STRUCTURE BASED SUBSTRATE IDENTIFICATION OF PROTEIN KINASE

Submitted in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
BIOINFORMATICS
By

Mandeep Singh, Rashika Singhal

(151501, 151508)

UNDER THE GUIDANCE OF

Dr. Narendra Kumar

Department of biotechnology and bioinformatics

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,WAKNAGHAT

May,2019

ACKNOWLEGMENT

The work on this project has been a good, inspiring and exciting experience for us throughout the whole year, though it was challenging sometimes but we successfully completed our work within time. We are using this opportunity to express our gratitude to everyone who supported throughout the project (Substrate idendtification of Protein Kinases), their aspiring guidance, invaluably constructive criticism and friendly advice helped us a lot during the whole time. We are sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

We would like to express our warm thanks to our guide Dr. Narendra Kumar for his support and guidance helping in completing this project. We would also like to thank the officials of Jaypee University of Information technology (JUIT), Waknaghat for their help and cooperation. At last we would like to thank our parents and dear friends/batch mates for believing us and encouraging us through the whole.

DECLARATION BY SCHOLAR

I hereby declare that the thesis entitled"**Substrate Identification of Protein Kinase"** submitted at the **Jaypee University of Information Technology , Waknaghat, India** is the record of work carried out by me under the guidance of "**Dr. Narendra Kumar**". I have not submitted this work elsewhere for any degree or diploma . I am fully responsible for the content of my B-Tech thesis.

Mandeep singh(151501)

Rashika singhal(151508)

Jaypee University of Information Technology Waknaghat, India

Date:

CERTIFICATE

This is to certify that the project report entitled "**Structure Based Substrate Identification of Protein Kinase**", submitted by Mandeep Singh and Rashika Singhal at Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other university or Institute for the award of this or any other degree or diploma.

Signature of Supervisor

Name of Supervisor: Dr. Narendra Kumar
Designation: Assistant Professor,
Jaypee University of Information Technology,
Waknaghat Solan, Himachal Pradesh

CONTENTS

LIST OF FIGURES

LIST OF TABLES

RATIONALE

A protein kinase is a kinase enzyme that alter other molecules, most of them are proteins, by chemically adding phosphate groups to them (phosphorylation). The chemical activity of a kinase includes transferring a phosphate group from a nucleoside triphosphate (usually ATP) and covalently appending it to specific amino acids with a free hydroxyl group. Most kinases follow up on serine and threonine(serine/threonine kinases), others follow up on tyrosine (tyrosine kinases), and a number follow up on all of them (dual-specificity kinases). The substrate is perceived by kinase through the interactions of residues around the phosphorylation site in the substrate with the residues in the protein kinase. As we know that peptide is present on the surface of the substrate hence, we need to collect the x-ray structures of kinase-substrate peptide complexes. Through, these structures we will be able to identify the interaction between the peptide and the protein by visualising the structure. We will use these interactions to predict whether the unknown substrate will bind or not. To evaluate the potential of unknown peptide to be a substrate, you can model it in the active site and score the amino-amino interactions. Summing the interaction scores will give you the score of peptide. Peptides with a score above a threshold may be categorised as the substrates. (These scores have to be benchmarked against known substrates).

