerful structure based approach which successfully predicts substrates for ten different kinase families, even though it does not use any kinase family specific substrate data for training. Moreover, because of the scoring by simple statistical pair potential it is not compute intensive and can be used for high throughput analysis of sequences in a genomic

Table 3.3: The protein kinases from human genome belonging to 33 families which show similarity to 22 kinase groups present in our dataset. For each group the prediction methods which performed with accuracy higher than 60% are also listed in the decreasing order of their prediction accuracy.

Group	Family	Genomic kinases	Template	Programs recommended
AGC	Akt	3	PKB	PPSP, Scansite, GPS, NetPhosK, MODPROPEP
	DMPK	2	ROCK	MODPROPEP, GPS
	GRK	7	GRK	GPS
	PDK1	1	PDK	MODPROPEP, PPSP
	PKA	5	PKA	PPSP, GPS, NetPhosK, MODPROPEP, Scansite
	PKC	9	PKC	MODPROPEP, PPSP, GPS
	PKG	2	PKG	MODPROPEP, PPSP
	PKN	3	PKC	MODPROPEP, PPSP, GPS
	RSK	8	PKB	PPSP, Scansite, GPS, NetPhosK, MODPROPEP
	SGK	3	PKB	PPSP, Scansite, GPS, NetPhosK, MODPROPEP
CAMK	CAMK1	5	CaMKII	GPS, PPSP
	CAMK2	4	CaMKII	GPS, PPSP
	CAMKL	1	ChK	MODPROPEP
	CASK	1	CaMKII	GPS, PPSP
	DAPK	5	DAPK,CaMKII	MODPROPEP, PPSP, GPS
	DCAMKL	2	DAPK,CaMKII	MODPROPEP, PPSP, GPS
	MLCK	3	DAPK	MODPROPEP, PPSP
	PHK	2	PHK	GPS, PPSP
	RAD53	1	ChK	MODPROPEP
CK1	CK1	7	CK1	GPS, NetPhosK
CMGC	CDK	18	CDK2,CDK1	GPS, PPSP
	CDKL	2	CDK2,CDK1	GPS, PPSP
	CK2	2	CK2	MODPROPEP, PPSP, NetPhosK, GPS, Predikin
	GSK	2	GSK3	PPSP, GPS
	MAPK	14	MAPK	GPS, PPSP
	RCK	1	CDK1	PPSP, GPS, Scansite
STE	STE11	7	MAP3K	MODPROPEP, PPSP
	STE20	21	PAK	PPSP, GPS, MODPROPEP
	STE7	6	MAP2K	MODPROPEP
	STE-Unique	2	MAP3K	MODPROPEP, PPSP
Other	IKK	2	IKK	PPSP
TKL	PLK	4	PLK1	GPS
	MLK	5	MAP3K	MODPROPEP, PPSP

scale. However, its utility can be further improved if percentile cut off for bracketing the correct phosphorylation site can be further lowered from 30%. As demonstrated earlier in case of protein structure prediction problems, multiscale modeling strategy can potentially help in improving the rank of the correct substrate. Therefore, scoring by pair potentials can be used as a first level of search, while sites ranked within top 30% can be reranked using all atom forcefield. For the ten kinase families for which MODPROPEP could successfully rank the correct binding site within top 30%, we carried out detailed all atom modeling of all the Ser/Thr containing peptides in the binding site of the respective kinases. Figure 3.7 shows the C^{α} RMSD values for the kinase and the peptide backbone from the respective template structures for all the energy minimized kinase peptide complexes analyzed in case of PKA. As can be seen, in all cases modeled complexes remain close to the template structure. Similar trend was also observed for kinase-peptide complexes belonging to other families. For each kinase-peptide complex, interaction energy between the kinase and the peptide was computed using MM-PBSA approach and all the modeled peptides were re-ranked as per their MM-PBSA binding energy values. Figure 3.8 shows the comparison of the ranking by MM-PBSA and pair potential for all the ten kinase families. As can be seen for PKA, out of the total of 332 substrate proteins, in case of 237 proteins MODPROPEP could rank the true phosphorylation site within top 30%. As per pair potential ranking, out of these 237 cases, in case of 128 the phosphorylation site was within top 10%, in case of additional 66 proteins the phosphorylation site was ranked within 10% to 20%. Thus, for a total of 194 cases the phosphorylation site was ranked within top 20% and in case of 43 proteins the phosphorylation site was ranked within 20% to 30% by MODPROPEP. Interestingly upon re-ranking by MM-PBSA approach the number of true phosphorylation sites within top 10% increased to 188 and the number of true phosphorylation sites within top 20% increased to 217. Thus by re-ranking there is a significant enrichment of true phosphorylation site in top 10% and 20% window. In Figure 3.8, similar results have been plotted for all 10 kinase families and since there were different number of substrate proteins for different families, they have been represented as percentage of the total number of substrate proteins considered for modeling by MM-PBSA. As can be seen, in case of 7 out of ten kinase families re-ranking has helped in increasing the number of cases where true phosphorylation site could be bracketed within top 10%

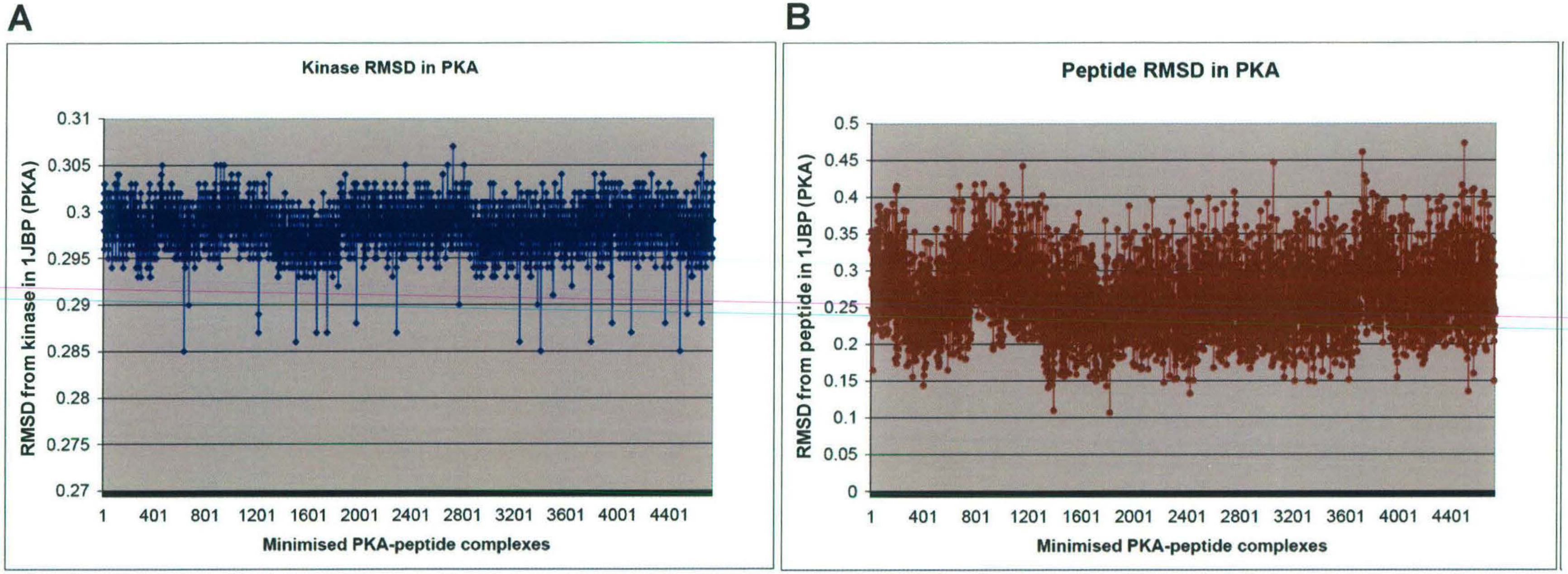


Figure 3.7: Distribution of the Cα RMSD values of **(A)** kinase, and **(B)** modeled peptide from their template structure 1JBP for all energy minimized kinase-peptide complexes in case of PKA.

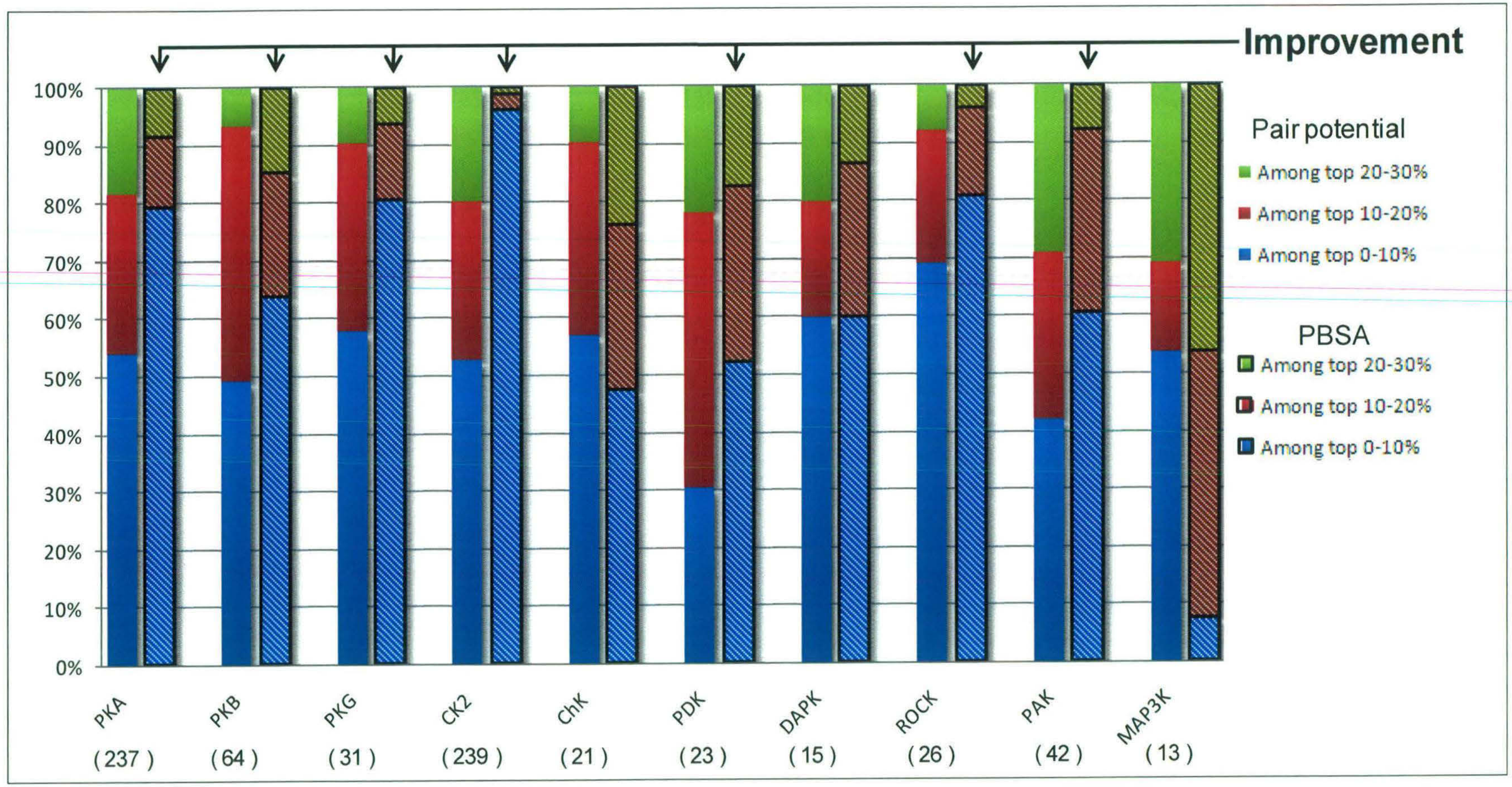


Figure 3.8: Comparison of the ranking of true phosphorylation site among top ranked 30% peptides by pair potential and MM/PBSA method. For each of the ten protein kinase groups for which BT matrix had good prediction accuracy, the distribution of the percentage of cases in which the actual phosphorylation site was among 0-10%, 10-20% and 20-30% of top scoring peptide is shown. The filled bars represent scoring by pair potential, while bars with stripes represent scoring by MM/PBSA methods.

window. Figure 3.9 shows the AUC values obtained from ROC analysis for the ranking using pair potential as well as MM-PBSA. As can be seen, AUC values have increased in all cases except for ChK and MAP3K, thus demonstrating the utility of re-ranking the pair-potential predictions using MM-PBSA approach. Figure 3.6 shows the representative ROC curves for CK2 and all kinases. Thus our results demonstrate that, prediction accuracy of MODPROPEP can be further improved if a multi-scale modeling approach involving re-ranking of pair potential predictions by MM-PBSA energy values is implemented.

3.4 DISCUSSION

In this chapter, we have developed a novel structure based approach for predicting substrates of protein kinases. The putative substrate peptides are modeled in the substrate binding pockets of kinases using the available crystal structures of kinase-peptide complexes as templates. The binding energy of these peptides in complex with the kinase are evaluated using a residue based statistical pair potential derived by Betancourt and Thirumlai. We have carried out detailed benchmarking of this approach on the experimental data available in the Phospho.ELM database and compared our results with those from a number of other phosphorylation site prediction tools. Our results indicate that, the structure based method developed in this work can predict more than 60% of the experimentally identified substrates for 10 protein kinases. The prediction accuracies for PKA, PKB, PKG, and PDK were well above 70% with PKG having the highest prediction accuracy of 81.5%. The other kinase groups for which our approach showed good prediction accuracy were ChK, CK2, DAPK, ROCK, and MAP3K. Our approach also outperformed all other prediction tools for PKG, PDK, ChK, CK2, DAPK, ROCK, and MAP3K. We also carried out receiver operating characteristic (ROC) curve analysis for analyzing the robustness of our structure based prediction approach. The area under curve (AUC) values for these 10 kinases ranged from 0.681 (MAP3K) to 0.838 (PKG). It is encouraging to note that, the prediction accuracy of our method is comparable to other sequence based methods like GPS, PPSP, SCANSITE and NetPhosK, even though it does not use any experimental phosphorylation site data for training unlike sequence based methods. We have also compared our prediction results with PREDIKIN which is the only other structure based approach, but uses a different scoring scheme. Our results clearly demonstrate the better prediction accuracy of our

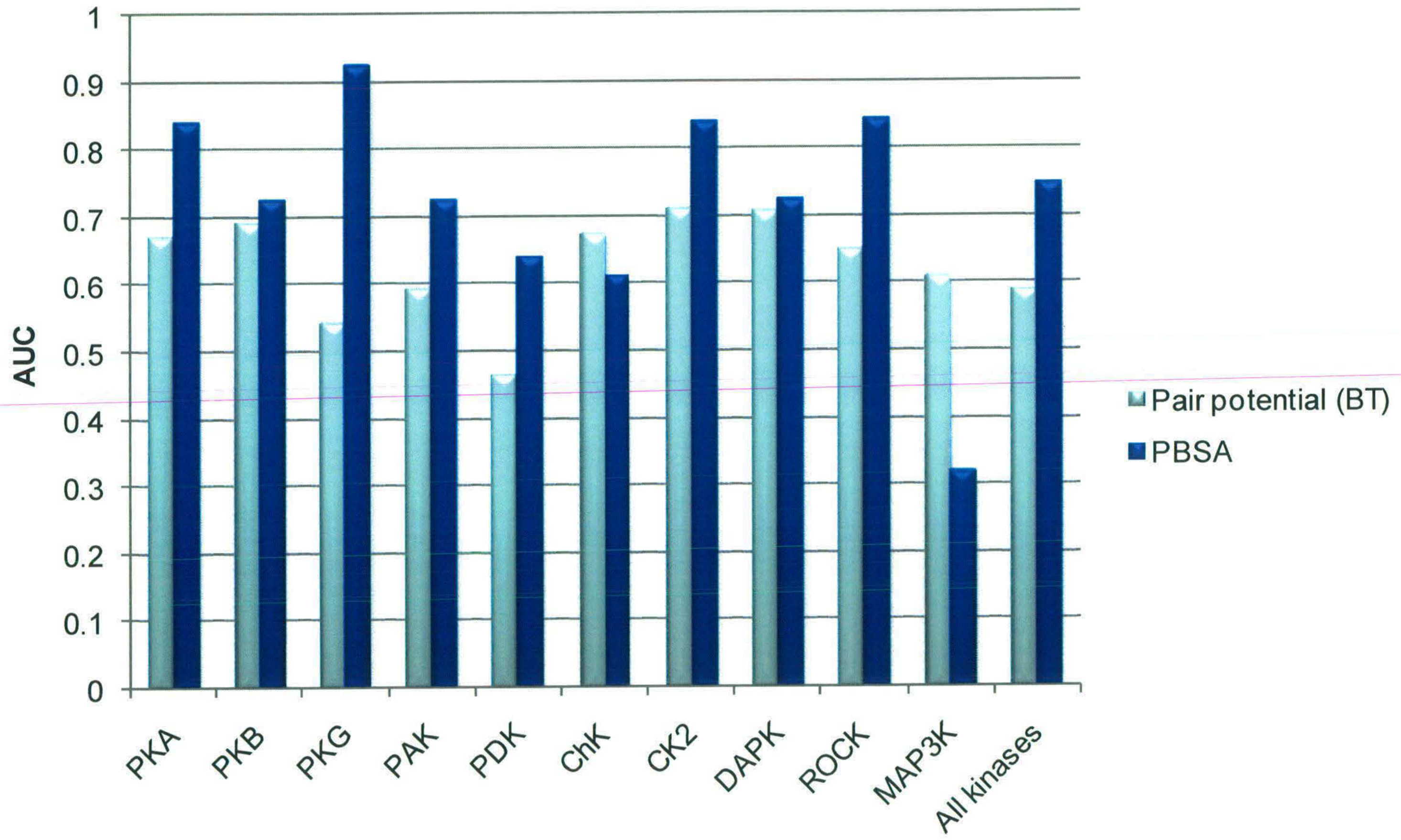


Figure 3.9: AUC values in ROC analysis for 30% top ranking peptides scored by BT pair potential matrix. These values are compared with the AUC value: obtained when the same set of peptides were re-ranked using the MM/PBSA method.

method compared to PREDIKIN. However, it may be noted that, we have compared our results to PREDIKIN 1.0 (Brinkworth *et al.*, 2003) as that was the available version when this work was carried out. The recent version of PREDIKIN (Saunders *et al.*, 2008; Saunders and Kobe, 2008) uses a scoring scheme derived from experimental phosphorylation site data and hence has higher prediction accuracy. Since it uses experimental data for its scoring scheme, it becomes similar in methodology to other profile based methods. Therefore, we have not compared our results with results from recent PREDIKIN server.

A common utility of phosphorylation site prediction programs is the prediction of substrates of genomic kinases for the prediction of phosphorylation networks in the organism. With the objective to identify the most suitable programs for different kinase families, we compared the accuracies of PREDIKIN, SCANSITE, NetPhosK, GPS, PPSP and MODPROPEP for various protein kinase families in human genome. Our structure based method, MODPROPEP was able to predict best or comparable to the closest competitor for 15 out of 33 kinases compared. This suggests that MODPROPEP can be used for identifying protein phosphorylation networks in various genomes. Since the scoring function used in MODPROPEP is not dependent on experimental data for any given kinase family, it can be potentially useful for kinase families for which no experimental data is available.

It must also be noted that, the predictions of MODPROPEP which have been compared with other methods are based on scoring by pair potential alone. We have also demonstrated that, use of a multi scale approach and re-ranking the high scoring peptides identified by pair potential using all atom MM/PBSA improves the percentile score of the true phosphorylation site. More detailed studies are necessary to test whether implementation of the complete multiscale approach can further improve the prediction accuracy of our structure based approach.

Chapter Four

Analysis of solvent accessible surface areas of phosphorylation sites in substrates of protein kinases

4.1 INTRODUCTION

The various available programs for prediction of phosphorylation sites by a given kinase, break the substrate proteins into all possible Ser/Thr/Tyr containing peptides and then evaluate the scores for the phosphorylation of each of these peptides. The high scoring peptides above certain cutoff value for the score are reported to be the most likely targets for phosphorylation. In the sequence based prediction methods, the score is computed in a probabilistic manner as per the probability of occurrence of various amino acids flanking the phosphorylatable Ser/Thr/Tyr. On the other hand, structure based methods calculate the score based on the binding affinity of each of Ser/Thr/Tyr containing peptides in complex with the protein kinase in question. All these programs assume that, all the peptides containing Ser/Thr/Tyr residues are equally accessible to the protein kinases for the phosphorylation reaction. The selectivity of the kinase for a target phosphorylation site is attributed exclusively to the residues flanking the phosphorylatable Ser/Thr/Tyr residues in the substrate protein. These residues, usually three on each side of the phosphorylation site make favorable contacts with the specificity determining residues in the protein kinase.

Although complementarities between the peptide residues and the kinase substrate binding pockets play a crucial role in the phosphorylation event, experimental evidences suggest that it is not the only factor determining the substrate selectivity. Other factors, such as the proximity of the substrate to the protein kinase play an equally important role (Ubersax and Ferrell, 2007). This process, called substrate recruitment, is any mechanism, which brings substrate and kinase in close proximity and thus increases the chances of substrate-kinase complex formation. One of the mechanisms of substrate recruitment involves the interaction between docking motifs on the substrate and interaction domain of the kinase (Biondi and Nebreda, 2003). These motifs are located far apart from the phosphorylation site in the substrate protein and increase the affinity of the substrate for the kinase many fold. In some substrates, phosphorylation event increases the affinity of the substrate for the next phosphorylation in the same substrate, this is a recurring theme in the phosphorylation by protein kinases. Localization of protein kinase in a specific subcellular compartment provides a further layer of specificity. Sometimes kinases interact with the substrate through an intermediary of scaffold protein, which acts as a platform for both interacting partners. Similarly, accessibility of the Ser/Thr/Tyr containing peptide stretch on a substrate protein also plays a major role in the substrate recognition by kinases. Some of the peptides, which otherwise fulfill the requirements of the high affinity to protein kinase, might be buried and hence not accessible for the protein kinase for the transfer of phosphate group. Therefore, in the absence of a conformation change in the substrate upon recruitment, only a subset of peptides, which is spatially located on the surface of the substrate protein, can potentially be phosphorylated.

Thus, a number of factors help in maintaining the high level of substrate specificity that is observed in the cellular phosphorylation networks. However, it is very difficult to incorporate all these effects in the prediction programs. Solvent accessibility of the phosphorylatable peptides can in principle be calculated, hence incorporation of the surface accessibility terms in the prediction algorithms might help in improving their prediction accuracy. However, none of the currently available computation programs for the substrate prediction with the exception of SCANSITE (Obenauer *et al.*, 2003), make use of the surface accessibility of the peptides while making predictions.

Large-scale high throughput mass spectrometric experiments have discovered a huge number of phosphorylated peptides which are substrates of protein kinases (Olsen *et al.*, 2006). Conventional kinase assays and peptide library experiments have also contributed to the known phosphorylation sites of the kinases (Songyang *et al.*, 1994). Phospho.ELM (Diella *et al.*, 2004) is a database, which catalogues these sites. Although substrate data stored in this database has been used for the development, or benchmarking of a number of prediction programs, no information about the accessibility of these peptides in their substrate protein is available. This, in part, is because of the crystal structures for most of the substrate proteins have not yet been solved. Phospho3D (Zanzoni *et al.*, 2007) is a database, which catalogues and stores the substrate protein whose structure has been solved along with the functional annotations at the phosphorylation site. It also stores the results of local structural alignment of substrates at the phosphorylation site. Although it provides the accessibility of phosphorylation sites, it is only for those few substrates whose structures are available in PDB. In another study, large-scale calculation of predicted values of solvent accessible area of known phosphorylation sites concluded that the phosphorylation sites are mostly accessible on the surface of the substrate protein, and are mostly found in the loop and hinge regions of the proteins (Gnad *et al.*, 2007). Although these two studies have shed some light on the surface accessibility of the known site of phosphorylation, one of these has information about a very small number of substrates, while the prediction of accessibilities from sequence alone becomes a limitation of the second study. Therefore, before inclusion of the solvent accessible area terms in the prediction algorithms, it is necessary to carry out an exhaustive analysis of the solvent accessibilities of phosphorylation sites in the known substrate proteins.

In this study, we have attempted to investigate whether inclusion of solvent accessibility probabilities of putative substrate peptides can help in improvement of prediction accuracy. We have calculated the solvent accessibilities of phosphorylation sites in the crystal structures or homology models of known substrate proteins. The accessibilities of the known phosphorylation sites have been compared with the accessibilities of the Ser/Thr/Tyr containing peptides which are not phosphorylated. Based on this analysis, we have attempted to estimate if statistically significant correlation exists between solvent accessibility and propensity for phosphorylation.

4.2 METHODS

4.2.1 Dataset of substrate protein of kinases

Substrate proteins of protein kinases catalogued in Phospho.ELM (Diella *et al.*, 2004) database (version 5.0, May 2006) were downloaded from UNIPROT database (http://www.uniprot.org). Information about the location of sites of phosphorylation on these proteins was also extracted. This version contains a total of 13563 phosphorylation sites in 4422 protein sequences.

4.2.2 Identification of structural homologs.

PDB (Berman *et al.*, 2000) database as on 27th November 2007 was used for identifying the structural homologs of substrate proteins. This PDB release consisted of 107691 polypeptide chains from 45658 unique structures. Amino acid sequences of these polypeptides chains were downloaded from RCSB website (http://www.rcsb.org). All PDB sequences were converted to a searchable database using the formatdb program of NCBI blast suite of softwares. For finding the structural homologs of the substrate proteins of various kinases, blast search was carried out against sequences in PDB with an e-value cutoff of 10^{-6}. The most significant BLAST hit for each substrate protein was selected as structural template. If the alignment between the template and the substrate protein showed a gap over the known phosphorylation site, the corresponding substrate sequence was removed from the data set. Figure 4.1 shows a flowchart depicting the protocol for mapping of phosphorylation sites onto the PDB structures. The PDB coordinate files corresponding to the structural templates of the substrate proteins were downloaded from PDB website.

