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
Introduction

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 $2.20 \times 10^{-16}$, $5.07 \times 10^{-6}$ and $2.34 \times 10^{-8}$ 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 forcefield 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.