CHAPTER 1

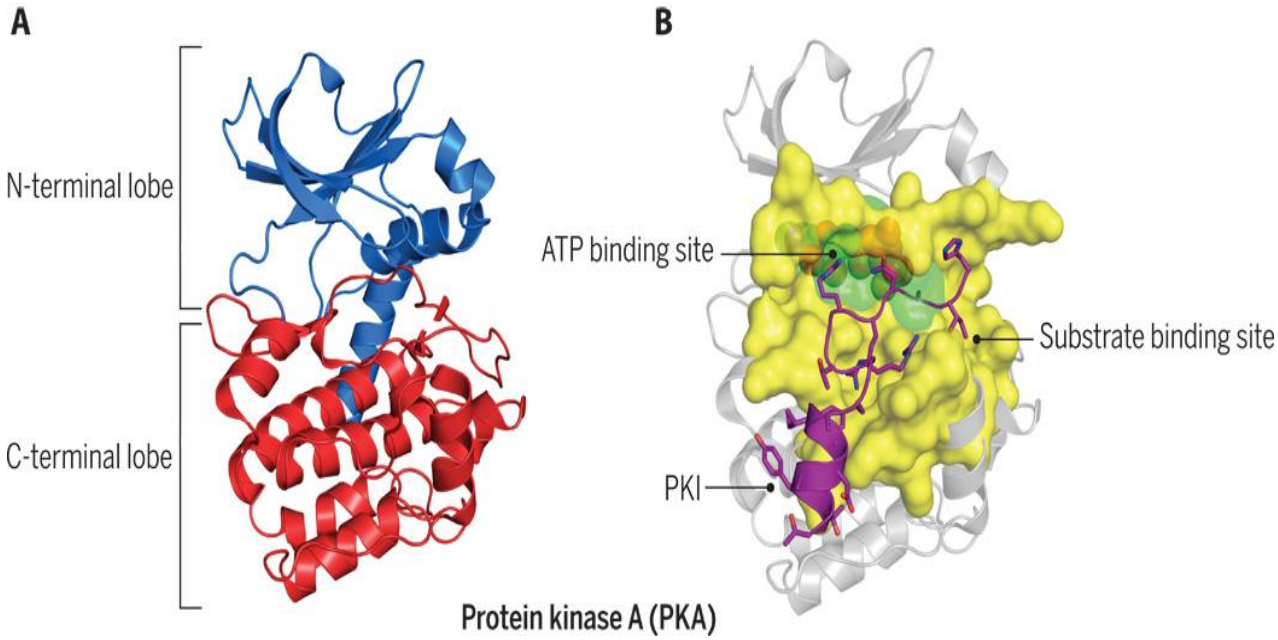
INTRODUCTION

General Background

Protein kinases play a nearly universal role in cellular regulation and are rising as an important category of recent drug targets, nonetheless the cellular functions of most human kinases mostly remain unknown. Aspects of substrate recognition common to any or all kinases in the ATP nucleotide binding site have been exploited in the generation of analog-specific mutants for exploring kinase function and discovering novel super molecule(protein) substrates. Protein Kinase can modify the function of a protein in nearly each possible way[1]. The human protein kinase gene family comprises of 518 members along with 106 pseudogenes[2]. The SER/THR protein kinases are in enormous majority. The SER/THR protein kinases interact with various substrates starting from enzymes, including other kinases, to transcription factors, receptors, and different regulatory proteins. Thus, mechanisms to assure specificity should be present. However, from rising structural knowledge it is becoming apparent that the ways in which protein kinases interact with their substrates local to the active site are comparatively few. Instead, docking interactions, in pockets or grooves outside the active site of the kinase, are used to identify substrates and different interacting proteins. Protein phosphorylation is a common posttranslational modification and is concerned with several physiological and pathophysiological processes. Among other diseases, the liberation of protein kinase activities may cause cellular transformation and cancer.

Thus, kinases are major drug targets. Seeing how kinases interact with their substrates may explain the processes that leads to ailment, just as help in the advancement of better, more specific kinase inhibitors with improved clinical achievements.

Protein kinases are made up of non-conserved regulatory domains and a conserved catalytic core of around 250 amino acid residues that binds and anchors substrates and is in charge for catalysis. The catalytic domain comprises of two lobes known as N and C (also called small and large lobes, respectively), named for their N- or C-terminal position, respectively, within the domain. The N-lobe comprises of five-stranded, anti-parallel β sheets that are a vital part of the adenosine triphosphate (ATP) binding site, whereas the C-lobe is generally coiled or helical .The active-site cleft, that contains the ATP binding site, lies between the two lobes. In an activated kinase, the lobes converge to make a deep cleft where ever the adenine ring of ATP binds such that the phosphate is positioned at the fringes where the transfer of the phosphoryl group takes place, whereas the adenosine moiety is buried in an exceedingly hydrophobic region of the pocket. Adjoining to the ATP binding pocket is a shallow crevice called the substrate binding site (SBS) that anchors the substrate and accurately positions the phosphorylatable residue.