4.2.3 Calculation of solvent accessible area of phosphorylation sites

The fragments of the structural templates aligning with the heptameric sequences of the substrate proteins containing the phosphorylation sites were identified. The absolute and relative solvent accessible surface areas of these heptameric peptide fragments in the template structure were calculated using the NACCESS (Hubbard and Thorntan, 1993) program. The solvent accessible surface areas of all other Ser/Thr/Tyr containing heptapeptides which are not phosphorylated were also calculated. Apart from the solvent accessible surface areas of the heptameric peptide fragments, the relative and absolute accessible surface areas were also calculated for the central Ser/Thr/Tyr residues.

4.2.4 Development of web server

All the computational steps discussed above have been implemented in a user friendly web server pAccess (http://www.nii.ac.in/paccess.html), for automatically carrying out the analysis of the solvent accessible areas of Ser/Thr/Tyr containing peptides present in putative substrate proteins of various kinases. Given the sequence of the substrate protein of a kinase as input (Figure 4.2), the software identifies its structural homolog in PDB, and calculates the solvent accessible surface area of all seven residue long peptides that contain Ser/Thr/Tyr as the central residues. All of these peptides are compared against the experimentally known phosphorylation sites catalogued in Phospho.ELM database. Any of these Ser/Thr/Tyr containing peptide that is already known to be a phosphorylation site, is reported along with the name of protein kinase responsible for its phosphorylation. The accessibilities of the Ser/Thr/Tyr containing peptides are displayed using a jmol applet (Figure 4.3).

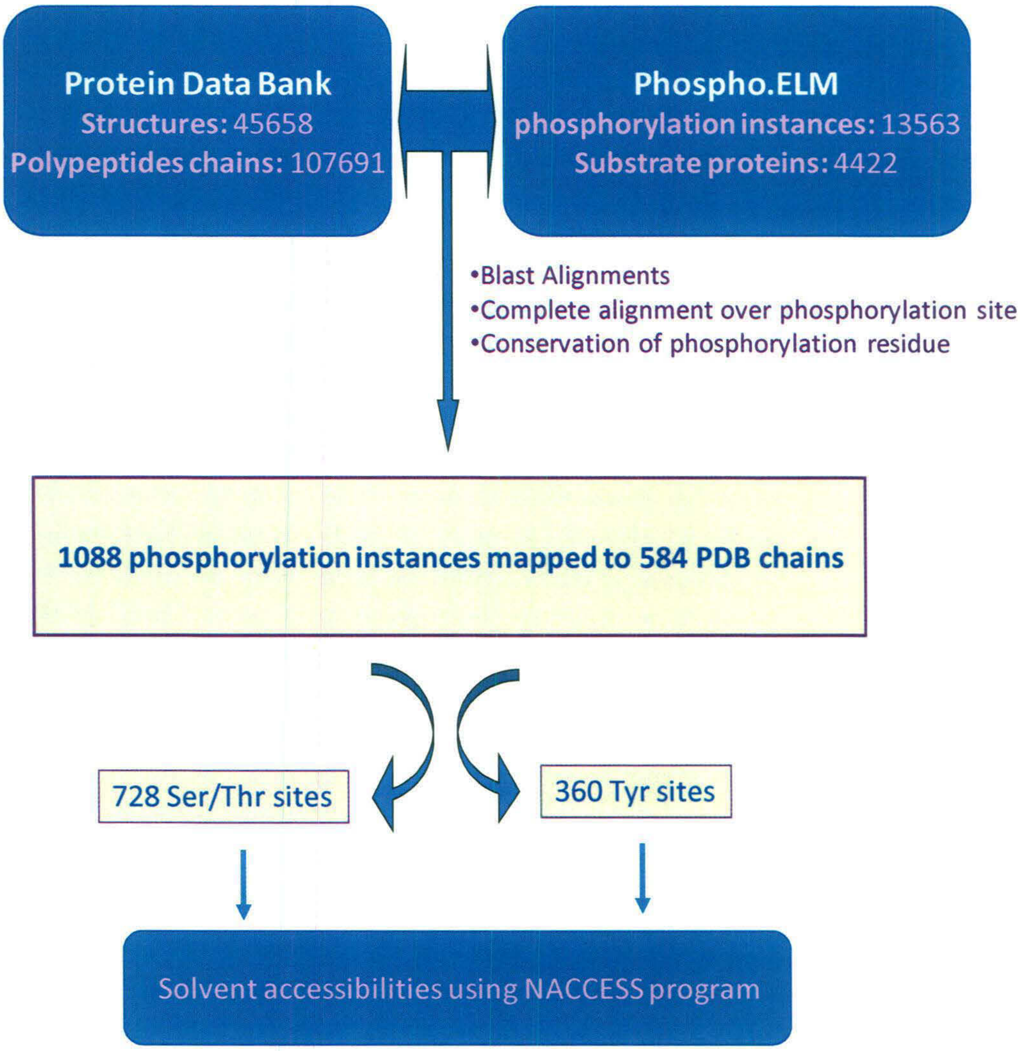
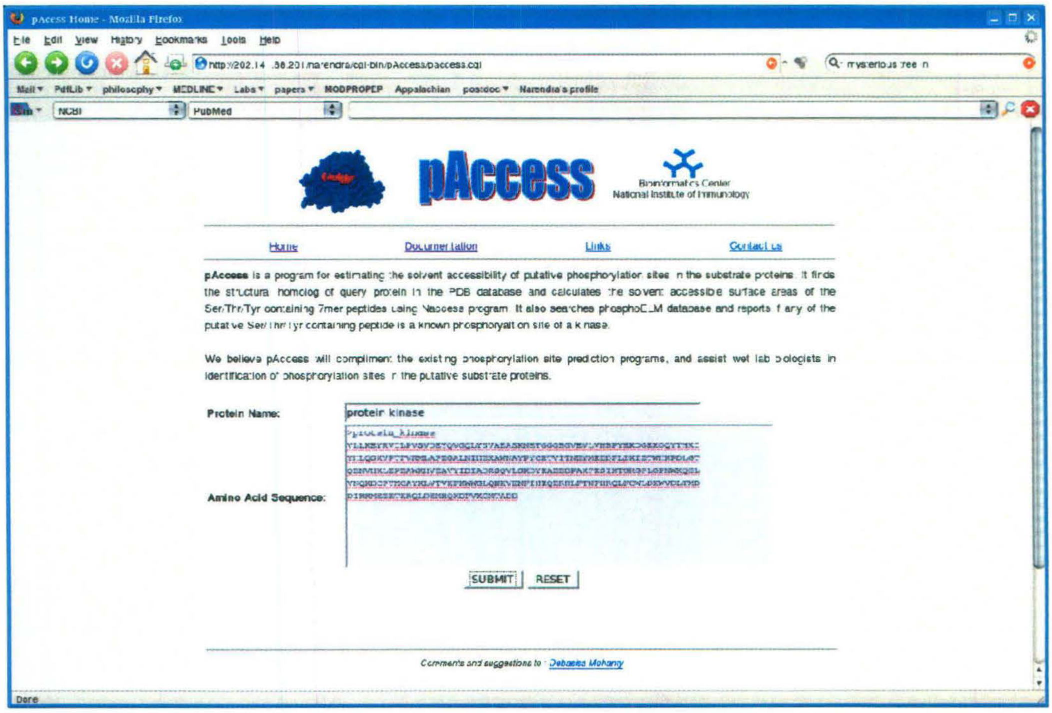

Protein Data Bank
Structures: 45658
Polypeptides chains: 107691
Phospho.ELM
phosphorylation instances: 13563
Substrate proteins: 4422
•Blast Alignments
•Complete alignment over phosphorylation site
•Conservation of phosphorylation residue
1088 phosphorylation instances mapped to 584 PDB chains
728 Ser/Thr sites
360 Tyr sites
Solvent accessibilities using NACCESS program

Figure 4.1: Flowchart depicting mapping of phosphorylation sites in the substrate protein to the structural homolog identified from PDB.

Figure 4.2: A snapshot of the query interface of the pAccess program.

Figure 4.3: A snapshot of the result page of pAccess program for the analysis of solvent accessible surface area values for a query protein kinase substrate protein. The absolute and relative surface area values of all heptapeptides in the query protein containing Ser/Thr/Tyr as the central residue are reported. If any of these peptides is an experimentally known phosphorylation site, the substrate protein containing that peptide and the protein kinase responsible for the phosphorylation are reported. Clicking on the peptide shows the peptide in the template structure in an interactive Jmol applet at the right hand side of the page.


pAccess Home - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://202.14 .58.201/narendra/cgi-bin/pAccess/paccess.cgi
Mail PdfLib philosophy MEDLINE Labs papers MODPROPEP Appalachian postdoc Narendra's profile
NCBI PubMed
pAccess
Bioinformatics Center
National Institute of Immunology
Home Documentation Links Contact us
Done

pAccess is a program for estimating the solvent accessibility of putative phosphorylation sites in the substrate proteins. It finds the structural homolog of query protein in the PDB database and calculates the solvent accessible surface areas of the Ser/Thr/Tyr containing 7mer peptides using Naccess program. It also searches phosphoELM database and reports if any of the putative Ser/Thr/Tyr containing peptide is a known phosphoryaltion site of a kinase.

We believe pAccess will compliment the existing phosphorylation site prediction programs, and assist wet lab biologists in identification of phosphorylation sites in the putative substrate proteins.

Protein Name: protein kinase

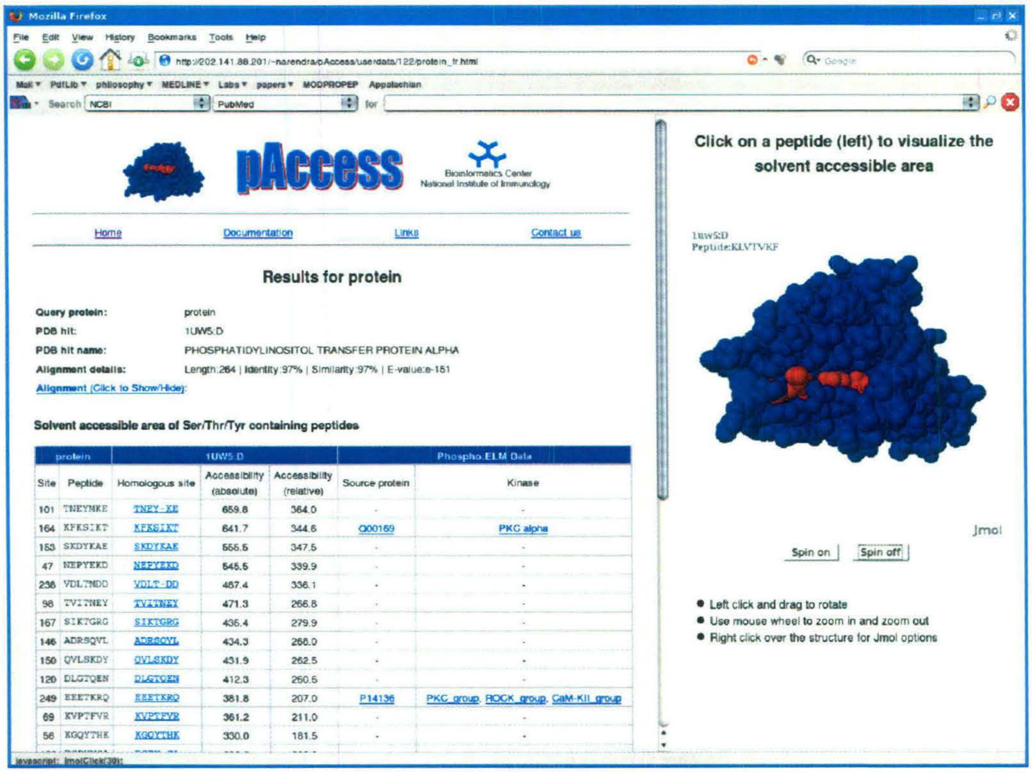
Amino Acid Sequence:

```
>protein_kinase
VLLKEYRVILPVSVDEYQVGQLYSVAEASKNETGGGEGVEVLVNEPYEKDGEKGQYTHK
IYLQQKVPTFVRMLAPEGALNIHEKAWNAYPYCRTVITNEYMKEDFLIKIETWHKPDLGT
QENVHKLEPEAWKHVEAVYIDIADRSQVLSKDYKAEEDPAKFKSIKTGRGPLGPNWKQEL
VNQKDCPYMCAYKLVTVKFKWWGLQNKVENFIHKQEKRLFTNFHRQLFCWLDKWVDLTMD
DIRRMEEETKRQLDEMRQKDPVKGMTADD
```

SUBMIT | RESET

Comments and suggestions to : Debasisa Mohanty

Figure 4.2


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://202.141.88.201/~narendra/pAccess/userdata/122/protein_fr.html
Google
Mail PdfLib philosophy MEDLINE Labs papers MODPROPEP Appalachian
Search NCBI PubMed for
pAccess
Bioinformatics Center
National Institute of Immunology
Home Documentation Links Contact us
1UW5:D
Peptide:KLVTVKF
Jmol
Spin on
Spin off
javascript: jmolClick(30);

Results for protein

Query protein: protein
PDB hit: 1UW5:D
PDB hit name: PHOSPHATIDYLINOSITOL TRANSFER PROTEIN ALPHA
Alignment details: Length:264 | Identity:97% | Similarity:97% | E-value:e-151
Alignment (Click to Show/Hide):

Solvent accessible area of Ser/Thr/Tyr containing peptides

protein		1UW5:D			Phospho.ELM Data	
Site	Peptide	Homologous site	Accessibility (absolute)	Accessibility (relative)	Source protein	Kinase
101	TNEYMKE	TNEY-KE	659.8	364.0	-	-
164	KFKSIKT	KFKSIKT	641.7	344.6	O00169	PKC alpha
153	SKDYKAE	SKDYKAE	555.5	347.5	-	-
47	NEPYEKD	NEPYEKD	545.5	339.9	-	-
236	VDLTMDD	VDLT-DD	467.4	336.1	-	-
98	TVITNEY	TVITNEY	471.3	266.8	-	-
167	SIKTGRG	SIKTGRG	436.4	279.9	-	-
146	ADRSQVL	ADRSQVL	434.3	266.0	-	-
150	QVLSKDY	QVLSKDY	431.9	262.5	-	-
120	DLGTQEN	DLGTQEN	412.3	260.6	-	-
249	EEETKRQ	EEETKRQ	381.8	207.0	P14136	PKC_group, ROCK_group, CaM-KII_group
69	KVPTFVR	KVPTFVR	361.2	211.0	-	-
56	KGQYTHK	KGQYTHK	330.0	181.5	-	-

Click on a peptide (left) to visualize the solvent accessible area

- Left click and drag to rotate
- Use mouse wheel to zoom in and zoom out
- Right click over the structure for Jmol options

Figure 4.3

4.3 RESULTS

4.3.1 Structural homologs of substrate proteins of kinases

Out of 4422 substrate protein sequences catalogued in the Phospho.ELM database, structural homologues could be identified for 1425 proteins containing 2860 phosphorylation sites. The structural templates for these 1425 substrate proteins corresponded to 990 PDB structures. A careful examination of the sequence alignments of substrate proteins with the structures in PDB showed that, in some of the alignments the region corresponding to the phosphorylation site contained a gap in the structure, or the phosphorylatable Ser/Thr/Tyr residue was replaced by a non-phosphorylatable amino acid. In such cases, the sequence of the substrate protein was removed from the data set. After applying this filter, the data set consisted of 1088 phosphorylation instances from 719 substrate proteins which could be mapped on to 584 polypeptide chains from 571 PDB entries. These included 526, 202 and 360 instances containing serine, threonine and tyrosine as phosphorylation residues respectively. Figures 4.4 and 4.5 show the distribution of the alignment length and percentage identity respectively. As can be seen from Figure 4.4, 85.8% of all hits show an alignment over a length of 100 or more amino acids. Similarly, the percentage identities of the hits were also significant as about 44% of hits showed a percentage identity in the range of 90-100%, while 90% of the hits showed identity above 30% (Figure 4.5). Thus, the structural templates have good homology with the substrate proteins over a significant length of alignment. Hence, these 719 substrate proteins are likely to have structures similar to those of the templates in PDB. Therefore, the solvent accessibilities of the Ser/Thr/Tyr containing peptide stretches in the substrate proteins can be estimated based on accessibility of the corresponding structural fragments in the identified PDB hits.

4.3.2 Solvent accessibility of phosphorylation sites

Solvent accessible surface areas of the phosphorylation sites on the substrate proteins were calculated from their crystal structures when available, or from their structural homologs. Seven amino acids long peptide stretches with phosphorylation site as the central residue were selected for the calculation of solvent accessible surface areas by NACCESS computer program. For each substrate protein, apart from the phosphorylation site, the accessible surface areas of all other Ser/Thr/Tyr containing potential sites which are not phosphorylated, were also calculated. In

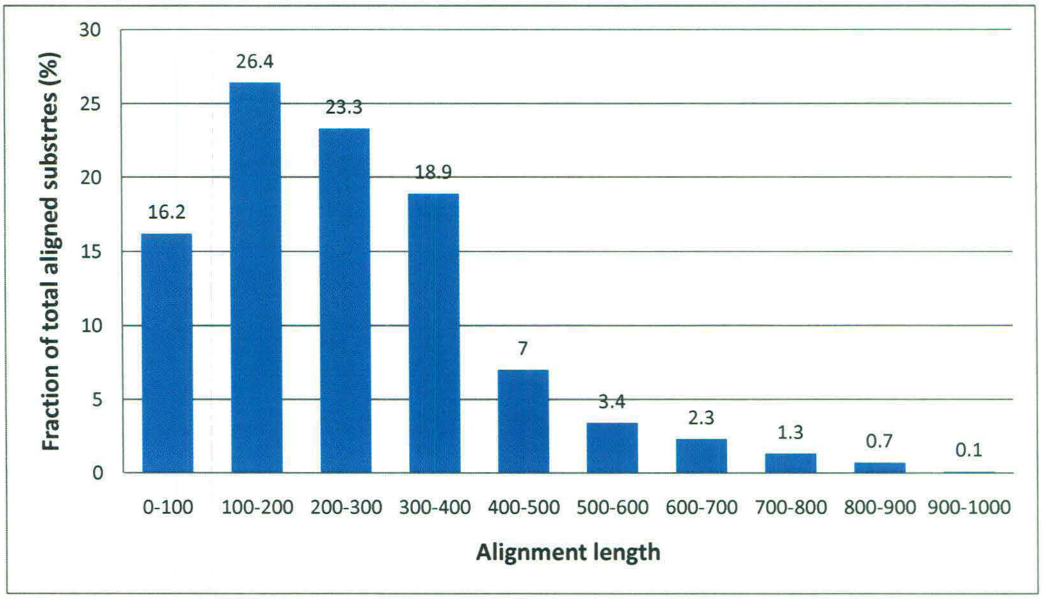


Figure 4.4: The distribution of length of sequence alignments between kinase substrates and homologous proteins in PDB sequence database.

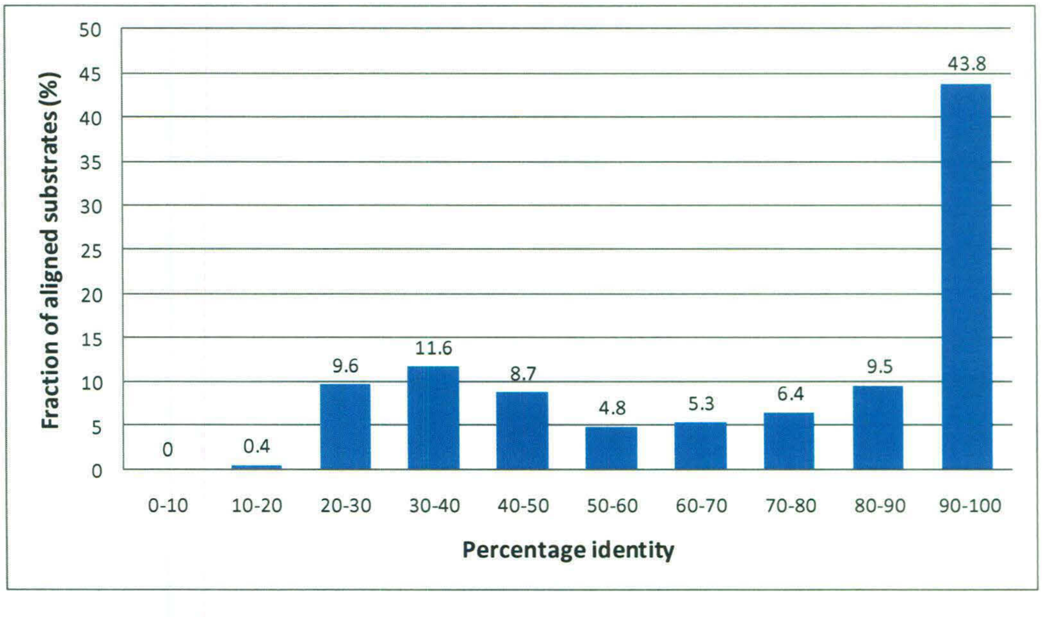
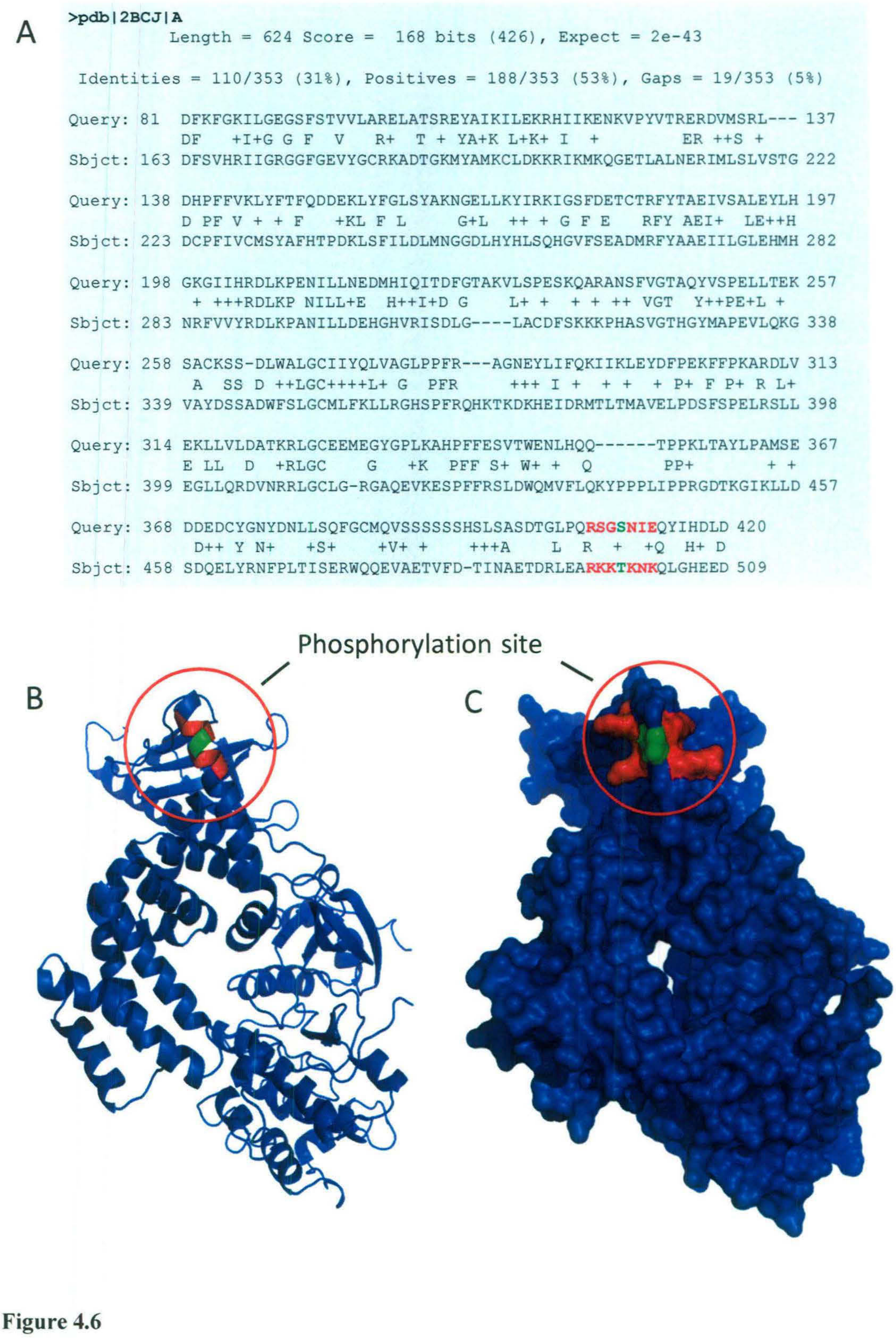


Figure 4.5: Distribution of the sequence identities of the kinase substrate proteins which showed significant alignment with protein sequences in PDB database.

every case, absolute as well as relative surface areas were calculated. The relative surface area values represent the surface accessibility in comparison to the accessibility of a residue in extended conformation when it is surrounded by alanine residues on both sides. Figure 4.6 shows an example of the analysis of solvent accessibility of a representative query substrate protein 3-phosphoinositide dependent protein kinase-1 whose solvent accessible surface area of phosphorylation site is calculated from the structural homolog with PDB ID 2BCJ. Figure 4.7 shows the average solvent accessible surface areas for the central Ser/Thr/Tyr residues as well as the heptameric peptide fragment corresponding to sites which are known to be phosphorylated as well as for those which are not phosphorylated. As can be seen, in general the average solvent accessible surface area of the phosphorylation site residues is higher as compared to the accessibilities of Ser/Thr/Tyr containing peptides that are not phosphorylated. The statistical significance of the difference between the accessibilities of phosphorylation site residues and non-phosphorylated residues was judged by Wilcoxon test. The difference in average relative solvent accessible area between phospho and non-phospho residues was statistically significant as judged by p-values of 2.20×10^{-16}, 5.07×10^{-6} and 2.34×10^{-8} for Ser, Thr and Tyr residues. Similarly, p-values for the difference between the average relative solvent accessible area of phospho-peptide and non-phospho peptides were 2.20×10^{-16}, 2.05×10^{-13} and 4.77×10^{-12} for Ser, Thr and Tyr containing peptides respectively. Figure 4.7 also shows that the difference between the average solvent accessibility values for the absolute area was also significant with great degree of confidence as judged by Wilcoxon test p-values. Thus, based on average accessibility values the phosphorylation site residues/peptides were found to be more exposed to the solvent. Figure 4.8 shows a comparison of the distribution of absolute accessibilities of phosphorylation site containing peptides with their non-phosphorylatable counterparts. As can be seen from Figure 4.8, at high accessibility values (> 400 Å^2) the percentage of phosphorylated peptides are higher than the percentage of peptides which are not phosphorylated, while the trend is reversed in the low accessibility range. However, a surprising observation from this analysis is the occurrence of a significant number of phosphorylated peptides even at accessibility values below 200 Å^2.