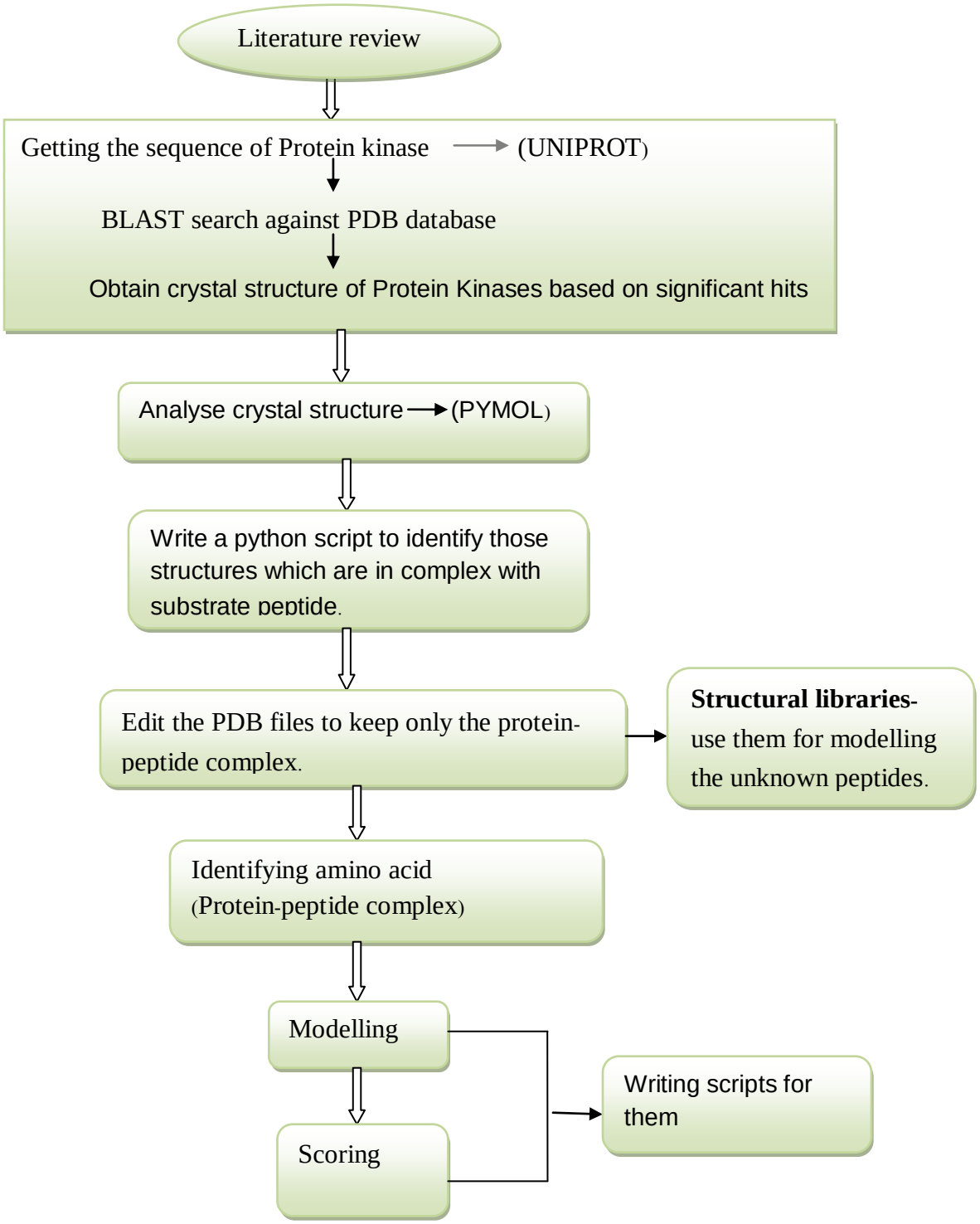


<u>Figure 1.1</u> Structure of protein kinases binding to its substrate. Image adapted from [2].

Catalysis is interceded by opening and closing of this active-site cleft. Substrates are anchored and positioned close to this cleft in order that the hydroxyl group of the phosphoryl table residue (termed P0) can accept the phosphate. Flanking regions helps in stabilizing the active kinase and are also essential for catalysis. Tyrosine kinases have a profound cleft crevice around P0 than serine/threonine (Ser/Thr) kinases to more readily oblige a massive side chain. .An increase in the catalytic activity of kinases typically results in cancer therefore, their activation must be firmly regulated.[3]

CHAPTER 2

MATERIAL AND METHOD



2.1 Data Retrieval

2.1.1 UNIPROT

It is freely accessible and collaboration of databases and contains huge information about the biological and molecular function. It provides researchers with a extensive, high-quality and freely accessible resource of protein sequence and functional information. Therefore, we have utilised this database to retrieve the information of protein kinase(PDB ID: Q9NYL2).

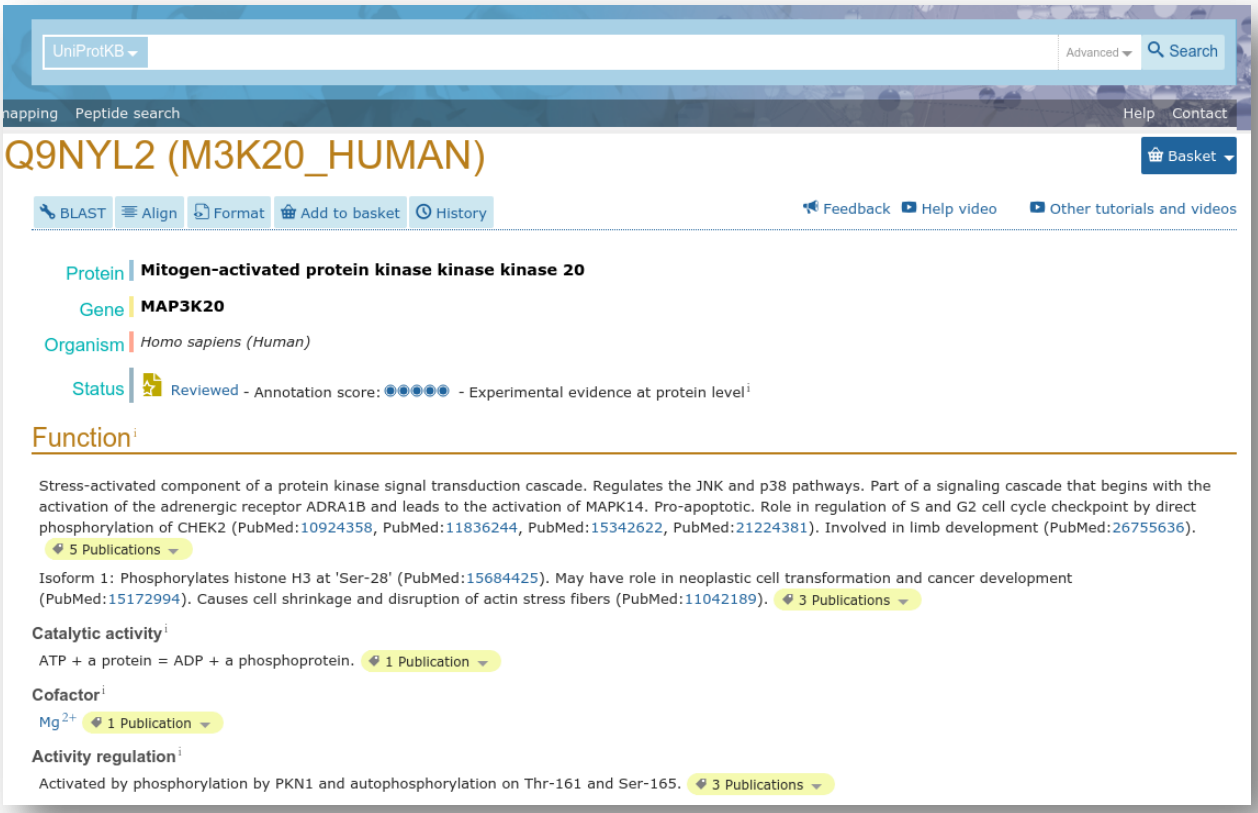


Figure 2.1: Protein Kinase in Uniprot

2.1.2 BLAST

Basic local alignment tool (BLAST) one of the most favoured choices for searching and aligning sequences. Blast identify region of similarity between various biological sequences. The program

does the comparision between nucleotide or protein sequences with sequence databases and evaluates the statistical significances.Therefore, performed blast against the protein data bank.