Since the results shown in Figure 4.8 is based on analysis of accessibility values in crystal structures homologous to various kinase substrates, we decided to

Figure 4.6: An example of analysis of solvent accessibility of heptapeptide containing site of phosphorylation in kinase substrate protein, calculated by NACCESS program from the homologous protein identified by BLAST search of PDB sequences. (A) Sequence alignment of query substrate 3-phosphoinositide dependent protein kinase-1 (O15530) with the sequence of structural homologue guanine nucleotide binding protein:GPCR kinase 2 (PDB identifier: 2BCJ). The phosphorylation site residue has been shown in green and three residues on each side are shown in red. (B) Cartoon representation of 2BCJ. Phosphorylation site and the residues surrounding it are shown in red and green respectively. (C) Surface representation of 2BCJ. The peptide stretch corresponding the phosphorylation site and surrounding residues have been shown in green and red. The surface area of this region was calculated by NACCESS program.



```
>pdb|2BCJ|A
          Length = 624 Score =  168 bits (426), Expect = 2e-43

 Identities = 110/353 (31%), Positives = 188/353 (53%), Gaps = 19/353 (5%)

Query: 81  DFKFGKILGEGSFSTVVLARELATSREYAIKILEKRHIIKENKVPYVTRERDVMSRL--- 137
            DF   +I+G G F  V   R+  T + YA+K L+K+ I  +        ER ++S +
Sbjct: 163 DFSVHRIIGRGGFGEVYGCRKADTGKMYAMKCLDKKRIKMKQGETLALNERIMLSLVSTG 222

Query: 138 DHPFFVKLYFTFQDDEKLYFGLSYAKNGELLKYIRKIGSFDETCTRFYTAEIVSALEYLH 197
            D PF V + + F   +KL F L     G+L  ++ + G F E   RFY AEI+  LE++H
Sbjct: 223 DCPFIVCMSYAFHTPDKLSFILDLMNGGDLHYHLSQHGVFSEADMRFYAAEIILGLEHMH 282

Query: 198 GKGIIHRDLKPENILLNEDMHIQITDFGTAKVLSPESKQARANSFVGTAQYVSPELLTEK 257
             + +++RDLKP NILL+E  H++I+D G    L+ +  + + ++ VGT  Y++PE+L +
Sbjct: 283 NRFVVYRDLKPANILLDEHGHVRISDLG----LACDFSKKKPHASVGTHGYMAPEVLQKG 338

Query: 258 SACKSS-DLWALGCIIYQLVAGLPPFR---AGNEYLIFQKIIKLEYDFPEKFFPKARDLV 313
             A  SS D ++LGC++++L+ G  PFR     +++ I +  + +  + P+ F P+ R L+
Sbjct: 339 VAYDSSADWFSLGCMLFKLLRGHSPFRQHKTKDKHEIDRMTLTMAVELPDSFSPELRSLL 398

Query: 314 EKLLVLDATKRLGCEEMEGYGPLKAHPFFESVTWENLHQQ------TPPKLTAYLPAMSE 367
            E LL  D  +RLGC    G   +K  PFF S+ W+ + Q       PP+       + +
Sbjct: 399 EGLLQRDVNRRLGCLG-RGAQEVKESPFFRSLDWQMVFLQKYPPPLIPPRGDTKGIKLLD 457

Query: 368 DDEDCYGNYDNLLSQFGCMQVSSSSSSHSLSASDTGLPQRSGSNIEQYIHDLD 420
             D++ Y N+   +S+    +V+ +    +++A    L  R  +  +Q  H+ D
Sbjct: 458 SDQELYRNFPLTISERWQQEVAETVFD-TINAETDRLEARKKTKNKQLGHEED 509
```

Figure 4.6

Figure 4.7: The average solvent accessible surface area values as calculated by NACCESS program for phosphorylation site Ser, Thr and Tyr residues and heptapeptides consisting of three residues on each side of these residues. The surface accessible areas of these residues and heptapeptides are compared to the surface accessible areas of Ser, Thr and Tyr residues which are not phosphorylation sites (non-phospho), and heptapeptides containing these non-phospho residues as the central residues. For every case, relative and absolute surface accessible areas are calculated from either the crystal structure, or the structural homolog of the substrate proteins (see methods). **(A)** Relative solvent accessibilities of Ser, Thr and Tyr residues. The difference between phospho and non-phospho residues is statistically significant (Wilcoxon test: p values 2.20×10^{-16}, 5.07×10^{-6} and 2.34×10^{-8} for Ser, Thr and Tyr respectively). **(B)** Relative solvent accessibilites of heptapeptides containing Ser/Thr/Tyr as the central residues (Wilcoxon test: p values 2.20×10^{-16}, 2.05×10^{-13} and 4.77×10^{-12} for Ser, Thr and Tyr respectively). **(C)** Absolute solvent accessibilities of Ser , Thr and Tyr residues (Wilcoxon test: p values 1.19×10^{-10}, 2.30×10^{-5} and 5.20×10^{-8} for Ser, Thr and Tyr respectively). **(D)** Absolute solvent accessibilites of heptapeptides containing Ser/Thr/Tyr as the central residues (Wilcoxon test: p values 2.20×10^{-16}, 1.98×10^{-13} and 1.68×10^{-10} for Ser, Thr and Tyr respectively).

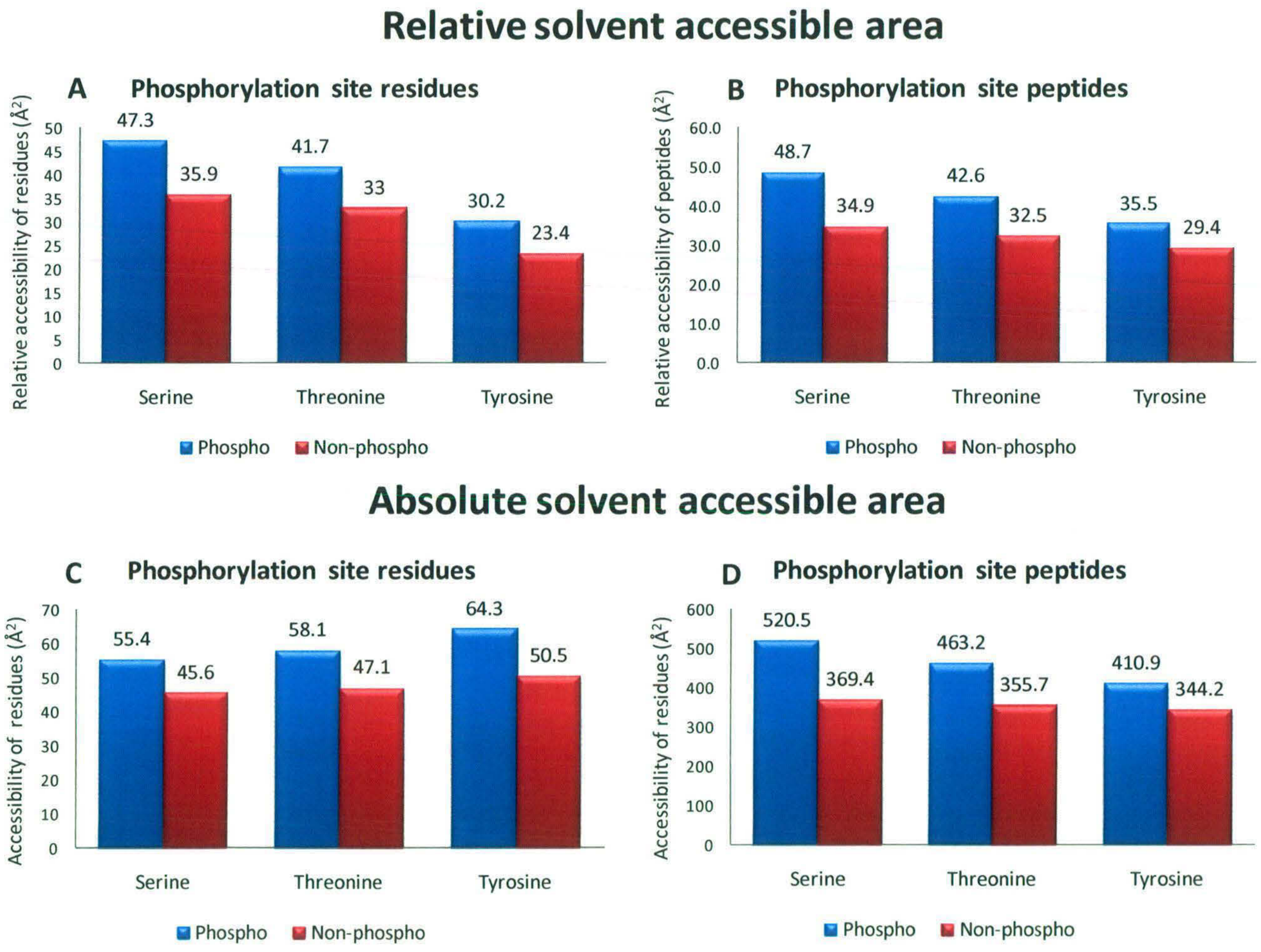

Relative solvent accessible area
A Phosphorylation site residues
Relative accessibility of residues (Å²)
50
45
40
35
30
25
20
15
10
5
0
47.3
35.9
41.7
33
30.2
23.4
Serine
Threonine
Tyrosine
🟦 Phospho 🟥 Non-phospho
B Phosphorylation site peptides
Relative accessibility of peptides (Å²)
60.0
50.0
40.0
30.0
20.0
10.0
0.0
48.7
34.9
42.6
32.5
35.5
29.4
Serine
Threonine
Tyrosine
🟦 Phospho 🟥 Non-phospho
Absolute solvent accessible area
C Phosphorylation site residues
Accessibility of residues (Å²)
70
60
50
40
30
20
10
0
55.4
45.6
58.1
47.1
64.3
50.5
Serine
Threonine
Tyrosine
🟦 Phospho 🟥 Non-phospho
D Phosphorylation site peptides
Accessibility of residues (Å²)
600
500
400
300
200
100
0
520.5
369.4
463.2
355.7
410.9
344.2
Serine
Threonine
Tyrosine
🟦 Phospho 🟥 Non-phospho

Figure 4.7

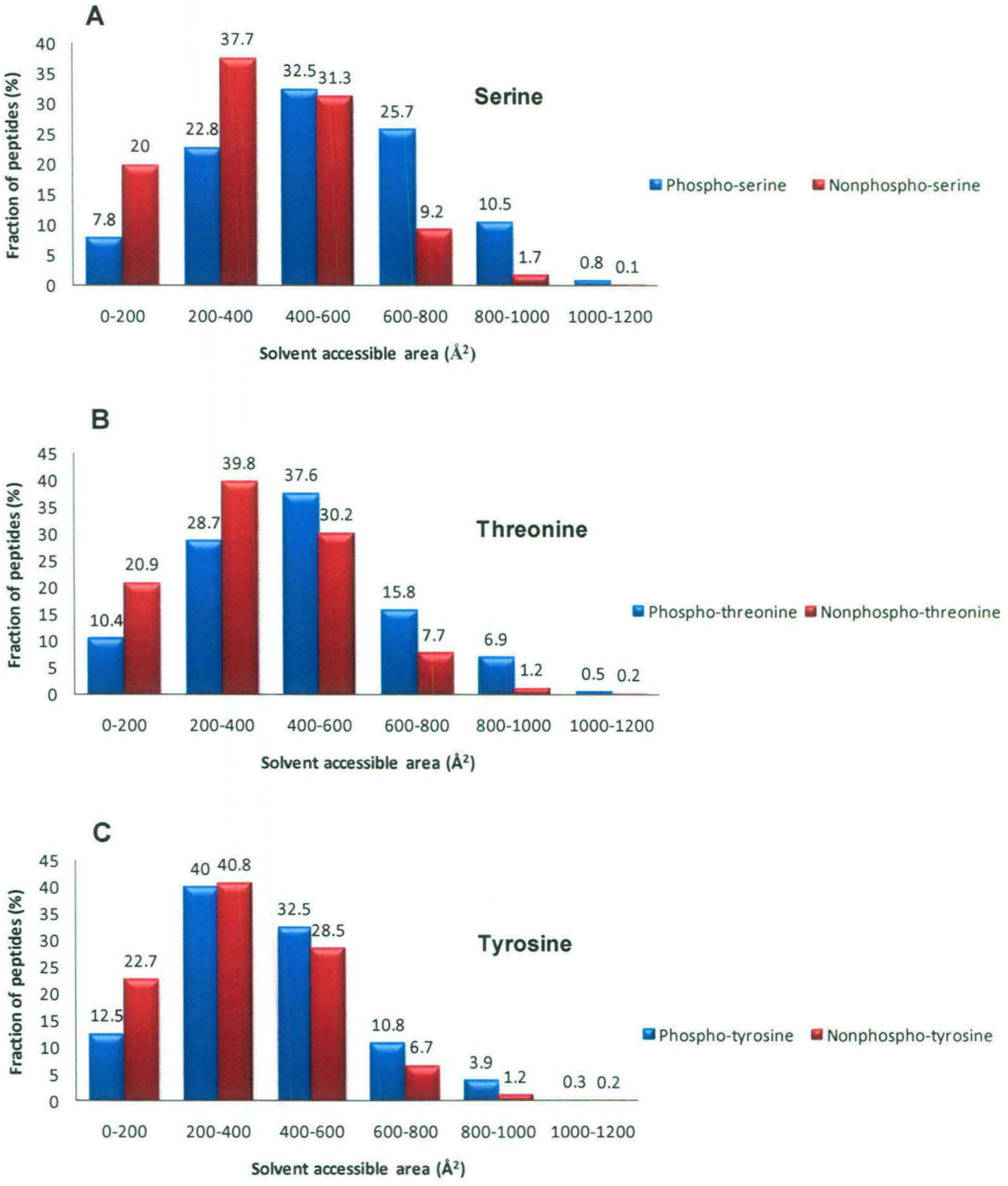


Figure 4.8: Distribution of solvent accessible surface areas of phosphorylation site containing peptides as compared to nonphospho-site containing peptides, when phosphorylation site is Ser (A), Thr (B) and Tyr (C).

nalyze separately the accessibilities of phosphorylation sites in substrate proteins for vhich crystal structures were available. Some of the protein kinase substrates in our lata set had structural hits from PDB with 100% match over alignment length of more han 200 amino acids. Table 4.1 lists 25 such proteins that showed 100% match with 'DB entries. In 19 of these proteins, the peptides harboring the phosphorylation site ıad accessibility values higher than 350 Å^2. However, for the remaining six proteins he known phosphorylation sites were less accessible than the other potential sites. 'his suggests that observation of known phosphorylation sites at low accessibility 'alues is not an artifact arising from structural homologs included in our analysis.

In our analysis of solvent accessibility of known phosphorylation sites, a lot of ites considered in dataset had been identified by high throughput mass spectroscopic xperiments. The protein kinases that phosphorylate these sites are not known and here is a possibility that some of these sites are false arising because of the ıncertainty in correctly assigning the sequence of the phosphorylation site from the nass spectrometry data. For these reasons, we reanalyzed the average solvent ccessible surface area and the distribution of the solvent accessible surface area after emoving the phosphorylation sites identified by mass spectrometry from the dataset. After removing such site, we were left with 223, 90 and 119 instances containing Ser, 'hr and Tyr residues respectively. Figure 4.9 shows the average relative and average bsolute solvent accessible surface area of phosphorylation site residues and the 7mer ıeptide harboring phosphorylation site as compared to their nonphosporylatable Ser, 'hr and Tyr containing counterparts when instances identified by mass spectrometry vere not considered. It can be seen that the average surface area values in every case emain more or less same as when such site were considered (Figure 4.7). Figure 4.10 hows the comparison of distribution of absolute solvent accessible area values of ıhosphorylation sites containing peptides with their nonphosphorylatable counterpart vhen the phosphorylation instances identified by mass specrtometry were not ncluded. It can be seen clearly from Figure 4.10, at low accessibility value (<400Å^2) he percentage of phosphorylated peptides has reduced as compared to the values vhen sites by mass spectrometry were included in the analysis (Figure 4.8).

We analyzed in details the substrate proteins having known phosphorylation ites at low accessibility. Table 4.2 lists substrate proteins with phosphorylation sites vhose relative surface areas are less than 10 Å^2. Many of these sites are buried in the espective proteins and phosphorylation of such sites might require conformational

Table 4.1: The solvent accessible surface areas of phosphorylation site in protein kinase substrates which show 100% match over more than 200 alignment length with a PDB structure. The absolute solvent accessible area of heptapeptide containing phosphorylation site, and absolute and relative areas of phosphorylation site Ser/Thr/Tyr residues are calculated by NACCESS.

Table headers: **(A)** Substrate protein containing the phosphorylation site for protein kinase in column **J**. **(B)** Swissprot / TrEmbl code for the substrate protein. **(C)** Residue number of phosphorylation site Ser/Thr/Tyr in the protein sequence of substrate. **(D)** Heptapeptide sequence containing the phosphorylation site (red). **(E)** PDB ID of structural match of substrate protein. **(F)** Absolute solvent accessible surface area of heptapeptide (in column D) as calculated by NACCESS program. **(G)** Absolute solvent accessible area of phosphorylation site Ser/Thr/Tyr residue (colored red in column D). **(H)** Relative solvent accessible area of phosphorylation site Ser/Thr/Tyr residue. **(I)** Average relative solvent accessible area of all Ser/Thr/Tyr residues in the structural match. **(J)** Protein kinase responsible for phosphorylation of the substrate at the phosphorylation site.

A	B	C	D	E	F	G	H	I	J
Catalase	P04040	230	eav:ckf	1QQW	169.6	14.5	6.8	30.2	Abl, Abl2
Phosphoglycerate mutase	P18669	22	nrf.'gwy	1YJX	185.2	14.7	12.6	16.3	PAK1
Proto-oncogene tyrosine	P12931	215	ggf·its	2H8H	253.4	44.0	20.7	31.3	Src
Serine/threonine-protein kinase 6	O14965	266	llg.'age	2J4Z	282.9	88.4	11.9	34.2	Aurora A
Serine/threonine-protein kinase 6	O14965	278	fgw.'vha	2J4Z	305.5	41.2	6.2	34.2	Aurora A
Glucose-6-phosphate isomerase	P06744	184	wyv.'nid	1NUH	323.1	13.8	18	30.3	CK2 group
Casein kinase II subunit alpha	P68400	255	edl·dyi	1JWH	374.9	79.5	37.4	30.6	Lyn, Fgr
Dihydropteridine reductase	P09417	223	nrp.'sgs	1HDR	381.3	86.0	75.9	27.5	CaM-KII group
E3 ubiquitin-protein ligase CBL	P22681	371	yel·cem	1FBV	393.7	13.2	40.9	29.4	EGFR, InsR
Retinoic acid receptor RXR-alpha	P19793	249	tet·vea	1LBD	402.5	32.9	34.9	41.5	MAP2K4
Phosphoglycerate mutase 1	P18669	117	wrr.'ydv	1YJX	418.0	20.9	73.9	16.3	PAK1
C-Rel proto-oncogene protein	P16236	266	rrp.'dqe	1GJI	419.4	86.0	73.8	30.6	PKA group
14-3-3 protein eta	Q04917	58	rrs.'wrv	2C74	445.7	47.6	40.7	31.2	SDK1
Syntaxin-1A	Q16623	188	ssi.'kqa	1DN1	487.4	13.6	11.7	38.0	DAPK group
Band 3 anion transport protein	P02730	303	maq:rge	1HYN	492.8	48.5	41.6	29.3	CK1 alpha

Table 4.1 (continued on text page)

A	B	C	D	E	F	G	H	I	J
Retinoic acid receptor RXR-alpha	P19793	260	nps.'pnd	1LBD	500.6	56.0	25.8	41.5	MAPK1, MAPK3, MAPK group
Interferon regulatory factor 3	Q14653	385	gga'sle	1QWT	507.4	46.9	59.2	39.0	IKK-epsilon, TBK1
Interferon regulatory factor 3	Q14653	386	gas.len	1QWT	518.1	97.6	30.3	39.0	IKK group
Serine/threonine-protein kinase 6	O14965	391	skp.'ncq	2J4Z	524.5	48.8	41.9	34.2	Aurora A
Catalase	P04040	385	van:qrd	1QQW	526.9	54.9	66	30.2	Abl, Abl2
Serine/threonine-protein kinase 6	O14965	226	qkl.'kfd	2J4Z	574.7	69.0	48	34.2	Aurora A
Serine/threonine-protein kinase 6	O14965	287	srr:tlc	2J4Z	581.1	91.9	40.3	34.2	Aurora A
Tyrosine-protein kinase BTK	Q06187	550	dde:tss	1K2P	615.6	52.1	83.8	36.7	Lyn, BTK
Proteasome subunit alpha type 3	P25788	242	ake.ike	1IRU	623.6	35.3	24.5	36.3	CK2 group
Eukaryotic translation initiation factor 4E	P07260	15	env.'vdd	1AP8	748.6	113.3	97.2	43.0	CK2 group

Table 4.1 (continued from previous page)

Table 4.2: The substrate protein of protein kinases with relative surface accessible area of phosphorylation site less than 10 Å^2. All these substrates show a significant alignment over 200 amino acids with the sequence of a match from the PDB database. The relative solvent accessible area was calculated by NACCESS program.

Table headers: **(A)** Swissprot / TrEmbl code for substrate protein containing the phosphorylation site for protein kinase in column **H**. **(B)** Residue number of phosphorylation site Ser/Thr/Tyr in the protein sequence of substrate. **(C)** Heptapeptide sequence containing the phosphorylation site (red). **(D)** PDB ID of structural match of substrate protein. **(E)** Alignment length of the substrate with the structural match **(F)** The percentage identity of substrate protein with structural match over the alignment length. **(G)** Relative solvent accessible surface area of phosphorylation site Ser/Thr/Tyr residue (shown in red in column C) as calculated by NACCESS program. **(H)** Protein kinase responsible for phosphorylation of the substrate at the phosphorylation site.

A	B	C	D	E	F	G	H
P08575	1216	feqYqfl	1YGU	603	93	0.0	Csk
P10275	791	rhlSqef	2PNU	250	98	0.0	PKB group
P16452	247	wigSvdi	2Q3Z	693	49	0.0	PKA group
P39748	187	tfgSpvl	1UL1	355	98	0.0	CDK1
P51617	209	knvTnnf	2OID	326	48	0.0	IRAK1
Q63270	138	radSlqk	2B3Y	887	97	0.0	PKC group
Q9UEW8	373	plhTsrv	2OZA	360	42	0.0	Wnk1
P18031	153	ktyYtvr	1L8G	297	100	0.1	InsR
P68369	432	ekdYeev	1SA1	438	97	0.1	Syk
P17655	369	rgsTagg	1KFX	699	91	0.2	PKA group
P18031	66	dndYina	1L8G	297	100	0.3	EGFR, InsR
P31152	196	wyrSprl	2I6L	296	80	0.4	MAPK4
Q9P286	573	ksdSill	2F57	297	98	0.4	PAK5
P22001	135	irfYelg	2R9R	389	83	0.5	Src
P04167	128	rrfSlat	2Q6N	464	89	0.6	PKA group
P35270	213	retSvdp	1Z6Z	257	97	0.6	CCDPK
P18206	1100	nlqSvke	1TR2	1129	85	0.8	PKC alpha
P31946	131	rylSeva	2C23	230	97	0.8	PKC zeta
P43403	474	nrhYaki	2OZO	605	88	0.8	Lck
P11799	1748	griSnys	1KOA	430	57	1.0	CaM-KII group
Q9UQM7	286	rqeTvec	2V7O	297	94	1.2	CaM-KII alpha
P29322	839	erpYwem	2QOK	278	84	1.4	EphA8
Q63270	711	sygSrrg	2B3Y	887	97	1.9	PKC group
P61763	158	sfySphk	1DN1	589	94	3.1	CDK group
O43293	225	kqeTltn	1YRP	276	99	3.5	DAPK3
P05532	934	revSfyy	1P4O	380	45	3.8	Fyn
Q00169	164	kfkSikt	1UW5	264	97	3.8	PKC alpha
Q29502	197	nveSlld	2H9V	326	47	3.8	PAK2
P18031	50	rdvSpfd	1L8G	297	100	3.9	PKB group CLK1
Q9UM73	1278	rdiYetd	1P4O	295	63	3.9	ALK

Table 4.2 (continued on next page)

A	B	C	D	E	F	G	H
P68104	432	mrqTvav	2B7C	439	89	4.0	PKC group
O96017	516	lhtSrvl	2OZA	319	51	4.1	CHK2
P13569	660	rrnSilt	1XMI	283	93	4.6	PKA group PKG/cGK group
P37173	424	trrYmap	2QLU	299	65	4.8	TGFbR2
P18031	152	iktYytv	1L8G	297	100	5.0	InsR
P22681	371	yelYcem	1FBV	388	100	6.2	EGFR, InsR
O00571	322	lvaTpgr	2I4I	413	98	6.3	CDK1
P55211	153	dlaYils	1JXQ	269	79	6.6	Abl
P04040	230	eavYckf	1QQW	499	100	6.8	Abl, Abl2
Q14934	676	rkrSqpq	2AS5	287	78	7.1	MAPK1, MAPK3, RSK-2
Q07497	601	lveYlkl	2OZO	306	53	7.6	EphB5
P35241	564	kykTlre	2I1K	581	73	7.7	ROCK group
P35241	573	kgnTkrr	2I1K	581	73	7.9	ROCK group
Q13882	342	dneYtar	2H8H	445	61	8.0	Brk
P17655	368	rrgStag	1KFX	699	91	8.4	PKA group
P61763	574	hilTpqk	1DN1	589	94	8.8	CDK group
P48025	130	vrdYvrq	2OZO	252	76	9.0	Lyn, Syk
P53355	289	rreSvvn	1WMK	311	90	9.4	RSK-2, RSK group
P22607	724	nelYmmm	2PSQ	299	91	9.8	FGFR3
P41743	271	driYamk	1ZRZ	340	90	9.8	Src

Table 4.2 (continued from previous page)

Figure 4.9: The average solvent accessible surface area values of known phosphorylation sites, without including sites from high throughput mass spectrometry experiments, as compared to non-phosphorylation sites. (A) Relative solvent accessible surface area of Ser/Thr/Tyr residues. (B) Relative solvent accessible area of heptapeptides containing Ser/Thr/Tyr as the central residue. (C) Absolute solvent accessible surface area of Ser/Thr/Tyr residues. (D) Absolute solvent accessible area of heptapeptides containing Ser/Thr/Tyr as the central residues.