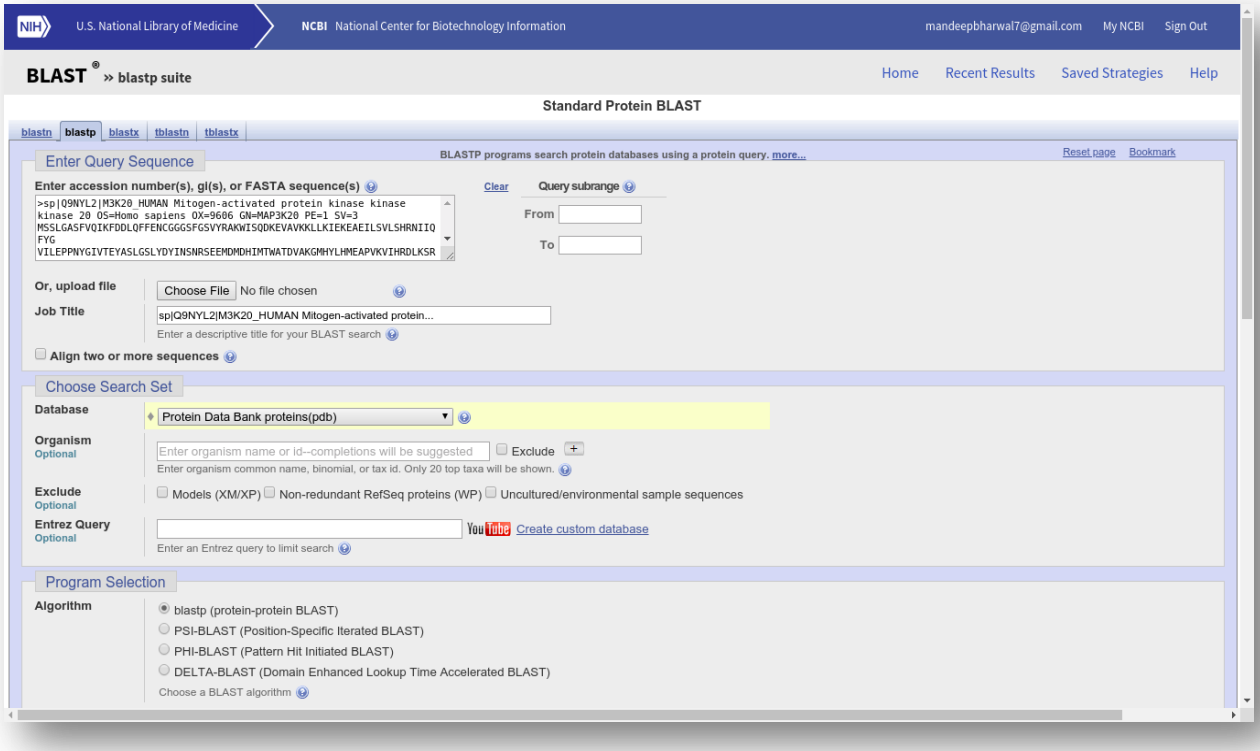


Figure 2.2 : BLAST against PDB

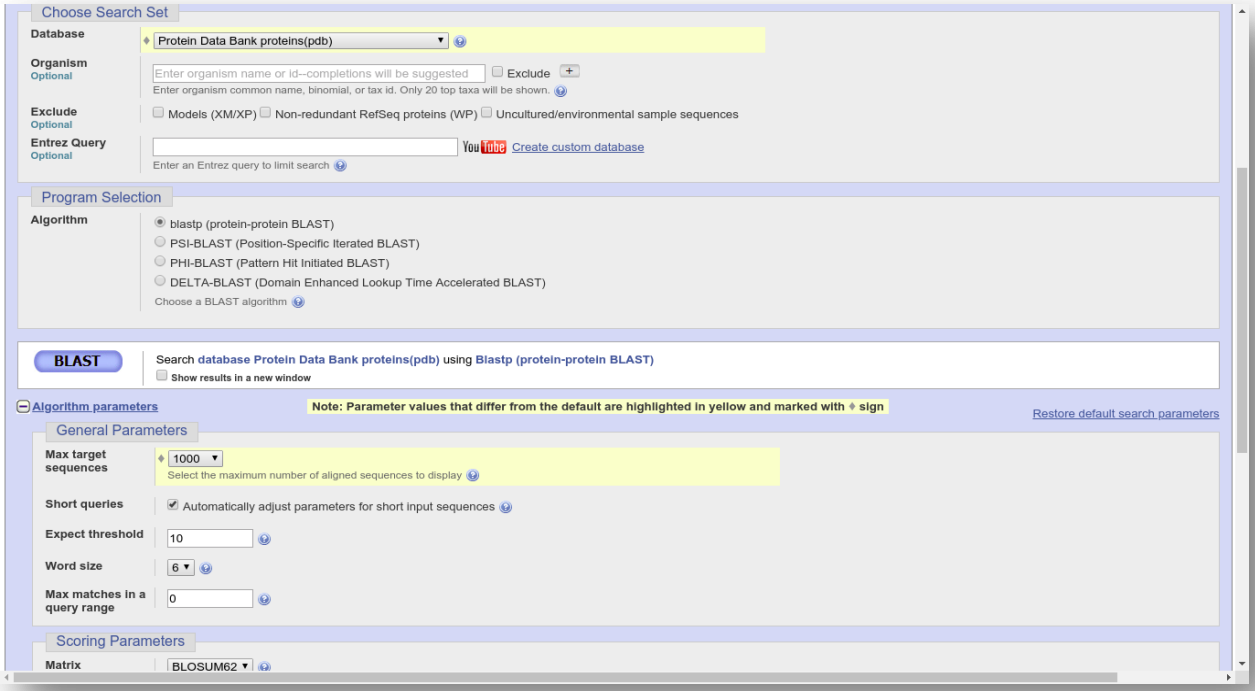


Figure2.3: Blast for1000 target sequences setting E-value 10.

- Performed blast with E-Value equals to 10 and retrieved first initial 1000 sequences. And further wrote the python script to
  - Download the PDB files,
  - To separate the chain ID and chains
  - File cleaning(kept only atoms)
  - To count the number of residues in chains
  - To retrieve the short peptides which have length less than 15.

- Visualized the pdb ids using PYMOL software and thus identified the protein peptide complex.

Table 2.1 Showing substrate binding with protein kinase

PDB IDs	Title of the structure	Size of the Substrates	Known Substrate binding to kinases	Type
1ZYS	CO-CRYSTAL STRUCTURE OF CHECKPOINT KINASE CHK1 WITH A PYRROLO-PYRIDINE 2 INHIBITOR	5	Chain ID B	SER/THR
4JDJ	CRYSTAL STRUCTURE OF SERINE/THREONINE-PROTEIN KINASE PAK 4 F461V 2 MUTANT IN COMPLEX WITH PAKTIDE T PEPTIDE SUBSTRATE	6	Chain ID E	SER/THR
1QMZ	PHO SPHORYLATED CDK2-CYCLYIN A-SUBSTRATE PEPTIDE COMPLEX	7	Chain ID B	TYR
2PHK	THE CRYSTAL STRUCTURE OF A PHOSPHORYLASE KINASE PEPTIDE SUBSTRATE 2 COMPLEX: KINASE SUBSTRATE RECOGNITION	7	Chain ID C	SER/THR
4FIF	CATALYTIC DOMAIN OF HUMAN PAK4 WITH RPKPLVDP PEPTIDE	7	Chain ID B	SER/THR
2Y8O	CRYSTAL STRUCTURE OF HUMAN P38ALPHA	8	Chain ID I	Dual specificity
3O71	CRYSTAL STRUCTURE OF ERK2/DCC PEPTIDE COMPLEX	8	Chain ID I	Dual Specificity
4UX9	CRYSTAL STRUCTURE OF JNK1 BOUND TO A MKK7 DOCKING MOTIF	8	Chain ID C	Dual Specificity
5ETF	STRUCTURE OF DEAD KINASE MAPK14 WITH BOUND THE KIM DOMAIN OF MKK6	8	Chain ID C	Dual specificity
5DE2	STRUCTURAL MECHANISM OF NEK7 ACTIVATION BY NEK9-INDUCED DIMERISATION	9	Chain ID B	Serine/THR
1LEW	CRYSTAL STRUCTURE OF MAP KINASE P38 COMPLEXED TO THE DOCKING SITE ON 2 ITS NUCLEAR SUBSTRATE MEF2A	10	Chain ID B	SER/THR
1O6K	STRUCTURE OF ACTIVATED FORM OF PKB KINASE DOMAIN S474D WITH 2 GSK3 PEPTIDE AND AMP-PNP	10	Chain ID C	SER/THR
1O6L	CRYSTAL STRUCTURE OF AN ACTIVATED AKT/PROTEIN KINASE	10	Chain ID F	SER/THR