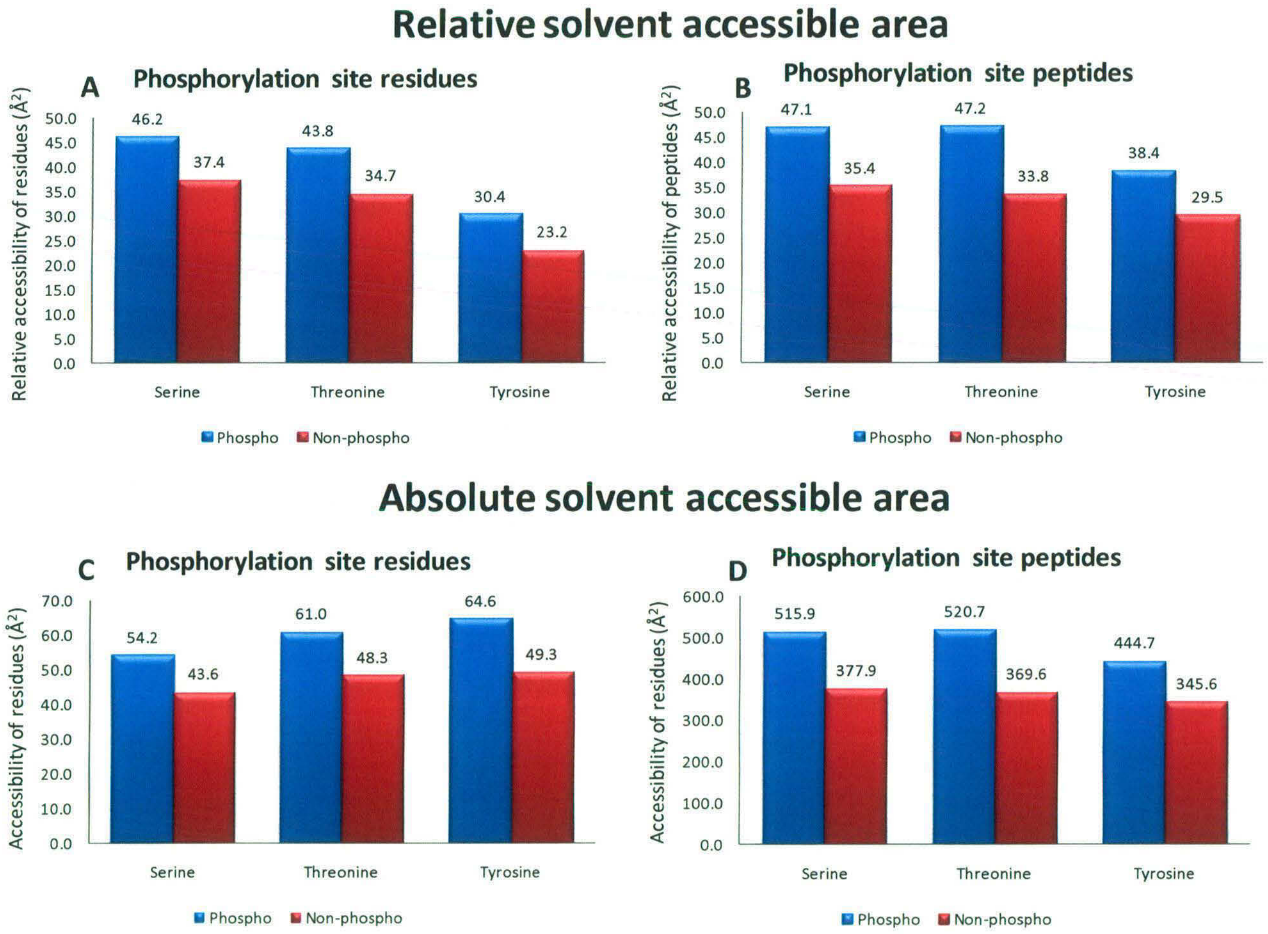

Relative solvent accessible area
A
Phosphorylation site residues
Relative accessibility of residues (Å²)
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
46.2
37.4
43.8
34.7
30.4
23.2
Serine
Threonine
Tyrosine
🟦 Phospho 🟥 Non-phospho
B
Phosphorylation site peptides
Relative accessibility of peptides (Å²)
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
47.1
35.4
47.2
33.8
38.4
29.5
Serine
Threonine
Tyrosine
🟦 Phospho 🟥 Non-phospho
Absolute solvent accessible area
C
Phosphorylation site residues
Accessibility of residues (Å²)
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
54.2
43.6
61.0
48.3
64.6
49.3
Serine
Threonine
Tyrosine
🟦 Phospho 🟥 Non-phospho
D
Phosphorylation site peptides
Accessibility of residues (Å²)
600.0
500.0
400.0
300.0
200.0
100.0
0.0
515.9
377.9
520.7
369.6
444.7
345.6
Serine
Threonine
Tyrosine
🟦 Phospho 🟥 Non-phospho

Figure 4.9

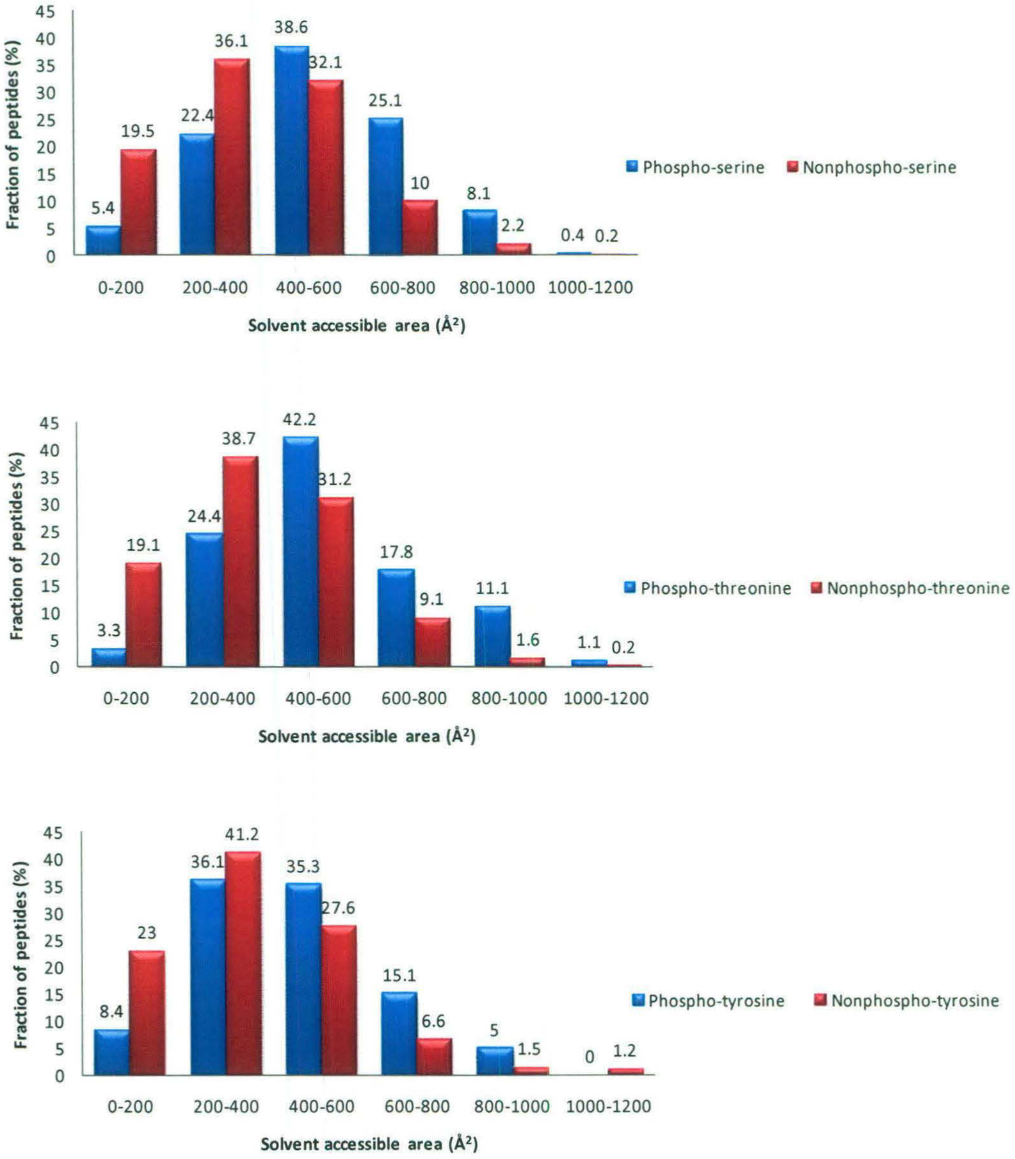
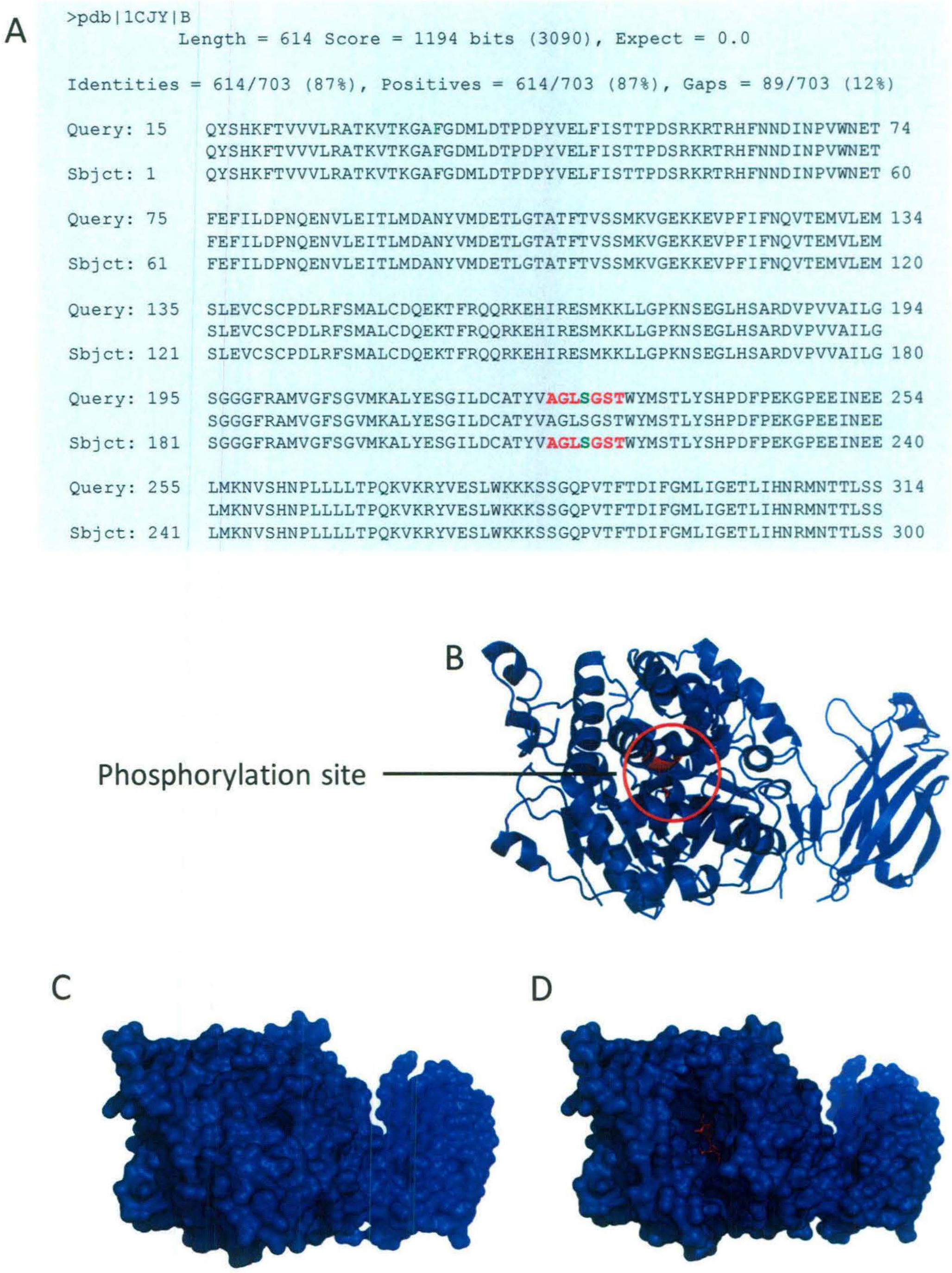


Figure 4.10: Distribution of solvent accessible surface area of phosphorylation site containing peptides, without including the sites identified from high throughput mass spectrometric experiments, as compared to nonphospho-site containing peptides.

Figure 4.11: An example of substrate protein cytosolic phospholipase A2 containing a phosphorylation site with very low solvent accessible area. The crystal structure of human cytosolic phospholipase A2 has been identified from the PDB by blast alignments. (A) Alignment of cytosolic phospholipase A2 (P47712) with the sequence of PDB entry 1CJY. (B) Cartoon representation of 1CJY. The Phosphorylation site and the residues surrounding it are shown in green and red respectively. (C) Surface representation of 1CJY showing that the phosphorylation site and the surrounding residues are buried in the protein. (D) The residues covering the phosphorylation site have been removed from the suface represention of 1CJY to show the buried phosphorylation site (shown in red).



```
>pdb|1CJY|B
          Length = 614 Score = 1194 bits (3090), Expect = 0.0

Identities = 614/703 (87%), Positives = 614/703 (87%), Gaps = 89/703 (12%)

Query: 15    QYSHKFTVVVLRATKVTKGAFGDMLDTPDPYVELFISTTPDSRKRTRHFNNDINPVWNET 74
             QYSHKFTVVVLRATKVTKGAFGDMLDTPDPYVELFISTTPDSRKRTRHFNNDINPVWNET
Sbjct: 1     QYSHKFTVVVLRATKVTKGAFGDMLDTPDPYVELFISTTPDSRKRTRHFNNDINPVWNET 60

Query: 75    FEFILDPNQENVLEITLMDANYVMDETLGTATFTVSSMKVGEKKEVPFIFNQVTEMVLEM 134
             FEFILDPNQENVLEITLMDANYVMDETLGTATFTVSSMKVGEKKEVPFIFNQVTEMVLEM
Sbjct: 61    FEFILDPNQENVLEITLMDANYVMDETLGTATFTVSSMKVGEKKEVPFIFNQVTEMVLEM 120

Query: 135   SLEVCSCPDLRFSMALCDQEKTFRQQRKEHIRESMKKLLGPKNSEGLHSARDVPVVAILG 194
             SLEVCSCPDLRFSMALCDQEKTFRQQRKEHIRESMKKLLGPKNSEGLHSARDVPVVAILG
Sbjct: 121   SLEVCSCPDLRFSMALCDQEKTFRQQRKEHIRESMKKLLGPKNSEGLHSARDVPVVAILG 180

Query: 195   SGGGFRAMVGFSGVMKALYESGILDCATYVAGLSGSTWYMSTLYSHPDFPEKGPEEINEE 254
             SGGGFRAMVGFSGVMKALYESGILDCATYVAGLSGSTWYMSTLYSHPDFPEKGPEEINEE
Sbjct: 181   SGGGFRAMVGFSGVMKALYESGILDCATYVAGLSGSTWYMSTLYSHPDFPEKGPEEINEE 240

Query: 255   LMKNVSHNPLLLLTPQKVKRYVESLWKKKSSGQPVTFTDIFGMLIGETLIHNRMNTTLSS 314
             LMKNVSHNPLLLLTPQKVKRYVESLWKKKSSGQPVTFTDIFGMLIGETLIHNRMNTTLSS
Sbjct: 241   LMKNVSHNPLLLLTPQKVKRYVESLWKKKSSGQPVTFTDIFGMLIGETLIHNRMNTTLSS 300
```

Figure 4.11

shift in the protein, before they can be accessed and phosphorylated by the protein kinases. For example, phosphorylation site containing peptides AGLSGST in cytosolic phospholipase A2 has been found to be buried as judged from its solvent accessibility area from crystal structure (Figure 4.11). It would be interesting to analyze the phosphorylation mechanism of such proteins. It would also be interesting to see if the binding of such substrates to protein kinase leads to the unfolding and binding of phosphorylation site, or interaction with other cellular factors unfolds the protein and makes the phosphorylation site accessible to kinase.

As different protein kinases are known to be differentially regulated, with each one having a unique mechanism of regulation, we investigated if certain specific kinase families are responsible for phosphorylation of these sites which are buried to a large extent. Out of 1088 phosphorylation sites mapped to crystal structures, 628 sites did not have any information about the kinase responsible for their phosphorylation. The remaining 460 sites are phosphorylated by 134 kinase types as reported in Phospho.ELM database. Among these mapped phosphorylation sites in the crystal structure templates, we calculated the average accessible surface area for various kinase types. However, sufficient number of phosphorylation instances were not

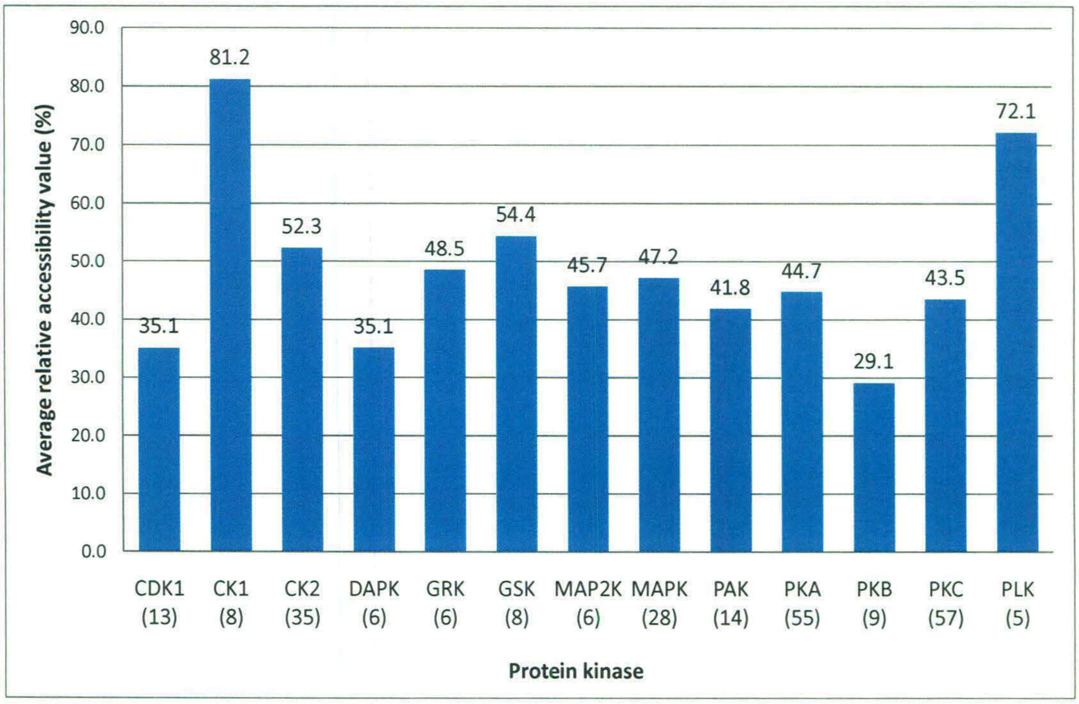


Figure 4.12: Average relative solvent accessible areas for various protein kinase groups. Only those kinase groups have been shown which had more than 5 (values in brackets below the kinase group name) phosphorylation instances mapped to a structural template.

mapped for most of the protein kinases. Therefore, we grouped together the phosphorylation instances assigned to similar kinases belonging to a specific group. We selected those groups which contained more than five phosphorylation instances. For this study, we considered only Ser/Thr kinases. A total of 237 phosphorylation instances could be grouped into 13 Ser/Thr kinase groups containing 56 individual kinases. Figure 4.12 and Figure 4.13 show the average accessibilities, and distribution of accessibilities of phosphorylation sites respectively for these kinase groups. CDK1, DAPK and PKB are three groups whose substrates have lowest average surface accessible areas among all kinases in our study. Based on this, it is tempting to speculate that, these kinase groups might employ a mechanism involving slight unfolding/conformational shift in their substrate proteins. On the other hand, phosphorylation sites in substrates of some kinases like CK1, GSK and PLK show a very high average accessible area, suggesting that the substrates of these kinases can be easily accessed by the protein kinase catalytic site.

4.4 DISCUSSION

In order to understand how protein kinases maintain the specificity towards the selection of their substrates, it is imperative to investigate the factors other than sequence and structural complimentary of the binding region of the substrates. Most phosphorylation site prediction methods take into consideration only the sequence or structure of a short peptide stretch flanking the phosphorylatable amino acid. However, in reality the protein kinases phosphorylate a folded and structurally intact substrate protein in its functional form. Therefore, solvent accessibility of the phosphorylatable amino acid and its flanking residues is a major factor, which is expected to affect the phosphorylation of a substrate protein. In this chapter, we have analyzed the solvent accessible areas of known phosphorylation sites in various substrate proteins of different kinases. Solvent accessible surface areas of phosphorylation sites have been carried out in the crystal structures of the substrate protein or their structural homologues. The results of our analysis indicate that, the phosphorylation site residues are significantly more exposed to the solvent as compared to the other sites containing Ser/Thr/Tyr residues. Our results are in agreement with earlier studies which have also suggested that phosphorylation sites are mostly found in regions of proteins that are likely to adopt loop or coil secondary structural states and prefer to be exposed to the solvent (Gnad *et al.*, 2007). Our

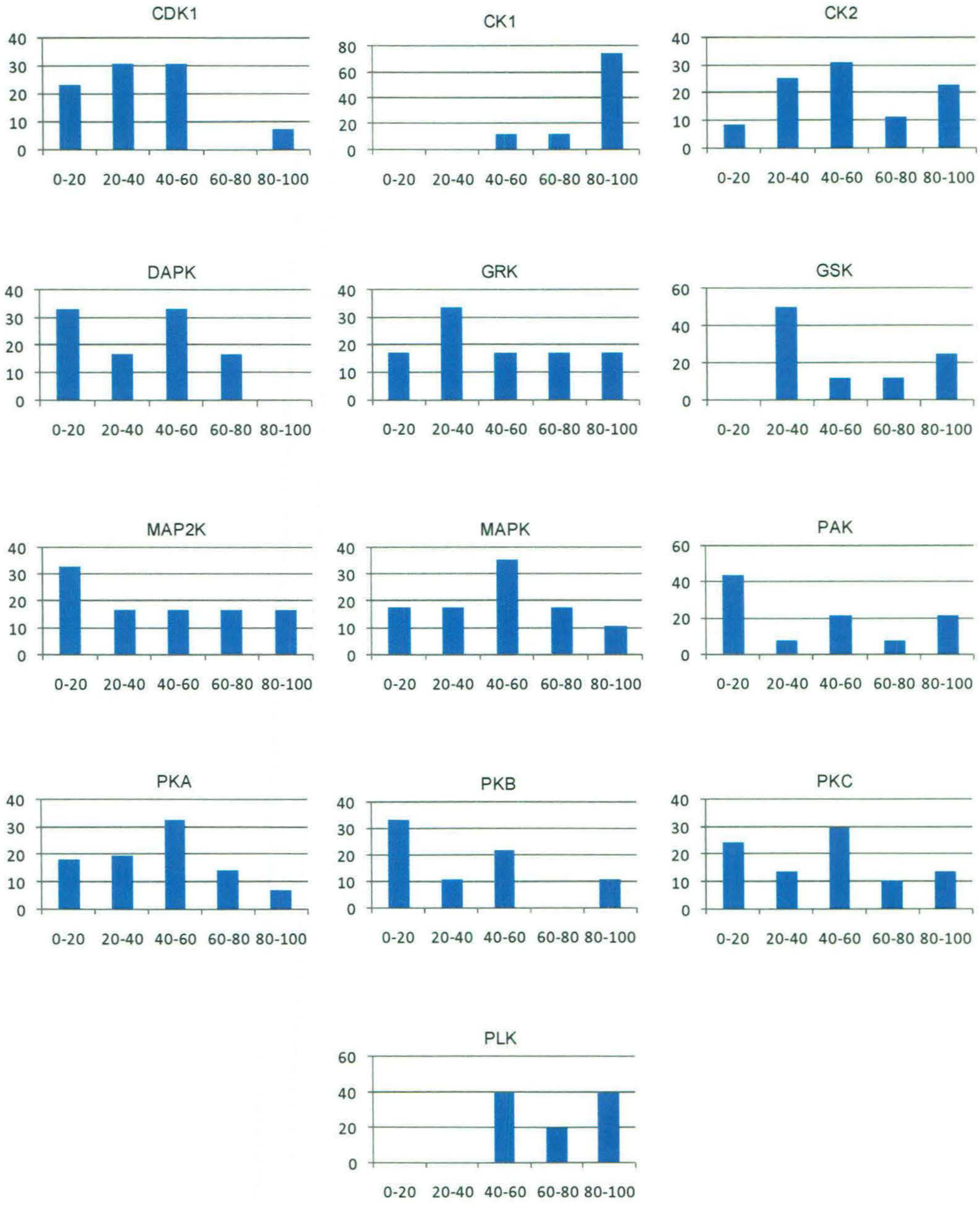


Figure 4.13: Distribution of relative solvent accessible surface areas of phosphorylation sites of various protein kinases as calculated by NACCESS program from the crystal structures of homologous proteins of their substrates. On each panel, X-axis has relative accessible area, and Y-axis has the number of substrates for the protein kinase.

current results based on the actual accessibility values from crystal structures reaffirm the same.