2G01	PYRAZOLOQUINOLONES AS NOVEL, SELECTIVE JNK1	10	Chain ID C	TYR
3CQU	CRYSTAL STRUCTURE OF AKT-1 COMPLEXED WITH SUBSTRATE PEPTIDE 2 AND INHIBITOR	10	Chain ID C	SER/THR
2JDO	STRUCTURE OF PKB-BETA (AKT2) COMPLEXED WITH ISOQUINOLINE-5- 2 SULFONIC ACID (2-(2-(4-CHLOROBENZYLOXY) ETHYLAMINO)ETHYL)	10	Chain ID C	SER/THR
3E87	CRYSTAL STRUCTURES OF THE KINASE DOMAIN OF AKT2 IN COMPLEX WITH ATP- 2 COMPETITIVE INHIBITORS	10	Chain ID F	SER/THR
3O17	CRYSTAL STRUCTURE OF JNK1-ALPHA1 ISOFORM	10	Chain ID C	TYR
3OCB	AKT1 KINASE DOMAIN WITH PYRROLOPYRIMIDINE INHIBITOR	10	Chain ID J	TYR
3PTG	DESIGN AND SYNTHESIS OF A NOVEL, ORALLY EFFICACIOUS TRI-SUBSTITUTED	10	Chain ID K	TYR
3QHR	STRUCTURE OF A PCDK2/CYCLINA TRANSITION-STATE MIMIC	10	Chain ID F	SER/THR
3VUD	CRYSTAL STRUCTURE OF A CYSTEINE-DEFICIENT MUTANT	10	Chain ID F	TYR
3VUG	CRYSTAL STRUCTURE OF A CYSTEINE-DEFICIENT MUTANT M2 IN MAP KINASE JNK1	10	Chain ID F	TYR
3VUH	CRYSTAL STRUCTURE OF A CYSTEINE-DEFICIENT MUTANT M3 IN MAP KINASE JNK1	10	Chain ID D	TYR
2FYS	CRYSTAL STRUCTURE OF ERK2 COMPLEX WITH KIM PEPTIDE DERIVED 2 FROM MKP3	11	Chain ID P	Dual
3P4K	THE THIRD CONFORMATION OF P38A MAP KINASE OBSERVED IN PHOSPHORYLATED 2 P38A AND IN SOLUTION	11	Chain ID V	SER/THR
3V3V	STRUCTURAL AND FUNCTIONAL ANALYSIS OF QUERCETAGETIN, A..	11	Chain ID C	SER/THR
2B9H	CRYSTAL STRUCTURE OF FUS3 WITH A DOCKING MOTIF	12	Chain ID B	SER/THR
2Q0N	STRUCTURE OF HUMAN P21 ACTIVATING KINASE 4 (PAK4) IN COMPLEX WITH A 2 PEP cons.	12	Chain ID B	SER/THR
2XRW	LINEAR BINDING MOTIFS FOR JNK AND FOR CALCINEURIN ANTAGONISTICALLY 2 CONTROL THE NUCLEAR SHUTTLING OF NFAT4	12	Chain ID B	SER/THR

2XS0	LINEAR BINDING MOTIFS FOR JNK AND FOR CALCINEURIN ANTAGONISTICALLY 2 CONTROL THE NUCLEAR SHUTTLING OF NFAT4	12	Chain ID B	SER/THR
4XBU	IN VITRO CRYSTAL STRUCTURE OF PAK4 IN COMPLEX WITH INKA PEPTIDE	13	Chain ID B	SER/THR
5N37	CAMP-DEPENDENT PROTEIN KINASE A FROM CRICETULUS	13	Chain ID I	DUAL
5V62	CRYSTAL STRUCTURE OF PHOSPHOLAMBAN (1-19):PKA C-SUBUNIT:AMP-PNP:MG2+ 2 COMPLEX	14	Chain ID B	DUALSpecificity
3O7L	CRYSTAL STRUCTURE OF AKT-1 COMPLEXED WITH SUBSTRATE PEPTIDE 2 AND INHIBITOR	15	Chain ID B	SER/THR

- **Calculating amino acid-amino acid contact preferences at the interface of protein kinases and their substrate.**

  After visualizing the protein peptide complexes in pymol, Calculated amino acid – amino acid contact preferences at the interface of the protein kinases and their substrate. Identified all possible amino acid - amino acid contacts from the crystal structures of protein kinase- substrate peptide complexes. The two residues were said to be in contact if they were less than 6Angstrom apart.All the residue residue contacts between peptide and protein were identified in all the complexes in the data set using the python program. Binding preferences of all amino acids were calculated as the log ratios of observed / expected frequencies. Observed frequencies were calculated from the count of amino acid-amino acid contact pair at the interface. Expected frequencies were calculated from the frequency of individual amino acid at the interface. A 20 X 20 matrix was calculated representing the binding preferences of residues which is specific to protein kinase – peptide interface.

Table 2.2 Distances among the protein kinase and its substrate less than 6 amstrong.

RESIDUES	RES SEQ	CHAIN	RESIDUES	RES SEQ	CHAIN	DISTANCE
LEU	5	B ⟶	ASP	161	A	5.8393087
ARG	6	B ⟶	GLU	160	A	5.752933
ARG	6	B ⟶	ASP	161	A	4.815948
ARG	6	B ⟶	CYS	162	A	5.742791
VAL	7	B ⟶	GLU	160	A	4.8825483
VAL	7	B ⟶	CYS	162	A	5.700053
cVAL	8	B ⟶	GLU	160	A	5.9848228

Table 2.3 Calculated distances of residues coming from protein kinase in structural library.

	ALA	ARG	ASN	ASP	CYS	GLU	GLN	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	89	23	10	21	12	22	12	14	10	15	26	22	13	19	12	9	11	7	22	26
ARG	23	53	11	21	4	16	9	14	12	17	23	14	9	11	10	11	10	1	15	17
ASN	10	8	47	12	5	9	9	11	12	10	27	21	3	11	6	12	6	2	7	15
ASP	20	18	13	61	5	11	12	15	6	13	32	24	5	10	9	22	8	7	14	19
CYS	8	2	4	4	19	5	5	8	2	10	13	9	6	2	4	5	2	1	3	8
GLU	22	14	10	10	5	82	10	15	10	29	34	28	15	20	13	18	12	4	16	18
GLN	12	9	10	12	5	9	44	9	4	16	28	14	2	10	5	6	4	1	11	19
GLY	14	15	12	18	8	18	10	57	4	24	22	12	11	10	7	13	8	4	10	15
HIS	9	12	12	6	3	10	5	4	29	12	11	14	8	8	6	11	4	2	4	4
ILE	16	19	10	13	12	29	15	25	13	93	34	33	9	12	15	16	7	6	17	25
LEU	28	23	27	40	20	37	27	25	10	32	136	46	25	24	10	36	9	10	19	45
LYS	22	14	21	25	9	28	14	12	14	33	46	107	13	26	11	19	10	6	14	33
MET	13	9	3	5	6	15	2	11	8	9	25	13	34	5	4	7	2	1	7	20
PHE	18	11	11	10	7	16	10	14	8	12	25	26	5	55	14	15	10	4	10	16
PRO	12	9	6	10	4	13	5	7	6	15	10	11	4	14	57	11	7	0	9	10
SER	9	11	12	21	7	20	6	17	11	14	24	17	7	13	12	46	12	4	13	10
THR	16	9	7	25	12	13	4	9	5	7	7	8	1	10	6	13	42	4	11	16
TRP	7	1	2	7	1	4	1	6	2	6	10	6	1	4	0	3	5	14	2	10
TYR	22	15	7	15	4	16	11	10	4	17	19	14	7	10	9	15	11	2	58	24
VAL	26	17	15	19	9	22	17	15	4	25	44	33	20	15	10	9	16	9	22	78