Although, in general, the phosphorylation sites are more exposed than their non-phosphorylatable counter parts, our analysis has revealed few interesting examples of substrate proteins on which the phosphorylation sites have a very low solvent accessible surface area indicating that they are buried inside. Such sites cannot be easily phophorylated by the protein kinases in the absence of a considerable conformational change in the structure of the substrate proteins. It will be interesting to analyze such protein individually in detail to find out how they are phosphorylated.

In this work, the accessibilities of the phosphorylation sites have been calculated from the crystal structure or the structural homologs. The use of homologs for surface area calculation is appropriate as structures are known to be more conserved than sequences and the homologues have been identified based on significant sequence similarity. In addition, given the biological importance of the phosphorylation event, the phosphorylation sites are also likely to be conserved. Therefore, the probability of finding the phosphorylation site on a substrate protein in same structural context as in the structural match is very high, and the surface areas calculated from it are likely to be true representative of the surface area in the actual substrate proteins.

The results from analysis of accessibilities of phosphorylation sites suggests that, the inclusion of the solvent accessibility terms in the phosphorylation site prediction programs might help in improving their prediction accuracy. However, our observation of several phosphorylation sites at very low accessibility values suggests that it would be difficult to fix a deterministic criteria based on accessibility value for identifying potential phosphorylation sites. Accessibility and structural flexibility of the potential phosphorylation sites can probably be combined with interactions between phosphorylation site and kinase to develop more powerful structure based methods for prediction of substrates for kinases.

Chapter Five

Structure based prediction of MHC binding peptides using rotamer library and statistical pair potentials

5.1 INTRODUCTION

Major histocompatibility complex (MHC) proteins are present on the surface of antigen presenting cells. Their function is to bind processed peptides and provide a continuous update of cellular and environmental composition for the scrutiny by T-cell receptors (TCR) on the surface of cytotoxic T-lymphocytes (CTL). T-cell receptors recognize the antigenic peptides in complex with the MHC molecule and trigger the immune response (Pamer and Cresswell, 1998). The resulting immune response plays a central role in defending the body against foreign pathogenic invasion. MHC class I molecules bind the peptides of cytosolic origin and presents them to CD8$^+$ T-cells which in turn kill the infected cells. MHC class II molecules bind the processed peptides derived from the endocytosed proteins, and present them to CD4$^+$ T-cells which triggers humoral and cellular responses against the foreign organisms. Of all the peptides present in the antigenic proteins from the pathogen, only a small set of peptides can bind and generate the immune response (Pamer and Cresswell, 1998). Binding of peptides to the MHC molecules is the most selective step in the antigen presentation, and hence the identification of such peptides has important implications in rational vaccine design and better understanding of immune system. Experimental approaches for identifying T-cell epitopes in antigens of viral or bacterial origin typically involve synthesis of overlapping peptides and assays for their binding to MHC proteins (Wilson *et al.*, 1999). Approaches involving combinatorial libraries have also been extensively used to study binding of peptides to MHC molecules (Doytchinova *et al.*, 2004; Sung *et al.*, 2002; Udaka *et al.*, 2000). However, the genetic diversity of MHC system is characterized by extensive degree of allelic polymorphism (Pamer and Cresswell, 1998; Parham *et al.*, 1995), making experimental approaches for identifying the antigenic peptides practically tedious, time consuming and cost-intensive if not impossible.

Availability of a large volume of data on peptide binding preference of various MHC alleles has facilitated development of automated computer programs for prediction of peptides which can bind to various MHC molecules. Such programs can carry out a systematic scanning of the antigen protein sequence for candidate T-cell epitopes. Some of the early computational methods were based on identification of conserved motifs in a set of known MHC binding peptides. These binding motifs characterized the peptide specificity of different MHC alleles in terms of dominant anchor positions with a strong preference for a restricted set of amino acids (Sette *et al.*, 1989). Even though motif based scanning can identify MHC binding peptides in large number of cases, it was soon discovered that they produce many false positive and probably also a similar number of false negatives. Secondly, a simplistic classification of a peptide as binder or non-binder based on presence or absence of a motif is not adequate in cases where different peptides bind to a given MHC allele with different affinities. These shortcomings led to the development of so-called matrix based methods (Parker *et al.*, 1994; Sturniolo *et al.*, 1999). BIMAS (Parker *et al.*, 1994), ProPred (Singh and Raghava, 2001), ProPred1 (Singh and Raghava, 2003), MMBPred (Bhasin and Raghava, 2003) and EPIPREDICT are matrix based epitope prediction methods that are similar to conventional bioinformatics analysis based on sequence profiles. RANKPEP (Reche *et al.*, 2002; Reche *et al.*, 2004) is a method that utilizes position specific scoring matrices (PSSM) for predicting binding peptides. SYFPEITHI (Rammensee *et al.*, 1999) is another widely used weighted matrix based program. More sophisticated mathematical models employing quantitative structure activity relationship (QSAR) have also been used in programs such as MHCPred (Guan *et al.*, 2003a). These are methods based on powerful algorithms used in machine learning and artificial intelligence. nHLAPred, and NetMHCpan 2.0 (Nielsen *et al.*, 2007) are two methods which use artificial neural networks (ANN). NetMHC 3.0 (Buus *et al.*, 2003; Nielsen *et al.*, 2003; Nielsen *et al.*, 2004) is a method which uses PSSM also, along with ANN for making predictions. Support vector machines (SVM) based models are employed by two methods, KISS (Jacob and Vert, 2008) and SVMHC (Donnes and Elofsson, 2002). These methods in general have better prediction accuracy compared to motif based methods due to their ability to take into account complex correlations between various positions. However, their dependence on large amount of allele specific experimental data for training remains a major limitation as these methods can not be used for newer alleles for which experimental information on peptide binding is not available. They also often fail to detect MHC binding peptides with non-canonical motifs. Thus, the prediction performance of all these sequence based methods is limited by the amount and nature of training data as well as potential weaknesses in their methodology.

In contrast to the large number of sequence based methods for prediction of MHC binding peptides, there are relatively fewer structure based approaches. These structure based methods attempt to predict the MHC binding peptides based on structural modeling of the bound conformation of the peptide in the MHC binding pocket and ranking various peptides as per their interaction energy with the MHC. The crystal structures of various MHC-peptide complexes are used as structural templates for modeling. Such approaches can in principle help in understanding the structural basis of allele specific MHC binding. Even though several variants of structure based approaches involving, conformational search, ligand docking (Tong *et al.*, 2004) and molecular dynamics simulations (Pohlmann *et al.*, 2004) have been reported, the computational complexity has limited these studies to few MHC-peptide complexes. Therefore, these methods have not been used for fast automated scanning of antigen sequences. One structure based method PREDEP (Schueler-Furman *et al.*, 2000), which permits fast automated scanning for potential MHC binding peptides uses a threading type approach. In this approach, the peptide structure in the MHC groove is used as a template upon which query peptide sequences are threaded. The compatibility between threaded peptide sequence and MHC binding pocket is evaluated by statistical pair potential derived from the analysis of amino acid packing in protein structures. Earlier studies (Altuvia *et al.*, 1997) using PREDEP (Schueler-Furman *et al.*, 2000) have demonstrated that, this method successfully identifies good binders only for MHC molecules with hydrophobic binding pockets, probably because of high emphasis on the hydrophobic contacts in the statistical pair potentials. In view of these results, alternate scoring methods have been proposed for better identification of ligands for MHC molecules with hydrophilic pockets. Apart from PREDEP (Schueler-Furman *et al.*, 2000), MHC-Thread (http://www.csd.abdn.ac.uk/~gjlk/MHC-Thread/) and adaptive double threading approach (Zaitlen *et al.*, 2008) are two other structure based methods which predict potential MHC-binding peptides by using threading principles for scoring the compatibility of the peptide with the MHC.

Even though sequence based *in silico* prediction tools are widely used for identifying MHC binding peptides on proteins, they can not be applied in case of MHC alleles for which no experimental data is available. Discovery of newer MHC alleles require development of novel computational methods which can give reliable clues about their substrate specificities even in absence of extensive experimental data. Structure based substrate prediction methods can in principle address these problems. High homology between various MHC sequences and conservation of structural fold despite divergence in sequence make them ideal candidates for application of structure based prediction methods. However, the survey of available literature indicates that potential of structure based methods have not been exploited to the full extent. The primary reason for the lack of reliable structure based substrate prediction method is the computational complexity associated with structural modeling of the protein-peptide complexes. Prediction methods based on the evaluation of the binding energy between a protein structure and a given peptide ligand require reliable scoring functions as well as adequate CPU time for efficient conformational search. Therefore, structure based methods like MHCPRED, MHC-thread etc. use structural information to a limited extent and only consider the binding pocket environment based on 3D structure of the MHC, rather than structural modeling of the substrate in the binding site.

The binding of peptides in the MHC groove is a problem of recognition of a peptide ligand by a protein receptor. Computational approaches have proved to be very useful in the identification of ligands for proteins (Hou *et al.*, 2006; Martin *et al.*, 2006). Recent work in the field of computational protein design (Mooers *et al.*, 2003; Oelschlaeger *et al.*, 2005) have demonstrated that use of rotamer library is a powerful approach for the rapid and accurate modeling of the amino acid side chains in the context of a given structural environment. This method has been employed successfully for redesigning a number of proteins. It has been used for the redesigning of the calmodulin protein for increased affinity towards its target protein by optimizing the binding interface (Shifman and Mayo, 2002). Recently, the same approach was used for predicting the metallo-beta-lactamase mutants that produced enhanced catalytic activity (Oelschlaeger *et al.*, 2005). Rotamer library has also been employed successfully to design the entire sequence for the small proteins (Dahiyat and Mayo, 1997), to design protein variants with improved stability (Malakauskas and Mayo, 1998; Marshall *et al.*, 2002; Street *et al.*, 2000), and to induce ligand binding in a large protein by stabilizing its active conformation (Shimaoka *et al.*, 2000).

Therefore, in this chapter, we report the development and benchmarking of a structure based substrate prediction method for MHC molecules using rotamer library approach. The methodology used is a multi scale approach, in which threading of putative peptides on the structural template of MHC-peptide complex crystal structure has been used to identify the high scoring peptides by scoring them by residue based statistical pair potential (Betancourt and Thirumalai, 1999; Miyazawa and Jernigan, 1996). Good binders, as judged by their pair-potential scores, are modeled in binding groove of their respective MHC protein by rotamer library approach and ranked as per their binding free energies calculated by MM/PBSA method. The results were benchmarked on the experimental data on substrate peptides of class I and class II MHCs cataloged in SYFPEITHI database. The modeling of peptide-MHC complexes have been carried out using the MODPROPEP which was described in chapter 2. In this chapter, we describe the results on benchmarking and evaluation of the prediction accuracy.

5.2 METHODS

5.2.1 Modeling MHC-peptide complexes from protein sequences of MHC alleles

The protein sequences of all characterized MHC alleles belonging to class I MHC, and class II MHC were obtained from IMGT/HLA database at EBI (http://www.ebi.ac.uk/imgt/hla/) (Robinson *et al.*, 2003). A total of 1591 and 746 alleles are currently available for class I and class II MHCs respectively (Table 5.1 and Table 5.2). The crystal structures of MHC-peptide complexes were also downloaded from the PDB website (http://www.rcsb.org) (Berman *et al.*, 2000). All the crystal structures were classified according to their allele and MHC class. Crystal structures containing bound unnatural synthetic peptides were removed from the

Table 5.1: List of human class I MHC allele sequences obtained from IMGT/HLA database.

Serial Number	Class I MHC Allele	Number of Alleles
1	HLA-A	494
2	HLA-B	833
3	HLA-C	264
Total		**1591**

Table 5.2: List of human class II MHC allele sequences obtained from IMGT/HLA database.

Serial Number	Class II MHC Allele	Number of Alleles
1	DRA	2
2	DRB	524
3	DQA1	25
4	DQB1	66
5	DPA1	15
6	DPB1	114
Total		**746**

dataset. A total of 148 structures are currently available for 32 class I MHC alleles (Table 5.3), while there are 42 structures for 10 class II alleles (Table 5.4). Apart from these 190 human MHC crystal structures, 73 murine MHC crystal structures belonging to 10 different alleles are also available in our structural library (Table 5.5). These crystal structures were used as templates for modeling of MHC-peptide complexes in case of alleles for which crystal structures were not available. In order to model a given MHC allele, the most homologous MHC crystal structure, as identified by BLAST alignment, is used as structural template. The backbone coordinate of the structural template is kept fixed and the side chains of the query allele are modeled using SCWRL (Canutescu *et al.*, 2003) program. SCWRL uses backbone dependent rotamer library approach for side chain fixing. Similarly, while modeling putative substrate peptides in the binding pocket of a given MHC allele, the backbone coordinates of the bound peptide in the original template structure are preserved and side chains are modeled using SCWRL. This procedure generates a MHC-peptide complex in which the MHC part is that of query allele modeled on the backbone of homologous structure, and peptide is exactly in the same backbone conformation as in the template peptide. All these steps are carried out by using MODPROPEP (Kumar and Mohanty, 2007) program.

Table 5.3: List of human class I MHC-peptide complexes obtained from PDB.

Serial Number	Class I MHC Allele	Number of Structures
1	A*0101	1
2	A*0201	66
3	A*0265	7
4	A*0312	2
5	A*0326	5
6	A*1101	4
7	A*6801	1
8	B*0733	1
9	B*0801	2
10	B*0802	1
11	B*1505	3
12	B*2705	1
13	B*2709	4
14	B*2713	4
15	B*3501	1
16	B*3508	5
17	B*3542	11
18	B*3709	2

Serial Number	Class I MHC Allele	Number of Structures
19	B*4402	1
20	B*4403	2
21	B*4405	1
22	B*5101	2
23	B*5301	2
24	B*5703	3
25	B*5711	1
26	B*5815	1
27	Cw*0304	1
28	Cw*0401	2
29	Cw*0602	1
30	Cw*0741	2
31	Cw*1202	7
32	Cw*1802	1
Total		**148**

Table 5.4: List of human class II MHC-peptide complexes obtained from PDB.

Serial Number	Class II MHC Allele	Number of Structures
1	HLA-DRB1*0301	1
2	HLA-DRB1*0101	23
3	HLA-DRB1*1501	2
4	HLA-DRB1*0401	2
5	HLA-DRB3*0101	1
6	HLA-DRB5*0101	4
7	HLA-DQA1*0501_HLA-DQB1*0201	1
8	HLA-DQA1*0302_HLA-DQB1*0302	3
9	HLA-DQA1*0102_HLA-DQB1*0602	1
10	HLA_DQA1*0303_HLA_DQB1*0310	4
Total		**42**

Table 5.5: List of mouse class I and class II MHC-peptide complexes obtained from PDB.

Serial Number	Mouse Allele	MHC Class	No of Structures
1	H2-D	Class I	1
2	H2-Db	Class I	16
3	H2-Dd	Class I	3
4	H2-Kb	Class I	34
5	H2-Ab	Class II	2
6	H2-Ad	Class II	2
7	H2-Ak	Class II	4
8	H2-A-nod	Class II	2
9	H2-Ek	Class II	8
10	H2-E	Class II	1
Total			**73**

5.2.2 Conformational analysis of peptides bound in the MHC binding groove in crystal structures

MHC proteins bind and stabilize the antigenic peptides in the peptide-binding groove formed between two alpha-helices on the floor of antiparallel beta-sheets. Availability of a large number of crystal structures in our structural library enabled us to perform a systematic analysis of the binding conformation of peptides in complex with the MHC proteins to investigate whether all peptides adopt similar extended conformation when they bind to the MHC molecule. We removed the peptide coordinates from the MHC-peptides complexes and grouped them in two broad groups belonging to class I and class II MHCs. Each group consisted of a set of peptides of different lengths in the same conformations as they were in complex with MHC proteins. Class I MHC set contained 8 mer, 9 mer, 10 mer and 11mer peptides, while the class II MHC set contained peptides up to 16mer in length. The long class II MHC peptides usually contained the overhanging regions on C and N terminal of peptides which did not make any contact with the MHC proteins. Such peptides had been already trimmed to remove the overhanging regions. For each set of peptides, the longer peptides are broken down into all possible peptides of shorter lengths. For example, a 10 mer peptide is broken into 2 overlapping 9mers, and 3 overlapping 8mers. For each MHC class groups, the peptides of same length were compared and their backbone RMSD values were calculated after their optimum superimposition.

The orientation of the peptides in the MHC binding pocket were also compared for various alleles of class I as well as class II MHCs. For a pair of MHC-peptide complex, the structures were superimposed using only the Cα atoms of MHC protein coordinates. This brought up the coordinates of structure bound peptides in a common frame of MHC protein. The peptide coordinates were used to calculate the RMSD values of peptides pair without superimposition. This procedure was repeated for all possible pairs which could be compared for each allele.

5.2.3 Scoring the binding energy in MHC-peptide complexes by statistical pair potentials

The interacting residue pairs between peptide and the MHC protein were identified using the criteria of any two atoms of the residue pair being at a distance less or equal to 4.5Å. For each interacting residue pair, a score was assigned using the residue based statistical pair potentials. Binding energy was computed by summing the scores for all possible interacting residue pairs present in the complex.

All these steps were also carried out using MODPROPEP. Two different statistical pair potentials, namely Miyazawa-Jernigan (MJ) (Miyazawa and Jernigan, 1996) and Betancourt-Thirumalai (BT) (Betancourt and Thirumalai, 1999) were used. For scanning a potential antigenic protein for putative MHC binding peptides, all possible peptides of length same as that of template bound peptide, were ranked as per their binding energy values.

5.2.4 Scoring of crystal structure bound peptides from their source proteins

If crystal structure is available for a peptide in complex with the MHC, the peptide is generally a preferred substrate for the corresponding MHC. Since a typical antigenic protein contains only a few MHC binding peptides, it is reasonable to assume that, the bound peptide has favourable interactions with the MHC, while all other peptides in the corresponding antigenic protein have unfavourable interactions with the particular MHC allele. Therefore, a predictive computational approach should be able to discriminate the structure bound peptide among all other peptides from its parent protein. In order to check the predictive power of the statistical pair potentials, we identified the source proteins for each of the peptides bound in the MHC-peptide complexes in our structural library. These protein sequences were downloaded from the SWISSPROT database (Boeckmann *et al.*, 2003). A few complexes in the library contained unnatural or synthetic peptide bound to MHC proteins. Such complexes were removed from the dataset for this exercise (Tables 5.6 and 5.7). The amino acid sequences of each of these source proteins were scanned for all possible overlapping peptides. All overlapping peptides were modeled in the peptide binding groove of MHC allele by SCWRL program using the backbone dependent rotamer library approach. The interaction energy for each peptide was calculated using the BT and MJ residue-residue statistical pair potential matrices. Apart from these two statistical pair potentials, each peptide MHC complex built by SCWRL was also ranked as per penalty for steric clashes i.e. peptide having minimal or no steric clash being preferred binder. All modeled peptides were ranked as per their interaction energy values. The prediction was considered to be correct if the template bound peptide was ranked among the top 30% of the peptides from its source protein in terms of interaction energy score. The procedure is repeated for all the MHC-peptide template structure present in the structural library using the program

MODPROPEP. For each of the three scoring methods, the percentage of MHC-peptide complexes for which the above mentioned procedure gives the correct prediction was calculated.

Table 5.6: List of crystal structures of class I MHC alleles which bind a peptide whose source protein could be identified in the SWISSPROT database.

Serial Number	**MHC Allele**	**Number of structures**
1	A*0101	1
2	A*0201	17
3	A*0265	1
4	A*0312	1
5	A*0326	2
6	A*1101	4
7	A*2402	1
8	A*6801	1
9	B*0733	1
10	B*0801	1
11	B*1505	3
12	B*2705	1
13	B*2709	3
14	B*2713	2
15	B*3501	1
16	B*3508	4
17	B*3542	3
18	B*3709	1
19	B*4402	1
20	B*4403	1
21	B*4405	1
22	B*5101	2
23	B*5301	2
24	B*5703	3
25	B*5711	1
26	Cw*0304	1
27	Cw*0602	1
28	Cw*0741	3
29	Cw*1202	3
30	Cw*1802	1
31	E*0101	1
32	G*0101	1
33	G*0104	1
34	H2-Db	5
35	H2-Dd	1
36	H2-Kb	7
Total		**84**

Table 5.7: List of crystal structures of class II MHC alleles which bind a peptide whose source protein could be identified in the SWISSPROT database.

Serial Number	**Class II MHC allele**	**Number of structures**
1	H2-Ad	1
2	H2-Ak	3
3	H2-Ek	2
4	HLA-DQA1*0102_HLA-DQB1*0602	1
5	HLA-DQA1*0302_HLA-DQB1*0302	3
6	HLA-DRB1*0101	7
7	HLA-DRB1*0401	1
8	HLA-DRB1*1501	1
9	HLA-DRB5*0101	3
Total		**22**

5.2.5 Benchmarking of MODPROPEP on peptide binding data obtained from SYFPEITHI database

MHC binding peptides catalogued in SYFPEITHI (Rammensee *et al.*, 1999) database were used to estimate the prediction accuracy of MODPROPEP. SYFPEITHI contains experimentally verified peptides known to bind with a number of different alleles of class I and class II MHC proteins. Most of the alleles bind to peptides of different length. We identified those alleles which have at least 20 substrate peptides in the database. A total of 30 such alleles (Table 5.8) were selected for testing the prediction accuracy of MODPROPEP. These 30 alleles included 28 class I MHC alleles and 2 class II MHC alleles. All of these alleles together bind to 1654 peptides as reported in SYFPEITHI database. The source proteins of these peptides were downloaded from SWISSPROT database. If the crystal structure was not available for a given allele, we modeled it using a homologous MHC crystal structure. The length of the bound peptide in the structural template was identified and the source protein containing the substrate peptide was broken into overlapping peptides of the same length as the template bound peptide. All these overlapping peptides were modeled in the binding pocket of the MHC allele and binding energy values were calculated using BT matrix. All peptides are sorted as per their binding energy and assigned a rank. The prediction was considered to be correct, if the known substrate peptide cataloged in SYFPEITHI was ranked among top 30%. The procedure is repeated for all source proteins containing known substrate peptides of a given MHC allele and prediction accuracy was calculated. Similar procedure was followed for evaluating the prediction accuracy of each of the 30 alleles listed in Table 5.8.

Table 5.8: List of MHC alleles, which have at least 20 known binding peptides catalogued in the SYFPEITHI database. These alleles were used for the calculation of accuracy of MODPROPEP for prediction of epitopes of MHC.

S. No	MHC Allele	No of epitopes in SYFPEITHI	Template used for prediction/modeling
1	HLA-A*0201 (10)	74	1HHH
2	HLA-A*0201 (9)	347	1AKJ
3	HLA-A*1101 (9)	38	1Q94
4	HLA-A*2402 (9)	28	2BCK
5	HLA-A*2601 (9)	57	1Q94
6	HLA-A*6801 (9)	25	1AKJ
7	HLA-B*0702 (9)	29	1M05
8	HLA-B*0801 (9)	21	1M05
9	HLA-B*1501 (9)	62	1A9B
10	HLA-B*1801 (9)	111	1A9B
11	HLA-B*2703 (9)	23	1HSA
12	HLA-B*2704 (9)	38	1HSA
13	HLA-B*2705 (9)	115	1OGT
14	HLA-B*2705 (10)	40	1JGD
15	HLA-B*2706 (9)	33	1K5N
16	HLA-B*2709 (9)	43	1K5N
17	HLA-B*3701 (9)	22	1SYV
18	HLA-B*3901 (9)	76	1M05
19	HLA-B*4001 (9)	24	1M05
20	HLA-B*4402 (9)	42	1M6O
21	HLA-B*4701 (9)	20	1N2R
22	HLA-B*4901 (9)	100	1E27
23	HLA-B*5101 (9)	33	1E27
24	HLA-Cw*0401 (9)	48	1QQD
25	HLA-DRB1*0401 (14)	45	1BX2
26	H2-Db (9)	35	1BZ9
27	H2-Kb (8)	46	1BQH
28	H2-Kd (9)	38	1BZ9
29	H2-Kk (8)	21	1BQH
30	H2-Ab (15)	20	1MUJ

5.2.6 Re-ranking of high scoring peptides using MM/PBSA

The high scoring 30% peptides from each of the test proteins containing known substrates were re-ranked using the all atom force-field and MM/PBSA method following a procedure similar to that described in chapter 3. AMBER9 molecular dynamics suite of programs was used for these calculations. The models of MHC-peptides complexes involving the top ranked 30% peptides were energy minimized using the sander program of AMBER9 suite. The binding free energy of these minimized complexes were evaluated using the MM/PBSA module of AMBER9. MHC-peptide complexes were rearranged as per their binding free energies. This procedure was repeated only for those MHC alleles, whose prediction accuracy was more than 60% as evaluated by Betancourt and Thirumalai pair potential matrix.

5.2.7 Receiver operating characteristic curve (ROC) calculations

The ROC curves were calculated for the predictions by BT matrix as well as MM/PBSA approach. AUC, specificity and sensitivity values were estimated for the predictions carried out for each of the MHC alleles. Here again the procedure followed was similar to the one described in chapter 3.