- **Finding the nature of interface in protein kinases when it binds to the peptide**

  Calculated the interaction frequency of each amino acids present on the interface of protein kinases and peptide. measured the expected values from these frequencies and thus computed the log score for every possible amino acid - amino acid interaction and prepared the pair potential matrix , using formula.

Log(Observed/Expected)

Table 2.4 Expected No. of context in the in the library.

	ALA	ARG	ASN	ASP	CYS	GLU	GLN	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
ARG	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
ASN	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
ASP	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
CYS	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
GLU	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
GLN	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
GLY	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
HIS	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
ILE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
LEU	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
LYS	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
MET	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
PHE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
PRO	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
SER	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
THR	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
TRP	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
TYR	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 2.5 Observed No. of context in the library.

	ALA	ARG	ASN	ASP	CYS	GLU	GLN	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.22	0.06	0.03	0.05	0.03	0.06	0.03	0.04	0.03	0.04	0.07	0.06	0.03	0.05	0.03	0.02	0.03	0.02	0.06	0.07
ARG	0.06	0.13	0.03	0.05	0.01	0.04	0.02	0.04	0.03	0.04	0.06	0.04	0.02	0.03	0.03	0.03	0.03	0.00	0.04	0.04
ASN	0.03	0.02	0.12	0.03	0.01	0.02	0.02	0.03	0.03	0.03	0.07	0.05	0.01	0.03	0.02	0.03	0.02	0.01	0.02	0.04
ASP	0.05	0.05	0.03	0.15	0.01	0.03	0.03	0.04	0.02	0.03	0.08	0.06	0.01	0.03	0.02	0.06	0.02	0.02	0.04	0.05
CYS	0.02	0.01	0.01	0.01	0.05	0.01	0.01	0.02	0.01	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.01	0.02
GLU	0.06	0.04	0.03	0.03	0.01	0.21	0.03	0.04	0.03	0.07	0.09	0.07	0.04	0.05	0.03	0.05	0.03	0.01	0.04	0.05
GLN	0.03	0.02	0.03	0.03	0.01	0.02	0.11	0.02	0.01	0.04	0.07	0.04	0.01	0.03	0.01	0.02	0.01	0.00	0.03	0.05
GLY	0.04	0.04	0.03	0.05	0.02	0.05	0.03	0.14	0.01	0.06	0.06	0.03	0.03	0.03	0.02	0.03	0.02	0.01	0.03	0.04
HIS	0.02	0.03	0.03	0.02	0.01	0.03	0.01	0.01	0.07	0.03	0.03	0.04	0.02	0.02	0.02	0.03	0.01	0.01	0.01	0.01
ILE	0.04	0.05	0.03	0.03	0.03	0.07	0.04	0.06	0.03	0.23	0.09	0.08	0.02	0.03	0.04	0.04	0.02	0.02	0.04	0.06
LEU	0.07	0.06	0.07	0.10	0.05	0.09	0.07	0.06	0.03	0.08	0.34	0.12	0.06	0.06	0.03	0.09	0.02	0.03	0.05	0.11
LYS	0.06	0.04	0.05	0.06	0.02	0.07	0.04	0.03	0.04	0.08	0.12	0.27	0.03	0.07	0.03	0.05	0.03	0.02	0.04	0.08
MET	0.03	0.02	0.01	0.01	0.02	0.04	0.01	0.03	0.02	0.02	0.06	0.03	0.09	0.01	0.01	0.02	0.01	0.00	0.02	0.05
PHE	0.05	0.03	0.03	0.03	0.02	0.04	0.03	0.04	0.02	0.03	0.06	0.07	0.01	0.14	0.04	0.04	0.03	0.01	0.03	0.04
PRO	0.03	0.02	0.02	0.03	0.01	0.03	0.01	0.02	0.02	0.04	0.03	0.03	0.01	0.04	0.14	0.03	0.02	0.00	0.02	0.03
SER	0.02	0.03	0.03	0.05	0.02	0.05	0.02	0.04	0.03	0.04	0.06	0.04	0.02	0.03	0.03	0.12	0.03	0.01	0.03	0.03
THR	0.04	0.02	0.02	0.06	0.03	0.03	0.01	0.02	0.01	0.02	0.02	0.02	0.00	0.03	0.02	0.03	0.11	0.01	0.03	0.04
TRP	0.02	0.00	0.01	0.02	0.00	0.01	0.00	0.02	0.01	0.02	0.03	0.02	0.00	0.01	0.00	0.01	0.01	0.04	0.01	0.03
TYR	0.06	0.04	0.02	0.04	0.01	0.04	0.03	0.03	0.01	0.04	0.05	0.04	0.02	0.03	0.02	0.04	0.03	0.01	0.15	0.06
VAL	0.07	0.04	0.04	0.05	0.02	0.06	0.04	0.04	0.01	0.06	0.11	0.08	0.05	0.04	0.03	0.02	0.04	0.02	0.06	0.20