5.3 RESULTS

5.3.1 Analysis of bound conformations of antigenic peptides in the crystal structures

In order to identify the degree of conservation in the binding modes of antigenic peptides in the MHC-peptide crystal structures, the peptide conformations in the individual crystal structure complexes were superimposed over each other to calculate their RMSDs. Since, the bound peptides were of different length, the larger peptides were truncated to all possible peptides of smaller length bound in the other crystal structure complexes. For example, the peptides bound to class I MHC vary in length from octamers to decamers. Hence, all decamer peptides were fragmented as 3 overlapping octamers and two overlapping nonamers. Figure 5.1A shows a distribution of the RMSD values of all possible peptide pairs of same length. Peptides of 8 amino acid length, in complex with class I MHC, bind in a very similar conformation as 89.3% of all pairs show RMSD values less than 1Å. Peptides of 9 amino acid length showed slightly greater deviations as their distribution showed

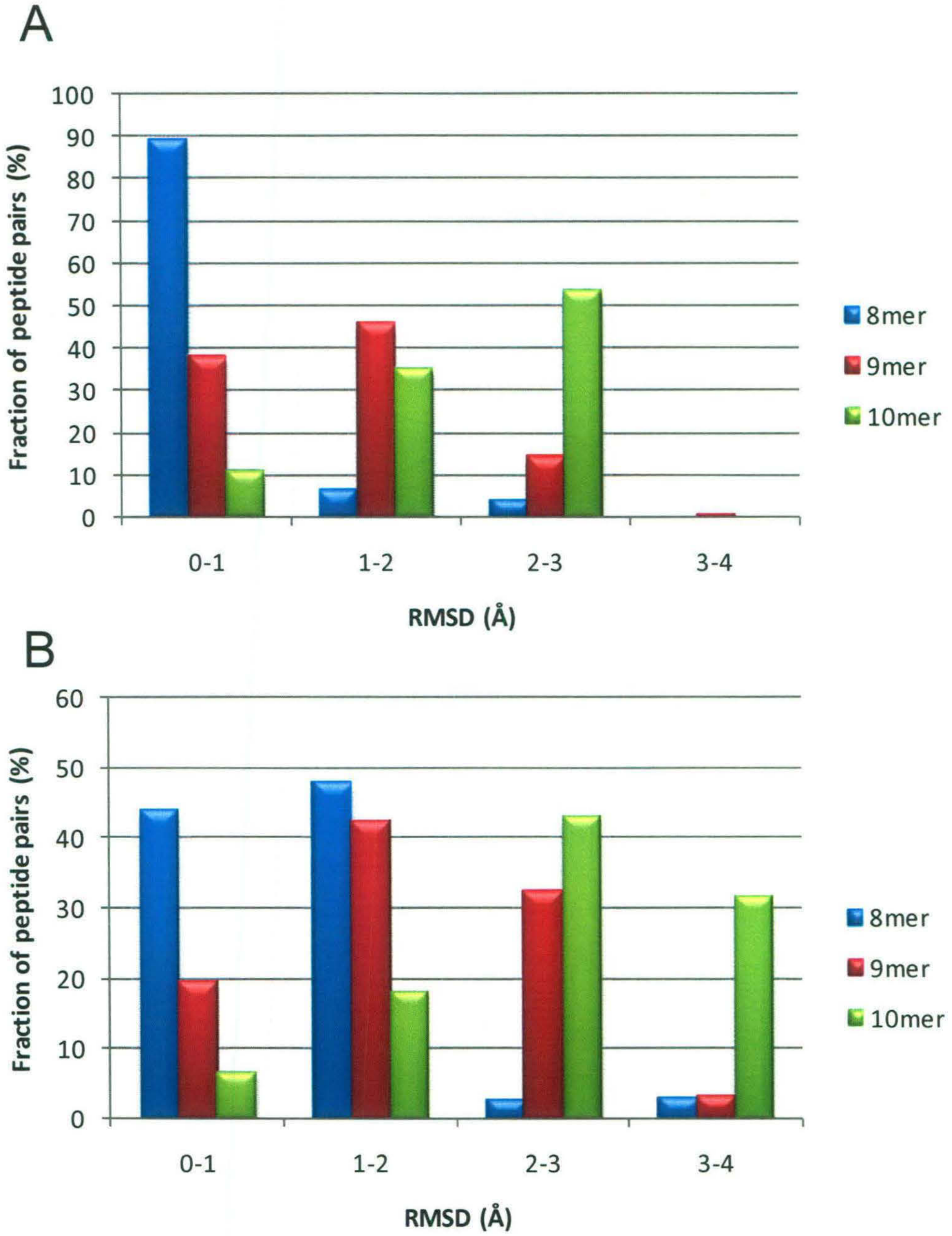


Figure 5.1: Distribution of backbone RMSD values when the peptide bound in MHC-peptide complexes were compared to each other. All peptides of same length were superimposed on each other and the RMSD value of every possible pair was calculated.

(A) Distribution for class I MHC bound peptides. **(B)** Distribution for class II MHC bound peptides.

46.2% peptide pairs show a RMSD deviation in the 1-2Å range. A small fraction of the pairs had deviation in the 3-4Å range. However, 38.4% have RMSD in the range of 1-2Å, thus a majority of 84.6% peptide pair had the RMS deviation less than 2Å. The peptides of length 10, however, showed a greater deviation than 8mer and 9mer peptides.

In case of Class II MHC (Figure 5.1B), bound peptides showed conservation in their binding modes for 8mer and 9mer peptides. The RMSD deviation for 8mer peptides was less than 1Å for 44.2% of all peptide pairs. An additional 48% of peptide pairs showed RMSD values in 1-2Å range. Only a small fraction of 5.7% peptide pairs showed as RMSD values more than 2 Å. Peptides of 9 mer length were less conserved in their binding modes than 8mer peptides. A fraction of 62.6% peptide pairs showed a RMSD of less than 2Å. The 10 mer peptides in case of MHC class II showed less conserved binding conformations. A significant fraction of 74.8% peptide pairs had an RMSD values more than 2Å.

We also carried out analysis of the orientations of the peptides in the binding pockets of MHC molecules. RMS deviations in the peptide coordinates were calculated, when only MHC part of the complexes were superimposed to bring the peptides to the same reference frame. Figure 5.2 shows an analysis of the orientations of the peptides bound to class I MHC allele A*0201. Our structure library had 61 complexes of A*0201 allele which have a 9 mer peptide bound in the MHC-complex. All of these crystal structures were superimposed over each other using the coordinates of Cα atoms of MHC protein alone. The peptide coordinates in the superimposed structures were used to calculate the deviation of every peptides with every other peptide. The clustering of the peptides using their RMSD values showed that all the peptides were bound in the similar conformation and at the same site, as they clustered together (Figure 5.2A). However, one peptide was bound in a slightly different conformation (shown in red color in Figure 5.2A). The distribution of the RMSD value of the peptides also confirmed this observation. Figure 5.2C shows the binding modes of three representative 8 mer, 9 mer and 10 mer peptides to the MHC allele A*0201. In order to show all three peptides in the same frame, the corresponding MHCs have been superposed and only a single MHC structure is shown for clarity. As can be seen, the 8mer peptide binds in a conformation very similar to the 9mer, but the 10mer peptide binds in a slightly different conformation. The difference in the conformation of the 10 mer can probably be attributed to the

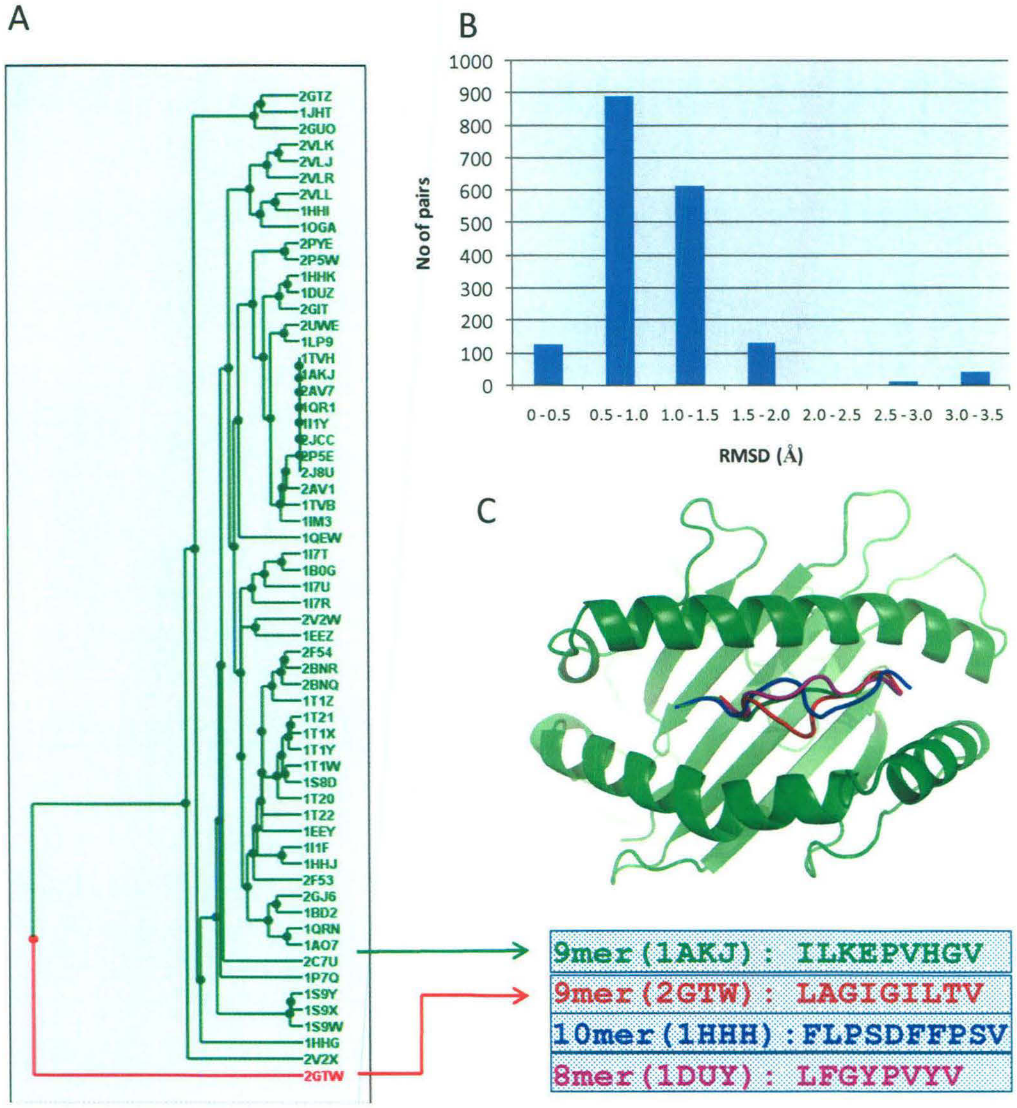
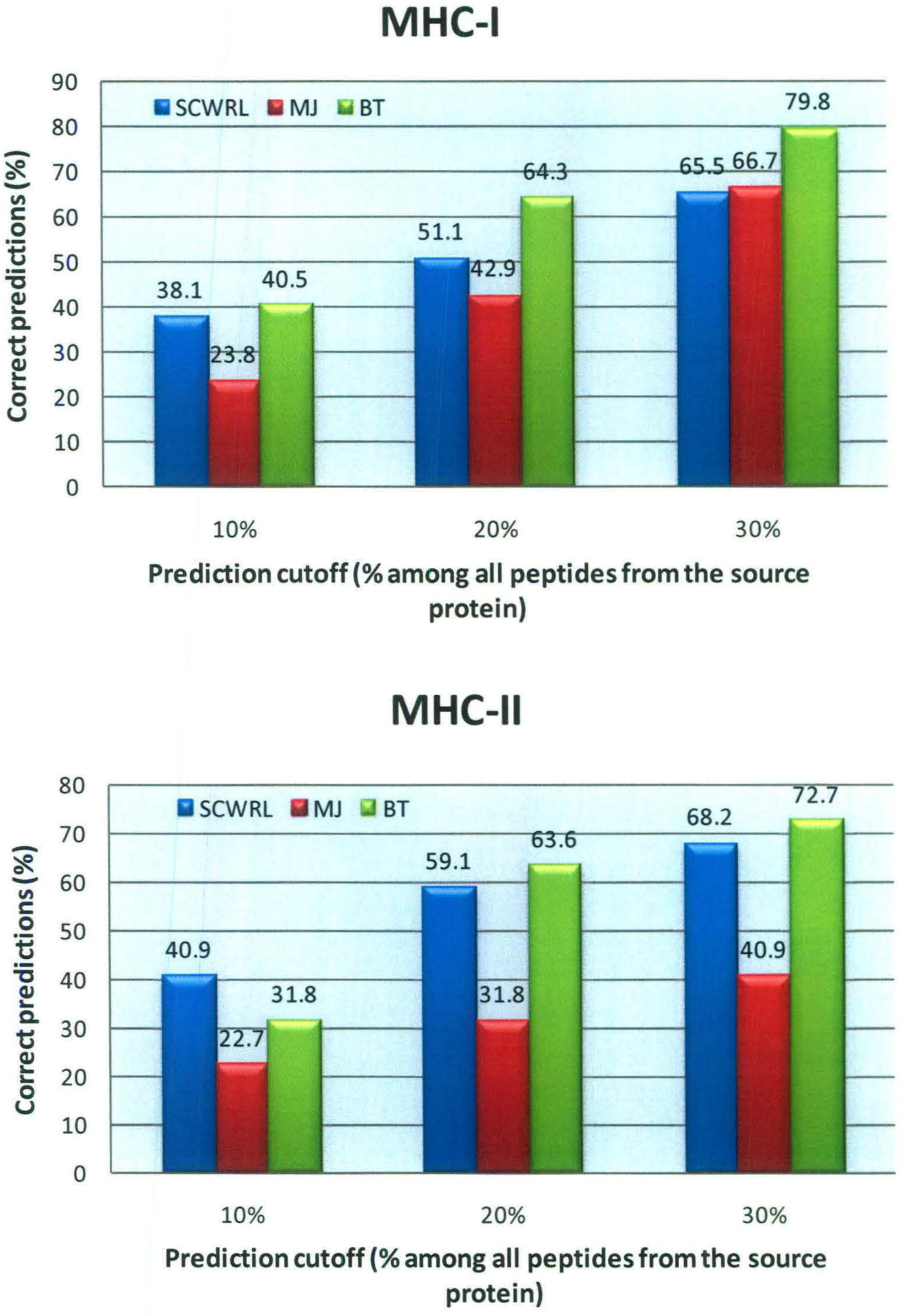


Figure 5.2: Peptide backbone conformation analysis of the class I allele HLA-A*0201 crystal structure bound peptides. (A) Clustering of 9mer peptides bound in 61 HLA-A*0201 crystal structures on the basis of their Cα RMSD values when every peptide was compared to every other peptide. Before RMSD calculation the peptides were brought in a common reference frame by superposition of MHC receptor coordinates. (B) Distribution of RMSD values of HLA-A*0201 bound peptides. (C) Backbone conformations of peptides bound in complex with the HLA-A*0201 crystal structures. Representative backbone conformations of 8mer (magenta), 9mer (green and red) and 10mer (blue) peptides are shown. Most of the 9mer peptides bind in the conformation same as that shown in the green color. However, in one crystal structure, the peptide traces an alternative path as shown in red color.

structural rearrangement required to accommodate the extra amino acid in the class I MHC pocket which is optimum for 9 mer peptides. Thus, the analysis of the bound conformations of a large number of peptides in MHC-peptide complexes demonstrated that, the peptides bind the MHC molecule in an essentially similar conformation and at the same site.

5.3.2 Scoring of crystal structure bound peptides in complex with MHC

We next proceeded to investigate whether the peptides bound in various MHC-peptide complexes can be distinguished based on their binding energy scores from all possible overlapping peptides of similar length present in the source proteins of the bound peptides. As described in the methods section, for a given MHC-peptide complex the bound peptide was considered as true positive binder and all other overlapping peptides of similar length from the corresponding source protein were considered as true negatives. Each of these peptides were modeled in complex with the corresponding MHC crystal structure and the binding energy score was calculated using Miyazawa-Jernigan (MJ) and Betancourt-Thirumalai (BT) residue based statistical pair potentials. The score was also calculated using residual steric clash function of SCWRL program as described in the methods section. The peptides were ranked as per their binding energy scores and the prediction was considered to be correct, if the bound peptide had a rank within top 30%. The cutoff value of 30 percentile for classifying a prediction as "correct" was arrived after several pilot studies. Figure 5.3 shows the prediction results for all the class I and class II MHC-peptide complexes, with percentile cutoff values of 10%, 20% and 30%. As can be seen in Figure 5.3A, only in 40.5% of the class I MHC-peptide complexes, the BT matrix could rank the bound peptide within top 10% of all possible peptides present in the source protein. On the other hand, in 79.8% of the cases BT matrix could rank the bound peptide within top 30%. Similarly, MJ matrix and SCWRL residual score could rank the bound peptide within top 30% in case of more than 65% of the class I MHC-peptide complexes. Thus, in case of class I MHC, BT matrix performed best among all three scoring schemes. Even at 20% cutoff level, the prediction accuracy of BT matrix was 64.3%. Similarly for class II MHC, BT matrix performed best at all cutoff levels (Figure 5.3B). The prediction accuracy of BT matrix was 72.7% at 30% cutoff level which is significantly better than prediction accuracy of 40.9% by MJ matrix.


MHC-I

MHC-II

Figure 5.3: Comparison of prediction accuracy at different cutoff values by SCWRL (van der Waals interactions), MJ matrix, and BT matrix.

The prediction accuracy based on SCWRL residual score was less than that by BT matrix at 30% cutoff level. The same trend was also observed at 20% cutoff level. Based on these results, the percentile cutoff of 30% was chosen for classifying the prediction for a given complex as correct. The choice of such larger percentile cutoff can also be justified by the fact that, ranking based on pair potential is used only for the first round of screening to shortlist peptides for detailed binding energy calculations using all atom forcefield. Thus, the results of benchmarking on MHC-peptide complex crystal structures indicated that BT matrix with 30% as percentile cutoff was most appropriate to ensure that true positive binder is sort listed in most cases. We next proceeded to investigate if the same criteria can be used to predict the known substrate specificity data on class I and class II MHCs cataloged in SYFPEITHI database.

5.3.3 Benchmarking on data obtained from SYFPEITHI database

In order to evaluate the predictive power of our structure based approach MODPROPEP, 30 different alleles, each having at least 20 known antigenic peptide substrates were selected. The source proteins of these peptides were scanned for all overlapping peptides. Since BT matrix performed best in predicting the crystal structure bound peptides, it was used as scoring matrix for benchmarking on SYFPEITHI data. A percentile cutoff level of 30% was chosen as criterion for correct prediction. Figure 5.4 shows the results of benchmarking on SYFPEITHI data. As can be seen, out of 30 alleles, 16 showed a prediction accuracy of more than 60%, i.e. in 16 alleles the known substrate peptide was found to be among top 30%, when all peptides from the source proteins were ranked using BT matrix. We also carried out the ROC analysis for these 16 alleles. A typical example of ROC analysis has been shown in Figure 5.5A for class I MHC allele A*0201 which contained maximum number of known antigenic peptides in SYFPEITHI data set. Figure 5.5B shows the distribution of AUC values for various alleles. Apart from AUC, the sensitivity and specificity for the 16 alleles are listed in Table 5.9. As can be seen, except for the allele H2-Db all others have AUC values higher than 0.7, and good sensitivity and specificity values. The AUC value for H2-Db was 0.597, with specificity and sensitivity of 98.5% and 43.8% respectively. The highest AUC value was 0.961 for the allele HLA-B*0801. The sensitivity and specificity value of HLA-B*0801 was

Figure 5.4: Benchmarking results of MODPROPEP for various MHC alleles on the peptide binding data in SYFPEITHI database. The BT matrix has been used for calculation of binding affinities of peptides to MHC alleles. A prediction was considered to be correct if the binding peptide was among top ranked 30% peptides among all peptides from its source protein. Number in the first bracket next to allele name indicates the length of peptide bound in the template structure. The total number of antigenic peptides used in prediction is mentioned in the last bracket. Red arrows indicate alleles with prediction accuracy more than 60%.

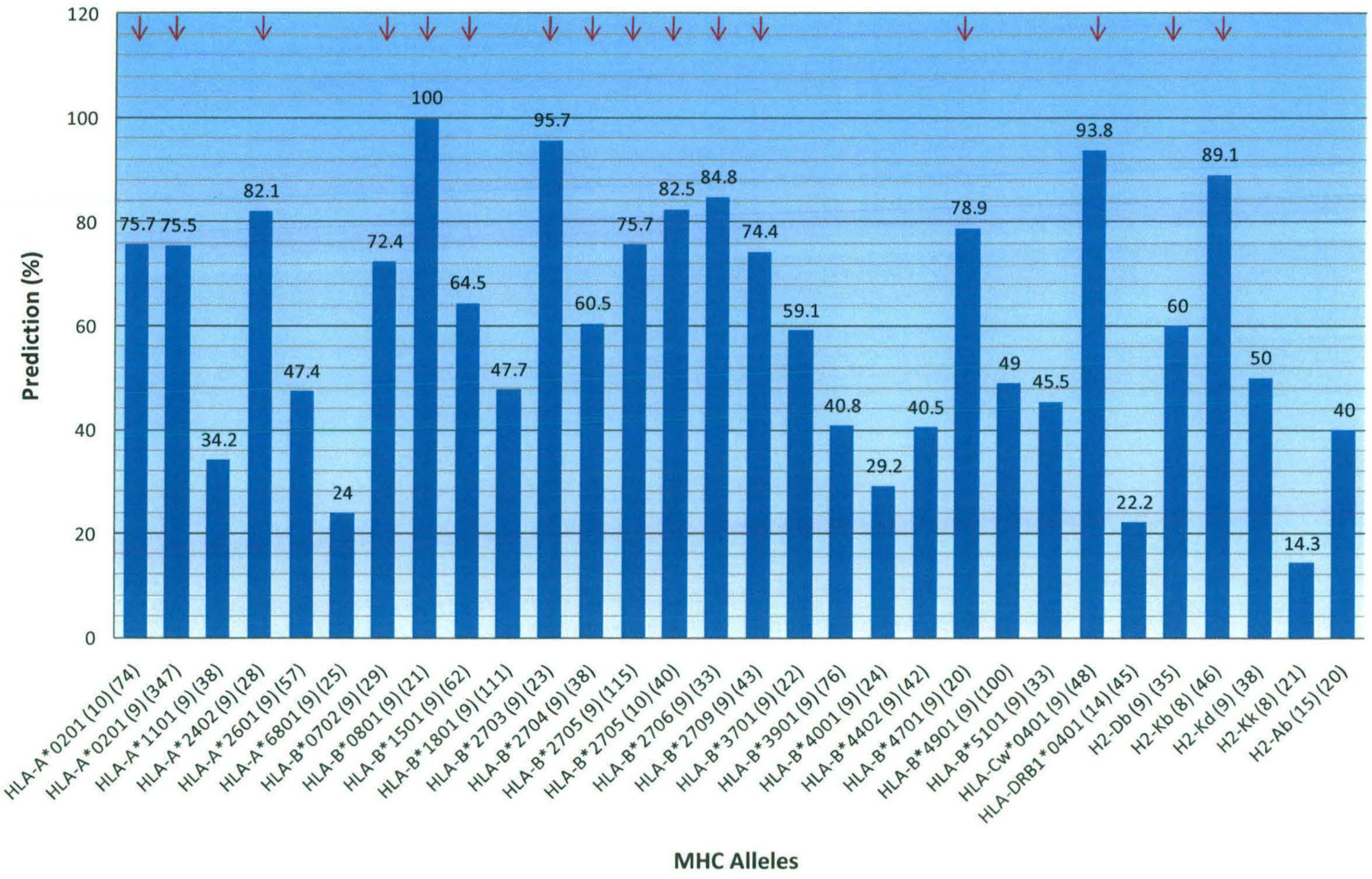

120
100
80
60
40
20
0
Prediction (%)
75.7
75.5
34.2
82.1
47.4
24
72.4
100
64.5
47.7
95.7
60.5
75.7
82.5
84.8
74.4
59.1
40.8
29.2
40.5
78.9
49
45.5
93.8
22.2
60
89.1
50
14.3
40
HLA-A*0201 (10) (74)
HLA-A*0201 (9) (347)
HLA-A*1101 (9) (38)
HLA-A*2402 (9) (28)
HLA-A*2601 (9) (57)
HLA-A*6801 (9) (25)
HLA-B*0702 (9) (29)
HLA-B*0801 (9) (21)
HLA-B*1501 (9) (62)
HLA-B*1801 (9) (111)
HLA-B*2703 (9) (23)
HLA-B*2704 (9) (38)
HLA-B*2705 (9) (115)
HLA-B*2705 (10) (40)
HLA-B*2706 (9) (33)
HLA-B*2709 (9) (43)
HLA-B*3701 (9) (22)
HLA-B*3901 (9) (76)
HLA-B*4001 (9) (24)
HLA-B*4402 (9) (42)
HLA-B*4701 (9) (20)
HLA-B*4901 (9) (100)
HLA-B*5101 (9) (33)
HLA-Cw*0401 (9) (48)
HLA-DRB1*0401 (14) (45)
H2-Db (9) (35)
H2-Kb (8) (46)
H2-Kd (9) (38)
H2-Kk (8) (21)
H2-Ab (15) (20)
MHC Alleles

Figure 5.4

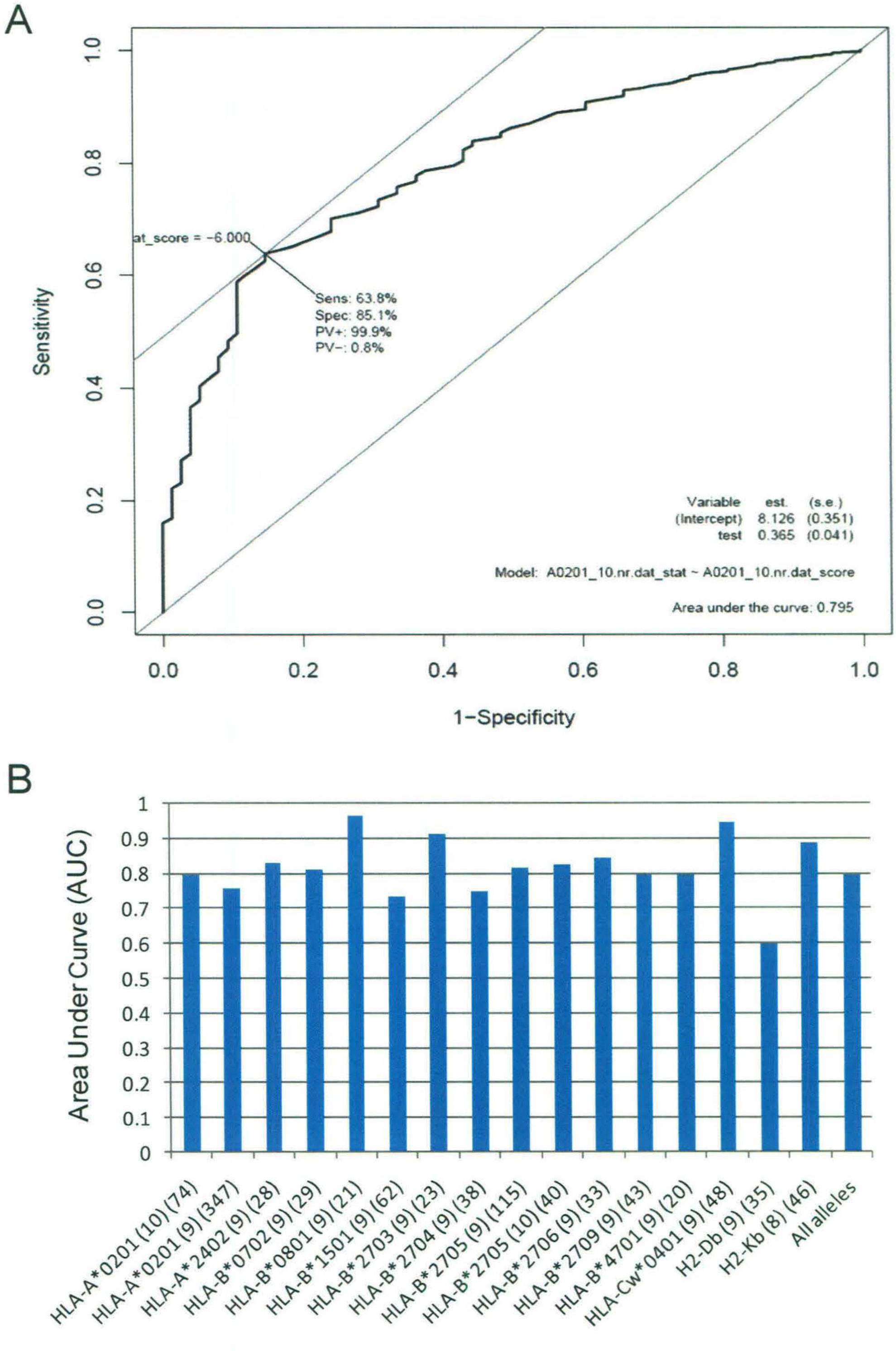


Figure 5.5: ROC analysis for prediction of binding peptides of MHC alleles. **(A)** The ROC curve for HLA-A*0201 allele. **(B)** Distribution of AUC values of 16 MHC alleles. Number in the first bracket next to allele name indicates the length of peptide bound in the template structure. The total number of antigenic peptides used in prediction is mentioned in the last bracket.

83.2% and 100% respectively. These values were consistent with the 100% prediction accuracy for this allele. When all alleles were considered together, the AUC value was 0.795 with specificity of 84.2% and sensitivity of 63.5%. Hence, the overall results from benchmarking on SYFPEITHI data suggests that, BT statistical pair potential can successfully identify potential MHC binding peptides for a large number of alleles. This result is specially encouraging, because our structure based approach MODPROPEP does not use any experimental data for training. Hence, it is likely that MODPROPEP can predict substrate peptides for new MHC alleles with high accuracy.

Table 5.9: List of Area under curve (AUC), sensitivity (Sn), and specificity (Sp) values for ROC analysis of 16 MHC alleles.

MHC Alleles	**Antigenic proteins**	**Prediction (%)**	**AUC**	**Sn (%)**	**Sp (%)**
HLA-A*0201 (10) (74)	74	75.7	0.795	63.8	85.1
HLA-A*0201 (9) (347)	347	75.5	0.758	73	70.9
HLA-A*2402 (9) (28)	28	82.1	0.831	65.6	92.6
HLA-B*0702 (9) (29)	29	72.4	0.808	60.3	100
HLA-B*0801 (9) (21)	21	100	0.961	83.2	100
HLA-B*1501 (9) (62)	62	64.5	0.735	60.9	77
HLA-B*2703 (9) (23)	23	95.7	0.912	76.6	95.7
HLA-B*2704 (9) (38)	38	60.5	0.748	46.3	100
HLA-B*2705 (9) (115)	115	75.7	0.814	66.1	86.1
HLA-B*2705 (10) (40)	40	82.5	0.824	65.4	86.8
HLA-B*2706 (9) (33)	33	84.8	0.841	66.1	97
HLA-B*2709 (9) (43)	43	74.4	0.795	57.2	95.1
HLA-B*4701 (9) (20)	20	78.9	0.794	83.8	73.7
HLA-Cw*0401 (9) (48)	48	93.8	0.943	84.9	91.7
H2-Db (9) (35)	35	60	0.597	43.8	98.5
H2-Kb (8) (46)	46	89.1	0.886	71.1	95.1
All alleles			0.795	63.5	84.2

5.3.4 Re-ranking of high scoring peptides using MM/PBSA

As per our multiscale strategy for identifying putative MHC binding peptides, in the first round we used residue based statistical pair potential matrix for screening of peptides which are most likely to have the favorable interaction with the MHC proteins. In case of a large number of MHC alleles, BT matrix was able to rank the true positive substrate peptide within top 30% of all possible peptides present in a given antigen. We wanted to investigate if re-ranking these top 30% high scoring peptides using all atom MM/PBSA approach can further improve the rank of the true positive peptides. For each of the 16 alleles, where BT matrix had a prediction accuracy above 60%, the top ranking 30% of the peptides were modeled using an all atom forcefield. MM/PBSA method was used for calculation of the binding free energies of these peptides. The MHC-peptide complexes were energy minimized before the calculation of binding energies. Figure 5.6 shows results of re-ranking of shortlisted peptides by MM/PBSA. For each allele, the left bar represents the percentages of substrates (out of those shortlisted in the first stage) for which the true binder could be ranked within 0%-10% (blue), 10%-20% (red) and 20%-30% (green) of all possible overlapping peptides in terms of BT matrix binding energy score. Since only top 30% of the peptides are shortlisted, in each case the total adds up to 100%. Right bar shows the same results after re-ranking using MM/PBSA binding free energy. Hence, an improvement in the rank of the true binder peptide will reflect as enhancement of percentage of substrates for which true binder could be ranked within top 10%. As can be seen from Figure 5.6, out of the 16 alleles, 8 alleles showed an improvement of the rank of known binder peptide. For example, in case of HLA-A*0201, only in case of 58.4% substrates, the known binder peptides were among top ranked 10% when scored by BT matrix. However, upon re-ranking by MM/PBSA method, in case of 72.1% of substrates, the known binder peptides fell among top ranked 10% peptides. Similarly, percentage of substrates for which the binder peptides were ranked within top 20%, increased from 84.4% to 91.6%, when the peptides shortlisted by BT matrix were re-ranked by MM/PBSA. Thus, these results suggests that re-ranking by MM/PBSA approach is indeed able to enhance the rank of the true binder peptide. Therefore, the multi scale structure based approach proposed in this thesis is indeed a promising approach for substrate peptides for new MHC alleles. To the best of our knowledge, this work also reports the most comprehensive benchmarking for prediction of the substrates for class I as well as class II MHCs covering a very significant number of antigens from a large number of alleles.

5.4 DISCUSSION

Rapid and correct identification of MHC-binding peptides is very important for development of knowledge based design of subunit vaccines. The limitations of experimental techniques, and the availability of a large amount of potential antigenic

Figure 5.6: Re-ranking of top ranked 30% peptides short listed by pair-potential matrix with a MM/PBSA based all atom force field. The first bar for each allele shows the fraction of cases in which known binding peptide is among top ranked 0-10%, 10-20% and 20-30% of all peptides from their respective source proteins. The second bar represents the redistribution when these 30% peptides are re-ranked by MM/PBSA method. Red arrows indicate the alleles for which the re-ranking by MM/PBSA led to improvement in the rank of known binding peptides. Number in the first bracket next to allele name indicates the length of peptide bound in the template structure. The total number of antigenic peptides used in prediction is mentioned in the last bracket.

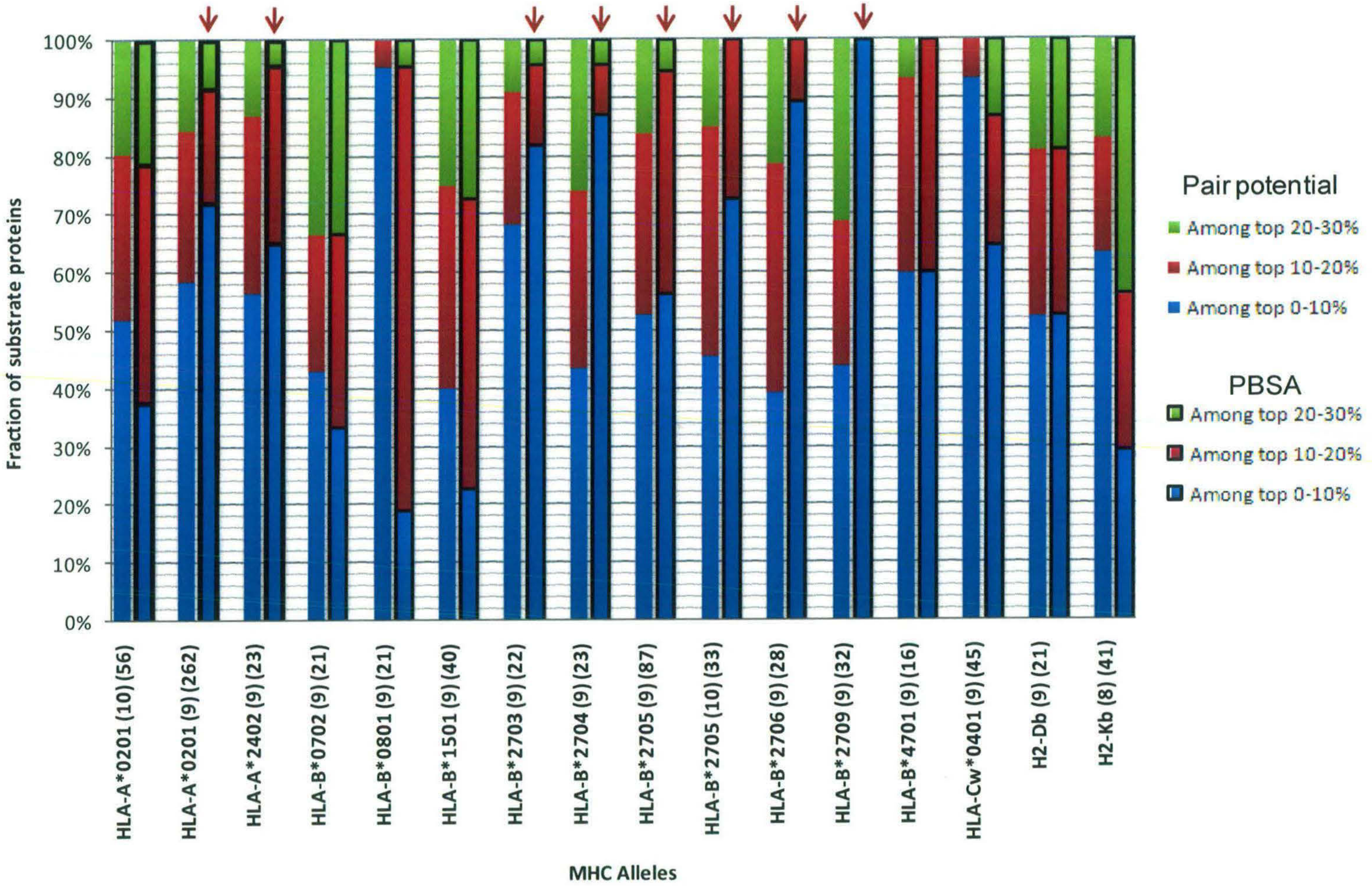

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Fraction of substrate proteins
HLA-A*0201 (10) (56)
HLA-A*0201 (9) (262)
HLA-A*2402 (9) (23)
HLA-B*0702 (9) (21)
HLA-B*0801 (9) (21)
HLA-B*1501 (9) (40)
HLA-B*2703 (9) (22)
HLA-B*2704 (9) (23)
HLA-B*2705 (9) (87)
HLA-B*2705 (10) (33)
HLA-B*2706 (9) (28)
HLA-B*2709 (9) (32)
HLA-B*4701 (9) (16)
HLA-Cw*0401 (9) (45)
H2-Db (9) (21)
H2-Kb (8) (41)
MHC Alleles
Pair potential
Among top 20-30%
Among top 10-20%
Among top 0-10%
PBSA
Among top 20-30%
Among top 10-20%
Among top 0-10%

Figure 5.6

protein data from genome sequences has led to the development of a number of computational methods for prediction of antigenic peptides in context to a number of class I and class II MHC alleles. Most of these methods are sequence motif, PSSM, artificial intelligence based methods which rely upon the training on already known MHC binding peptides. A very high degree of conservation in the MHC allele sequences and the structure fold holds the promise for the development of a structural based approach. The problem of identification of MHC binding peptides is essentially a problem of protein-peptide interaction rooted in the physical principles of identification of a favorable peptide by MHC protein. In this chapter, we have used the crystal structure of MHC-peptide complexes for prediction of MHC binding peptide where the MHC-peptide complexes are modeled using the existing complexes as structural templates. The overlapping peptides of a given length from an antigenic protein are modeled in complex with MHC and are scored by pair-potential matrix. The high scoring peptides are shortlisted and scored by the detailed all atom based MM/PBSA potential.

Our analysis of RMSD values of peptide backbone in the crystal structure bound conformation, when compared to each other in pairs, showed that mostly they bind in the MHC binding groove in more or less same conformation. This observation was same for class I as well as class II MHC proteins. Only a small fraction of peptide pairs exhibit slightly more deviation which may be attributed to few peptides which bind in an alternative conformation. The analysis of conformation of bound peptides, when only MHC protein coordinates are superimposed, also confirmed that most of the peptides bind in the same extended conformation. Detailed analysis of HLA-A*0201 revealed that 9mer peptides bound in the 60 complexes traverse the same path while only one 9mer peptide bind in different conformation.

The conservation in the binding conformation of the peptides implies that the contacting residues in MHC allele, which interact and stabilize peptide residues, can be identified with a good degree of confidence. As these same residues would also stabilize the peptide of unknown sequence, the compatibility of a peptide in the same conformation can be evaluated by employing a threading like approach. We used MODPROPEP, to identify the template crystal structure bound peptide from among all possible peptides from its source protein. To evaluate the interaction energy of a peptide, BT, and MJ pair potential matrix and van der Waals energy based SCWRL energy functions were used. BT matrix performed best among all three scoring schemes for class I as well as class II MHC complexes. BT matrix has been shown previously to perform well in cases involving hydrophilic as well as hydrophobic interface. In contrast, MJ matrix performs well at hydrophobic interface as it tends to score well for residues involving hydrophobic-hydrophobic interactions. Since, prediction accuracy was best when BT matrix was used for scoring, it implies that MHC-peptide complex involves abundant hydrophilic as well as hydrophobic residues as interaction partners between MHC and peptide.

When benchmarked on experimentally known binding peptide data using BT matrix, the prediction accuracy was found to be more than 60% for 16 alleles belonging to class I MHC. The prediction accuracy and AUC for HLA-B*0801 reached as high as 100% and 0.961 respectively. Therefore, BT pair potential matrix could rank the known binding peptides with high score as compared to other nonbinding peptides in the same protein. However, BT matrix could not predict so well for other 14 alleles. As the correct scoring of the peptide depends on the list of residue-residue contacting pairs between MHC and peptide, the correct identification of contacting residue pair is very important. For some models, when the structural similarity is not very high, or peptide bind in different conformation than that in the template crystal structure used for modeling, the incorrect residue pair identification may have led to the failure to predict known binding peptide among high ranking peptides. In addition, residue-residue pair potential is a coarse grained potential which works at the residue level and does not take into account the detailed atomic interaction while calculating the interaction energy. Scoring of binding energy of shortlisted peptides, using an all atom potential MM/PBSA, improved the ranks of the known binding peptide in case of 8 MHC alleles. Hence, all atom based MM/PBSA potential has shown promise as a good method for prediction of MHC binding peptides when coupled with BT matrix pair-potential. However, considering a large number of protein-peptide complexes involved while calculating the binding energy using MM/PBSA, we performed only a short energy minimization of 1000 steps on MHC-peptide complexes. For some complexes, this short minimization may be insufficient in removing the steric conflicts which might have occurred during side chain modeling by rotamer library. As a result, in such cases, the MM/PBSA binding energies would not be the true representative of actual binding affinity of peptide to MHC proteins. This might have led to the failure of MM/PBSA in improving the rank of binding peptide in some alleles. The full energy minimization of all peptides prior to the calculation of binding energy holds the promise in further improving the rank of binding peptide.

Summary and Conclusions

Proteins involved in majority of cellular processes usually perform their function by binding to some target proteins and forming protein-protein complexes. Hence, identification of the potential interacting partners of a given protein is crucial for understanding the molecular details of a variety of cellular processes. Experimental approaches for identification of protein-protein interactions indicate that often receptor protein recognizes short contiguous peptide stretches on the interaction partner. Among the various proteins which interact specifically with short peptide motifs, MHC and kinases represent two major protein families whose substrate specificities have been extensively studied by various experimental approaches. Correct identification of MHC binding peptides has important implications for modern epitope based vaccine design and also cancer therapy. Similarly, protein kinases are essential for regulation of the diverse cellular processes including metabolism, stress responses and cell-cycle control. Therefore, identification of substrate proteins for various kinases is crucial for understanding signaling networks in various organisms.

Even though sequence based *in silico* prediction tools are widely used for identifying MHC binding peptides and phosphorylation sites on proteins, they rely on identification of sequence motifs which are found in a known set of substrate peptides. Therefore, these methods can not be applied in case of MHC alleles or kinase subfamilies for which no experimental data is available. Discovery of newer MHC alleles and identification of large number of kinases in various genomes require development of novel computational methods which can give reliable clues about their substrate specificities even in absence of extensive experimental data. Structure based substrate prediction methods can in principle address these problems.

In this thesis, we have developed a novel structure based substrate prediction method for MHCs and kinases. This prediction method involves a multiscale approach, where at the first level putative high scoring substrate peptides are identified by threading of peptide sequences on the structural templates of kinase-peptide or MHC-peptide complexes and scoring them by residue based statistical pair potentials developed by Miyazawa and Jernigan or Betancourt and Thirumalai. High scoring peptides short listed by initial screening are modeled in the peptide binding pocket using rotamer library and detailed all atom molecular mechanics potentials, and their binding affinity is re-ranked using binding free energy values computed by MM/PBSA approach. In order to test the predictive ability of statistical pair potentials, detailed benchmarking was carried out on the experimentally identified substrate peptide data sets cataloged in Phospho.ELM and SYFPEITHI database. All these steps were carried out using MODPROPEP, a software developed for automating above mentioned computational tasks.

This software has been developed based on the observation that all the protein kinases shared a conserved structural fold despite their sequence divergence. Similar conservation of structures were also observed both for class I and class II MHC structures which share a higher degree of sequence identity within themselves. BLAST alignment of large number of protein kinases and MHC proteins available in sequence databases with the crystal structures also indicated that, homology models can be obtained for most of these sequences with reasonable accuracy. Comparison of the bound peptide structures indicated that in all of these three classes of proteins, the substrate peptides bind at a structurally homologous site on the conserved fold and the bound peptides maintain a more or less similar extended conformation. This suggested the possibility that bound peptides from peptide-protein complexes can be transformed to the protein structures lacking the bound peptide based on optimum superposition of the protein structures.

The prediction accuracy of MODPROPEP for prediction of phosphorylation sites was benchmarked on the experimentally verified phosphorylation sites catalogued in Phospho.ELM database (version 5.0, May 2006). It contained a total of 13603 phosphorylation instances in 4422 proteins by 263 kinase families. The analysis was restricted to Ser/Thr kinases having at least 20 phosphorylation instances in Phospho.ELM. Kinases which did not show significant homology with available structural templates were also excluded. Finally, out of the 38 substrate specific classes, 22 classes containing 70 kinase families were selected for benchmarking of our structure based substrate prediction program, MODPROPEP. They contained a total of 2457 phosphorylation instances in 1180 proteins by 70 kinase families. The results of MODPROPEP were compared with other prediction tools, namely, GPS, PPSP, SCANSITE, NetphosK and PREDIKIN. Each of the Ser/Thr containing heptameric peptides were modeled in the substrate binding pocket of respective kinase or the structural model. The binding energy of each peptide was calculated using the BT pair potential matrix. All modeled peptides were sorted as per their binding energy and assigned a rank. The prediction was considered to be correct if the peptide containing the known phosphorylation site had a rank within top 30%. This procedure was repeated for all the known substrate proteins for a given kinase group and percentage of substrates for which MODPROPEP gave correct prediction was evaluated. Similarly, prediction accuracy was calculated for all 22 kinase groups. It was encouraging to note that, our structure based method could predict more than 60% of the experimentally identified substrates for 10 protein kinases. The prediction accuracies for PKA, PKB, PKG, and PDK were well above 70% with PKG having the highest prediction accuracy of 81.5%. The other kinase groups for which MODPROPEP showed good prediction accuracy were ChK, CK2, DAPK, ROCK, and MAP3K. MODPROPEP also outperformed all other prediction tools for PKG, PDK, ChK, CK2, DAPK, ROCK, and MAP3K. However, for protein kinases PKA, PKB, and PAK performance of other prediction tools were better than MODPROPEP, even though MODPROPEP had prediction accuracy more than 60% for these kinase groups. We also carried out receiver operating characteristic (ROC) curve analysis for analyzing the robustness of our structure based prediction approach. The area under curve (AUC) values of these 10 kinases ranged from 0.681 (MAP3K) to 0.838 (PKG).

The prediction accuracy of MODPROPEP for the remaining 12 protein kinases was low, because many of the actual phosphorylation sites were not being ranked among the top 30% of all Ser/Thr containing peptides in these cases. The structural models of kinase-peptide complexes were analyzed in detail for each of these cases to understand the reasons for poor prediction accuracy. It was found that certain crucial residues which have been reported to be determinants of substrate specificity were not predicted as binding pocket residues based on our distance based criteria. Similarly, some residues were included in list of subsites even though they did not make significant contact with the peptide residues. Therefore, appropriate modifications were carried out to the list of binding pocket residues and a modified version of the algorithm was used to re-evaluate the prediction accuracy for all the kinase families. It was found that, for 10 kinase groups where prediction accuracy was more than 60% earlier, modified version of MODPROPEP also showed good prediction accuracy with further improvement in few cases. Prediction accuracy of PKA and PKB improved to 84.8% and 77.6% from 71.4% and 70.9% respectively. Most significant improvement was in case of PKC with prediction accuracy reaching 73.4% from 43.3%. However, for the remaining cases where earlier version of

MODPROPEP had shown low prediction accuracy, no significant improvement could be observed despite the modifications to the binding pocket residues.

Thus our results indicate that, BT pair potential was able to successfully rank the substrate peptides within top 30% of all Ser/Thr containing peptides for 11 kinase families. We also wanted to investigate whether modeling of peptides at an all atom level and evaluating their binding free energy by inclusion of electrostatic and desolvation terms can further improve the ranks of substrate peptides. The high scoring top 30% peptides obtained from pair potential based screening were modeled in the binding site of their respective kinases using rotamer library approach. The resulting protein-peptide complexes were energy minimized and their interaction free energy was calculated by MM/PBSA method. The peptides were ranked and arranged as per their binding free energy. The AUC values for 8 kinase families significantly improved for MM/PBSA approach when compared to the AUC values obtained from pair potential based approach. Thus MM/PBSA method was able to further improve the ranks of substrate peptides among all possible Ser/Thr containing peptides present in the substrate proteins

We also investigated whether inclusion of solvent accessibility probability of putative substrate peptides can help in improvement of prediction accuracy. Although some computational method for prediction of phosphorylation sites take into account solvent accessibility of peptides, the importance of solvent accessibility has not been analyzed thoroughly. To investigate the importance of the solvent accessibility in the phosphorylation event by protein kinase, we systematically analyzed the solvent accessibilities of phosphorylation sites in known substrate proteins, and compared with the accessibilities of sites which are not phosphorylated. Out of a total of 13563 phosphorylation sites in 4422 protein sequences, a total of 1088 phosphorylation instances could be mapped to 990 PDB entries. These included a 526, 202 and 360 instances containing serine, threonine and tyrosine as phosphorylation residue respectively. The solvent accessible surface area of these residues and the heptapeptides containing these as central residues were calculated using NACCESS program. The average relative solvent accessible area of phosphorylation site residues was found to be significantly more than their non-phosphorylated counterparts. The difference between phospho and non-phospho residues was statistically significant as judged by Wilcoxon test p-values of 2.20×10^{-16}, 5.07×10^{-6} and 2.34×10^{-8} for serine, threonine and tyrosine containing sites. These results suggest that incorporation of solvent accessibility term along with the current scoring function based on the residue-residue statistical energy can further improve the prediction accuracy. An internet based resource, pACCESS has been developed in this thesis for the analysis of the solvent accessible surface area of phosphorylation sites, which is made available at http://www.nii.ac.in/paccess.html.

Benchmarking was also carried out to evaluate the prediction accuracy of MODPROPEP for identifying substrate peptides recognized by class I and class II MHC proteins. In contrast to a limited number of crystal structures available for kinases, a large number of MHC-peptide complexes were available in PDB. Therefore, the structure based approach was first benchmarked on the available MHC-peptide complexes. We wanted to investigate whether the Betancourt-Thirumalai statistical potential can distinguish the MHC bound peptide from among the all possible overlapping peptides derived from the source protein of the bound peptide. It was encouraging to note that, in 79.8% of MHC class I crystal structures, the bound peptide was ranked among 30%. The prediction accuracy was found to be 72.7% in case of class II MHC-peptide complex dataset. Subsequently, MODPROPEP was also benchmarked using the peptide ligand information catalogued in the SYFPEITHI database. A total of 30 alleles which had more than 20 known peptide substrates, were selected for this analysis. MODPROPEP was used to scan the source protein of these peptides and the prediction was considered to be correct if the actual binder peptide was among the top 30% of all scored peptides. It was found that for 16 alleles, the prediction accuracy was more than 60%. The AUC values for these alleles ranged from 0.597 to 0.912. The sensitivity and specificity values were also promising for most of these alleles. The high scorer peptides obtained by the pair-potential were modeled by rotamer library approach in the binding pocket of their respective MHCs and ranked as per their binding energy using MM/PBSA method. Analysis of MM/PBSA results indicated that, in case of eight out of the 16 alleles, use of MM/PBSA for re-scoring resulted in improvement in the ranks of the experimentally identified substrate peptides. The improvement was most striking in case of class I MHC allele B*2709 in which MM/PBSA ranked all the 43 peptides among 10% top ranked peptides.

In this thesis, we have used a combination of rotamer library and scoring scheme using statistical pair potentials for developing a novel structure based computational approach for prediction of substrates for MHC and protein kinases.

Our benchmarking studies on large number of known substrates available in Phospho.ELM and SYFPEITHI indicate that, this computational method has good prediction accuracy for a number of kinase families and class I as well as class II MHC alleles. Our analysis also suggests that the prediction accuracy of phosphorylation site prediction programs can be further improved by use of propensity for solvent accessibility. The computational protocol MODPROPEP, developed in this thesis has also been made available as a user-friendly web based software at http://www.nii.res.in/modpropep.html. Apart from prediction, this software also provides an easy interface to analyze the protein-peptide interactions to understand the role of crucial residues in recognition of substrate peptides. Although currently we have benchmarked MODPROPEP on MHCs and kinases, this method being general in principle, can be extended easily to other protein-peptide complexes.

The work reported in this thesis has resulted in the following publication:

Kumar, N. and Mohanty, D. (2007). **MODPROPEP: a program for knowledge-based modeling of protein-peptide complexes.** *Nucleic Acids Res*, 35: W549-555.

Bibliography

Altuvia, Y., Schueler, O. and Margalit, H. (1995) Ranking potential binding peptides to MHC molecules by a computational threading approach. *J Mol Biol*, **249**, 244-250.

Altuvia, Y., Sette, A., Sidney, J., Southwood, S. and Margalit, H. (1997) A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Hum Immunol*, **58**, 1-11.

Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G. and Pandey, A. (2007) A curated compendium of phosphorylation motifs. *Nat Biotechnol*, **25**, 285-286.

Anamika, K., Bhattacharya, A. and Srinivasan, N. (2008) Analysis of the protein kinome of Entamoeba histolytica. *Proteins*, **71**, 995-1006.

Ansari, M. Z., Yadav, G., Gokhale, R. S. and Mohanty, D. (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res*, **32**, W405-413.

Apostolopoulos, V. and Lazoura, E. (2004) Noncanonical peptides in complex with MHC class I. *Expert Rev Vaccines*, **3**, 151-162.

Bauman, A. L. and Scott, J. D. (2002) Kinase- and phosphatase-anchoring proteins: harnessing the dynamic duo. *Nat Cell Biol*, **4**, E203-206.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.

Betancourt, M. R. and Thirumalai, D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci*, **8**, 361-369.

Bhasin, M. and Raghava, G. P. (2003) Prediction of promiscuous and high-affinity mutated MHC binders. *Hybrid Hybridomics*, **22**, 229-234.

Bhasin, M. and Raghava, G. P. (2007) A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J Biosci*, **32**, 31-42.

Biondi, R. M. and Nebreda, A. R. (2003) Signalling specificity of Ser/Thr protein kinases through docking-site-mediated interactions. *Biochem J*, **372**, 1-13.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. and Wiley, D. C. (1987) Structure of the human class I histocompatibility antigen, HLA-A2. *Nature*, **329**, 506-512.

Bjorkman, P. J. and Parham, P. (1990) Structure, function, and diversity of class I major histocompatibility complex molecules. *Annu Rev Biochem*, **59**, 253-288.

Blom, N., Gammeltoft, S. and Brunak, S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*, **294**, 1351-1362.

Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. and Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633-1649.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, **31**, 365-370.

Bossemeyer, D., Engh, R. A., Kinzel, V., Ponstingl, H. and Huber, R. (1993) Phosphotransferase and substrate binding mechanism of the cAMP-dependent protein kinase catalytic subunit from porcine heart as deduced from the 2.0 A structure of the complex with Mn2+ adenylyl imidodiphosphate and inhibitor peptide PKI(5-24). *Embo J*, **12**, 849-859.

Bradham, C. A., Foltz, K. R., Beane, W. S., Arnone, M. I., Rizzo, F., Coffman, J. A., Mushegian, A., Goel, M., Morales, J., Geneviere, A. M. *et al.* (2006) The sea urchin kinome: a first look. *Dev Biol*, **300**, 180-193.

Brinkworth, R. I., Breinl, R. A. and Kobe, B. (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc Natl Acad Sci U S A*, **100**, 74-79.

Brinkworth, R. I., Munn, A. L. and Kobe, B. (2006) Protein kinases associated with the yeast phosphoproteome. *BMC Bioinformatics*, **7**, 47.

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L. and Wiley, D. C. (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature*, **364**, 33-39.

Brown, N. R., Noble, M. E., Endicott, J. A. and Johnson, L. N. (1999) The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. *Nat Cell Biol*, **1**, 438-443.

Buus, S., Lauemoller, S. L., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A., Hilden, J., Holm, A. and Brunak, S. (2003) Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens*, **62**, 378-384.

Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T. and Manning, G. (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc Natl Acad Sci U S A*, **101**, 11707-11712.

Canutescu, A. A., Shelenkov, A. A. and Dunbrack, R. L., Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, **12**, 2001-2014.

Case, D. A., Darden, T. A., Cheatham, I., T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Pearlman, D. A., Crowley, M. *et al.* (2006) AMBER 9, University of California, San Francisco.

Chang, E. J., Begum, R., Chait, B. T. and Gaasterland, T. (2007) Prediction of cyclin-dependent kinase phosphorylation substrates. *PLoS ONE*, **2**, e656.

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

Cresswell, P. (1994) Assembly, transport, and function of MHC class II molecules. *Annu Rev Immunol*, **12**, 259-293.

Cujec, T. P., Medeiros, P. F., Hammond, P., Rise, C. and Kreider, B. L. (2002) Selection of v-abl tyrosine kinase substrate sequences from randomized peptide and cellular proteomic libraries using mRNA display. *Chem Biol*, **9**, 253-264.

Dahiyat, B. I. and Mayo, S. L. (1997) De novo protein design: fully automated sequence selection. *Science*, **278**, 82-87.

Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N. and Gibson, T. J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.

Diella, F., Gould, C. M., Chica, C., Via, A. and Gibson, T. J. (2008) Phospho.ELM: a database of phosphorylation sites--update 2008. *Nucleic Acids Res*, **36**, D240-244.

Donnes, P. and Elofsson, A. (2002) Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, **3**, 25.

Doytchinova, I. A., Walshe, V. A., Jones, N. A., Gloster, S. E., Borrow, P. and Flower, D. R. (2004) Coupling in silico and in vitro analysis of peptide-MHC binding: a bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes. *J Immunol*, **172**, 7495-7502.

Dunbrack, R. L., Jr. (2002) Rotamer libraries in the 21st century. *Curr Opin Struct Biol*, **12**, 431-440.

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

Goldberg, J. M., Manning, G., Liu, A., Fey, P., Pilcher, K. E., Xu, Y. and Smith, J. L. (2006) The dictyostelium kinome--analysis of the protein kinases from a simple model organism. *PLoS Genet*, **2**, e38.

Guan, P., Doytchinova, I. A., Zygouri, C. and Flower, D. R. (2003a) MHCPred: A server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res*, **31**, 3621-3624.

Guan, P., Doytchinova, I. A., Zygouri, C. and Flower, D. R. (2003b) MHCPred: bringing a quantitative dimension to the online prediction of MHC binding. *Appl Bioinformatics*, **2**, 63-66.

Hanks, S. K. and Hunter, T. (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *Faseb J*, **9**, 576-596.

Hanks, S. K. (2003) Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. *Genome Biol*, **4**, 111.

Hattotuwagama, C. K., Guan, P., Doytchinova, I. A., Zygouri, C. and Flower, D. R. (2004) Quantitative online prediction of peptide binding to the major histocompatibility complex. *J Mol Graph Model*, **22**, 195-207.

Hou, T., Chen, K., McLaughlin, W. A., Lu, B. and Wang, W. (2006) Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. *PLoS Comput Biol*, **2**, e1.

Huang, H. D., Lee, T. Y., Tzeng, S. W. and Horng, J. T. (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res*, **33**, W226-229.

Hubbard, S. and Thorntan, J. M. (1993) NACCESS Computer Program. Department of Biochemistry and Molecular Biology, University College, London.

Hunter, T. and Plowman, G. D. (1997) The protein kinases of budding yeast: six score and more. *Trends Biochem Sci*, **22**, 18-22.

Hunter, T. and Schulman, H. (2005) CaMKII structure--an elegant design. *Cell*, **123**, 765-767.

Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z. and Dunker, A. K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*, **32**, 1037-1049.

Ideker, T. and Sharan, R. (2008) Protein networks in disease. *Genome Res*, **18**, 644-652.

Jacob, L. and Vert, J. P. (2008) Efficient peptide-MHC-I binding prediction for alleles with few known binders. *Bioinformatics*, **24**, 358-366.

Janin, J., Bahadur, R. P. and Chakrabarti, P. (2008) Protein-protein interaction and quaternary structure. *Q Rev Biophys*, **41**, 133-180.

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

Kempler, K., Toth, J., Yamashita, R., Mapel, G., Robinson, K., Cardasis, H., Stevens, S., Sellers, J. R. and Battelle, B. A. (2007) Loop 2 of Limulus Myosin III Is Phosphorylated by Protein Kinase A and Autophosphorylation. *Biochemistry*, **46**, 4280-4293.

Kennelly, P. J. (2002) Protein kinases and protein phosphatases in prokaryotes: a genomic perspective. *FEMS Microbiol Lett*, **206**, 1-8.

Klein, J., Figueroa, F. and Nagy, Z. A. (1983) Genetics of the major histocompatibility complex: the final act. *Annu Rev Immunol*, **1**, 119-142.

Knighton, D. R., Zheng, J. H., Ten Eyck, L. F., Xuong, N. H., Taylor, S. S. and Sowadski, J. M. (1991) Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, **253**, 414-420.

Kobe, B., Kampmann, T., Forwood, J. K., Listwan, P. and Brinkworth, R. I. (2005) Substrate specificity of protein kinases and computational prediction of substrates. *Biochim Biophys Acta*, **1754**, 200-209.

Kruger, T., Schoor, O., Lemmel, C., Kraemer, B., Reichle, C., Dengjel, J., Weinschenk, T., Muller, M., Hennenlotter, J., Stenzl, A. *et al.* (2005) Lessons to be learned from primary renal cell carcinomas: novel tumor antigens and HLA ligands for immunotherapy. *Cancer Immunol Immunother*, **54**, 826-836.

Krupa, A. and Srinivasan, N. (2002) The repertoire of protein kinases encoded in the draft version of the human genome: atypical variations and uncommon domain combinations. *Genome Biol*, **3**, RESEARCH0066.

Krupa, A., Abhinandan, K. R. and Srinivasan, N. (2004) KinG: a database of protein kinases in genomes. *Nucleic Acids Res*, **32**, D153-155.

Kube, E., Becker, T., Weber, K. and Gerke, V. (1992) Protein-protein interaction studied by site-directed mutagenesis. Characterization of the annexin II-binding site on p11, a member of the S100 protein family. *J Biol Chem*, **267**, 14175-14182.

Kumar, N. and Mohanty, D. (2007) MODPROPEP: a program for knowledge-based modeling of protein-peptide complexes. *Nucleic Acids Res*, **35**, W549-555.

LaCount, D. J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J. R., Schoenfeld, L. W., Ota, I., Sahasrabudhe, S., Kurschner, C. *et al.* (2005) A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature*, **438**, 103-107

Levinson, N. M., Kuchment, O., Shen, K., Young, M. A., Koldobskiy, M., Karplus, M., Cole, P. A. and Kuriyan, J. (2006) A Src-like inactive conformation in the abl tyrosine kinase domain. *PLoS Biol*, **4**, e144.

Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A., Jorgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K. *et al.* (2007) Systematic discovery of in vivo phosphorylation networks. *Cell*, **129**, 1415-1426.

Linding, R., Jensen, L. J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M. B. and Pawson, T. (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res*, **36**, D695-699.

Lowe, E. D., Noble, M. E., Skamnaki, V. T., Oikonomakos, N. G., Owen, D. J. and Johnson, L. N. (1997) The crystal structure of a phosphorylase kinase peptide substrate complex: kinase substrate recognition. *Embo J*, **16**, 6646-6658.

Madhusudan, Trafny, E. A., Xuong, N. H., Adams, J. A., Ten Eyck, L. F., Taylor, S. S. and Sowadski, J. M. (1994) cAMP-dependent protein kinase: crystallographic insights into substrate recognition and phosphotransfer. *Protein Sci*, **3**, 176-187.

Malakauskas, S. M. and Mayo, S. L. (1998) Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol*, **5**, 470-475.

Mann, M., Ong, S. E., Gronborg, M., Steen, H., Jensen, O. N. and Pandey, A. (2002) Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol*, **20**, 261-268.

Manning, G., Plowman, G. D., Hunter, T. and Sudarsanam, S. (2002a) Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci*, **27**, 514-520.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T. and Sudarsanam, S. (2002b) The protein kinase complement of the human genome. *Science*, **298**, 1912-1934.

Manning, G., Young, S. L., Miller, W. T. and Zhai, Y. (2008) The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *Proc Natl Acad Sci U S A*, **105**, 9674-9679.

Marshall, S. A., Morgan, C. S. and Mayo, S. L. (2002) Electrostatics significantly affect the stability of designed homeodomain variants. *J Mol Biol*, **316**, 189-199.

Martin, L., Catherinot, V. and Labesse, G. (2006) kinDOCK: a tool for comparative docking of protein kinase ligands. *Nucleic Acids Res*, **34**, W325-329.

Matsumura, M., Fremont, D. H., Peterson, P. A. and Wilson, I. A. (1992) Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science*, **257**, 927-934.

Matsuoka, Y., Hughes, C. A. and Bennett, V. (1996) Adducin regulation. Definition of the calmodulin-binding domain and sites of phosphorylation by protein kinases A and C. *J Biol Chem*, **271**, 25157-25166.

Mawuenyega, K. G., Forst, C. V., Dobos, K. M., Belisle, J. T., Chen, J., Bradbury, E. M., Bradbury, A. R. and Chen, X. (2005) Mycobacterium tuberculosis functional network analysis by global subcellular protein profiling. *Mol Biol Cell*, **16**, 396-404.

McLachlin, D. T. and Chait, B. T. (2001) Analysis of phosphorylated proteins and peptides by mass spectrometry. *Curr Opin Chem Biol*, **5**, 591-602.

Miranda-Saavedra, D. and Barton, G. J. (2007) Classification and functional annotation of eukaryotic protein kinases. *Proteins*, **68**, 893-914.

Miyazawa, S. and Jernigan, R. L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, **256**, 623-644.

Mooers, B. H., Datta, D., Baase, W. A., Zollars, E. S., Mayo, S. L. and Matthews, B. W. (2003) Repacking the Core of T4 lysozyme by automated design. *J Mol Biol*, **332**, 741-756.

Morgan, D. O. (1997) Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annu Rev Cell Dev Biol*, **13**, 261-291.

Morrison, D. K., Murakami, M. S. and Cleghon, V. (2000) Protein kinases and phosphatases in the Drosophila genome. *J Cell Biol*, **150**, F57-62.

Neuberger, G., Schneider, G. and Eisenhaber, F. (2007) pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. *Biol Direct*, **2**, 1.

Niedner, R. H., Buzko, O. V., Haste, N. M., Taylor, A., Gribskov, M. and Taylor, S. S. (2006) Protein kinase resource: an integrated environment for phosphorylation research. *Proteins*, **63**, 78-86.

Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S. L., Lamberth, K., Buus, S., Brunak, S. and Lund, O. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*, **12**, 1007-1017.

Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S. and Lund, O. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, **20**, 1388-1397.

Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Roder, G., Peters, B., Sette, A., Lund, O. *et al.* (2007) NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE*, **2**, e796.

Nishikawa, K., Toker, A., Johannes, F. J., Songyang, Z. and Cantley, L. C. (1997) Determination of the specific substrate sequence motifs of protein kinase C isozymes. *J Biol Chem*, **272**, 952-960.

Obata, T., Yaffe, M. B., Leparc, G. G., Piro, E. T., Maegawa, H., Kashiwagi, A., Kikkawa, R. and Cantley, L. C. (2000) Peptide and protein library screening defines optimal substrate motifs for AKT/PKB. *J Biol Chem*, **275**, 36108-36115.

Obenauer, J. C., Cantley, L. C. and Yaffe, M. B. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, **31**, 3635-3641.

Oelschlaeger, P., Mayo, S. L. and Pleiss, J. (2005) Impact of remote mutations on metallo-beta-lactamase substrate specificity: implications for the evolution of antibiotic resistance. *Protein Sci*, **14**, 765-774.

Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P. and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, **127**, 635-648.

Pamer, E. and Cresswell, P. (1998) Mechanisms of MHC class I--restricted antigen processing. *Annu Rev Immunol*, **16**, 323-358.

Parang, K., Till, J. H., Ablooglu, A. J., Kohanski, R. A., Hubbard, S. R. and Cole, P. A. (2001) Mechanism-based design of a protein kinase inhibitor. *Nat Struct Biol*, **8**, 37-41.

Parham, P., Adams, E. J. and Arnett, K. L. (1995) The origins of HLA-A,B,C polymorphism. *Immunol Rev*, **143**, 141-180.

Parker, K. C., Bednarek, M. A. and Coligan, J. E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*, **152**, 163-175.

Pawson, T. (1994) Introduction: protein kinases. *Faseb J*, **8**, 1112-1113.

Pawson, T. and Scott, J. D. (1997) Signaling through scaffold, anchoring, and adaptor proteins. *Science*, **278**, 2075-2080.

Pawson, T. and Nash, P. (2000) Protein-protein interactions define specificity in signal transduction. *Genes Dev*, **14**, 1027-1047.

Pawson, T., Raina, M. and Nash, P. (2002) Interaction domains: from simple binding events to complex cellular behavior. *FEBS Lett*, **513**, 2-10.

Plowman, G. D., Sudarsanam, S., Bingham, J., Whyte, D. and Hunter, T. (1999) The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms. *Proc Natl Acad Sci U S A*, **96**, 13603-13610.

Pohlmann, T., Bockmann, R. A., Grubmuller, H., Uchanska-Ziegler, B., Ziegler, A. and Alexiev, U. (2004) Differential peptide dynamics is linked to major histocompatibility complex polymorphism. *J Biol Chem*, **279**, 28197-28201.

Prager, K., Wang-Eckhardt, L., Fluhrer, R., Killick, R., Barth, E., Hampel, H., Haass, C. and Walter, J. (2007) A structural switch of presenilin 1 by GSK-3beta mediated phosphorylation regulates the interaction with beta -catenin and its nuclear signaling. *J Biol Chem*, **282**, 14083-14093

Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A. and Stevanovic, S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213-219.

Rammensee, H. G., Falk, K. and Rotzschke, O. (1993) Peptides naturally presented by MHC class I molecules. *Annu Rev Immunol*, **11**, 213-244.

Reche, P. A., Glutting, J. P. and Reinherz, E. L. (2002) Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol*, **63**, 701-709.

Reche, P. A., Glutting, J. P., Zhang, H. and Reinherz, E. L. (2004) Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics*, **56**, 405-419.

Robinson, J., Waller, M. J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L. J., Stoehr, P. and Marsh, S. G. (2003) IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res*, **31**, 311-314.

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

Saunders, N. F. and Kobe, B. (2008) The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. *Nucleic Acids Res*, **36**, W286-W290

Schueler-Furman, O., Elber, R. and Margalit, H. (1998) Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes. *Fold Des*, **3**, 549-564.

Schueler-Furman, O., Altuvia, Y., Sette, A. and Margalit, H. (2000) Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci*, **9**, 1838-1846.

Schuler, M. M., Nastke, M. D. and Stevanovikc, S. (2007) SYFPEITHI: database for searching and T-cell epitope prediction. *Methods Mol Biol*, **409**, 75-93.

Sette, A., Buus, S., Appella, E., Smith, J. A., Chesnut, R., Miles, C., Colon, S. M. and Grey, H. M. (1989) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc Natl Acad Sci U S A*, **86**, 3296-3300.

Shifman, J. M. and Mayo, S. L. (2002) Modulating calmodulin binding specificity through computational protein design. *J Mol Biol*, **323**, 417-423.

Shimaoka, M., Shifman, J. M., Jing, H., Takagi, J., Mayo, S. L. and Springer, T. A. (2000) Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat Struct Biol*, **7**, 674-678.

Singh, H. and Raghava, G. P. (2001) ProPred: prediction of HLA-DR binding sites. *Bioinformatics*, **17**, 1236-1237.

Singh, H. and Raghava, G. P. (2003) ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics*, **19**, 1009-1014.

Smith, C. M., Shindyalov, I. N., Veretnik, S., Gribskov, M., Taylor, S. S., Ten Eyck, L. F. and Bourne, P. E. (1997) The protein kinase resource. *Trends Biochem Sci*, **22**, 444-446.

Songyang, Z., Blechner, S., Hoagland, N., Hoekstra, M. F., Piwnica-Worms, H. and Cantley, L. C. (1994) Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr Biol*, **4**, 973-982.

Songyang, Z., Lu, K. P., Kwon, Y. T., Tsai, L. H., Filhol, O., Cochet, C., Brickey, D. A., Soderling, T. R., Bartleson, C., Graves, D. J. *et al.* (1996) A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1. *Mol Cell Biol*, **16**, 6486-6493.

Street, A. G., Datta, D., Gordon, D. B. and Mayo, S. L. (2000) Designing protein beta-sheet surfaces by Z-score optimization. *Phys Rev Lett*, **84**, 5010-5013.

Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M. P., Sinigaglia, F. *et al.* (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol*, **17**, 555-561.

Sung, M. H., Zhao, Y., Martin, R. and Simon, R. (2002) T-cell epitope prediction with combinatorial peptide libraries. *J Comput Biol*, **9**, 527-539.

Toes, R. E., Nussbaum, A. K., Degermann, S., Schirle, M., Emmerich, N. P., Kraft, M., Laplace, C., Zwinderman, A., Dick, T. P., Muller, J. *et al.* (2001) Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J Exp Med*, **194**, 1-12.

Tong, J. C., Tan, T. W. and Ranganathan, S. (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci*, **13**, 2523-2532.

Trivedi, O. A., Arora, P., Vats, A., Ansari, M. Z., Tickoo, R., Sridharan, V., Mohanty, D. and Gokhale, R. S. (2005) Dissecting the mechanism and assembly of a complex virulence mycobacterial lipid. *Mol Cell*, **17**, 631-643.

Ubersax, J. A. and Ferrell, J. E., Jr. (2007) Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol*, **8**, 530-541.

Udaka, K., Wiesmuller, K. H., Kienle, S., Jung, G., Tamamura, H., Yamagishi, H., Okumura, K., Walden, P., Suto, T. and Kawasaki, T. (2000) An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. *Immunogenetics*, **51**, 816-828.

Wan, J., Kang, S., Tang, C., Yan, J., Ren, Y., Liu, J., Gao, X., Banerjee, A., Ellis, L. B. and Li, T. (2008) Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. *Nucleic Acids Res*, **36**, e22.

Ward, P., Equinet, L., Packer, J. and Doerig, C. (2004) Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. *BMC Genomics*, **5**, 79.

Watts, C. (1997) Capture and processing of exogenous antigens for presentation on MHC molecules. *Annu Rev Immunol*, **15**, 821-850.

Wilson, D. B., Pinilla, C., Wilson, D. H., Schroder, K., Boggiano, C., Judkowski, V., Kaye, J., Hemmer, B., Martin, R. and Houghten, R. A. (1999) Immunogenicity. I. Use of peptide libraries to identify epitopes that activate clonotypic CD4+ T cells and induce T cell responses to native peptide ligands. *J Immunol*, **163**, 6424-6434.

Wong, Y. H., Lee, T. Y., Liang, H. K., Huang, C. M., Wang, T. Y., Yang, Y. H., Chu, C. H., Huang, H. D., Ko, M. T. and Hwang, J. K. (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res*, **35**, W588-594.

Wooten, M. W. (2002) In-gel kinase assay as a method to identify kinase substrates. *Sci STKE*, **2002**, PL15.

Xue, Y., Zhou, F., Zhu, M., Ahmed, K., Chen, G. and Yao, X. (2005) GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res*, **33**, W184-187.

Xue, Y., Li, A., Wang, L., Feng, H. and Yao, X. (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, **7**, 163.

Yang, C. Y., Chang, C. H., Yu, Y. L., Lin, T. C., Lee, S. A., Yen, C. C., Yang, J. M., Lai, J. M., Hong, Y. R., Tseng, T. L. *et al.* (2008) PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*, **24**, i14-20.

Yang, J., Cron, P., Good, V. M., Thompson, V., Hemmings, B. A. and Barford, D. (2002) Crystal structure of an activated Akt/protein kinase B ternary complex with GSK3-peptide and AMP-PNP. *Nat Struct Biol*, **9**, 940-944.

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

Yellaboina, S., Goyal, K. and Mande, S. C. (2007) Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: comparison with high-throughput experimental data. *Genome Res*, **17**, 527-535.

Yewdell, J. W., Reits, E. and Neefjes, J. (2003) Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nat Rev Immunol*, **3**, 952-961.

Zaitlen, N., Reyes-Gomez, M., Heckerman, D. and Jojic, N. (2008) Shift-invariant adaptive double threading: learning MHC II-peptide binding. *J Comput Biol*, **15**, 927-942.

Zanzoni, A., Ausiello, G., Via, A., Gherardini, P. F. and Helmer-Citterich, M. (2007) Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res*, **35**, D229-231.

Zhang, H., Zha, X., Tan, Y., Hornbeck, P. V., Mastrangelo, A. J., Alessi, D. R., Polakiewicz, R. D. and Comb, M. J. (2002) Phosphoprotein analysis using antibodies broadly reactive against phosphorylated motifs. *J Biol Chem*, **277**, 39379-39387.

Zheng, J., Knighton, D. R., ten Eyck, L. F., Karlsson, R., Xuong, N., Taylor, S. S. and Sowadski, J. M. (1993) Crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MgATP and peptide inhibitor. *Biochemistry*, **32**, 2154-2161.

Zhou, F. F., Xue, Y., Chen, G. L. and Yao, X. (2004) GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem Biophys Res Commun*, **325**, 1443-1448.