- **Calculated the binding score of substrate**

  Since, the pair potential matrix we calulated is specific to protein kinase- peptide interface, it could be used for calculating the binding score of a potential unknown peptide (substrate) for classifying it into a binder or a non- binder.The binding score for the peptide would be calculated as the sum of all its interactions with the protein kinase.The cut off for the binding score would be decided based on the benchmarking on the known peptides of the protein kinases.

Table2.6 List of kinases and associated PDB ids

Protein Kinases	PDB IDS
CHK1	1ZYS
PAK 4	4JDJ,2QON,4XBU
CDK2	1QMZ,3QHR
PHK1	2PHK
MAP2K6	2Y8O
MAPK1	3O71,5V62,2FYS
MAPK8/ JNK1	4UX9,2XRW,3V3V, 3O17,2GO1,3VUD,3VUG,3VUH
MAPK14	5ETF
MAPK10	3PTG
NEK7	5DE2
MEF2A	1LEW
AKT 1	3CQU,3OCB,3O7L
AKT 2	1O6K,1O6L,2JDO,3E87,
PKA Alpha	5N37
MAPK	2B9H,3P4K

- **Phospho.ELM**

  It is a relational database designed to store in vivo and in vitro phosphorylation data extracted from the scientific literature and phosphor proteomic analyses. It consists of 42 574 serine, threonine and tyrosine non-redundant phosphorylation sites. The conservation of the phosphosites can be envisioned directly on the multiple sequence alignment which is used for the score calculation. In addition, it also includes information for the phosphorylated residue, i.e. conservation score (CS) and the surface accessibility score which are either anticipated or measured . The data can be obtained directly by a user-friendly web interface.[4]

  Retrieved the known substrates of the kinases through Phospho.ELM and prepared the table given below.

Table 2.7 Known substrates of PAK group.

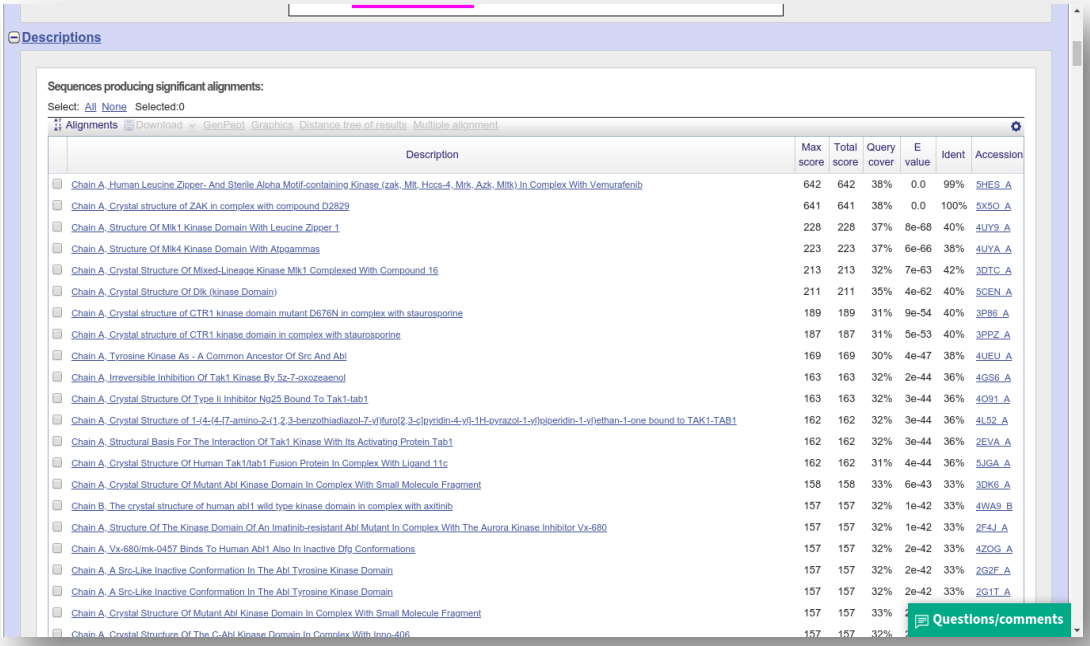
Accession	Residue	Position	Context
P35240	S	518	FKDTDMKRL**S**MEIEKEKVEY
P04049	S	338	KIRPRGQRD**S**SYYWEIEASE
P04049	S	338	KIRPRGQRD**S**SYYWEIEASE
P17600	S	605	GPAGPTRQA**S**QAGPVPRTGP
P08670	S	26	GPGTASRPSSSR**S**YVTTSTR
P08670	S	39	YVTTSTRTYSLGSALRP**S**TS
P08670	S	51	SALRPSTSR**S**LYASSPGGVY
P08670	S	56	STSRSLYASSPGGVYATR**S**S
P08670	S	56	STSRSLYASSPGGVYATR**S**S
P08670	S	66	PGGVYATRSSAVRLR**S**SVPG
P08670	S	73	RSSAVRLR**S**SVPGVRLLQDS

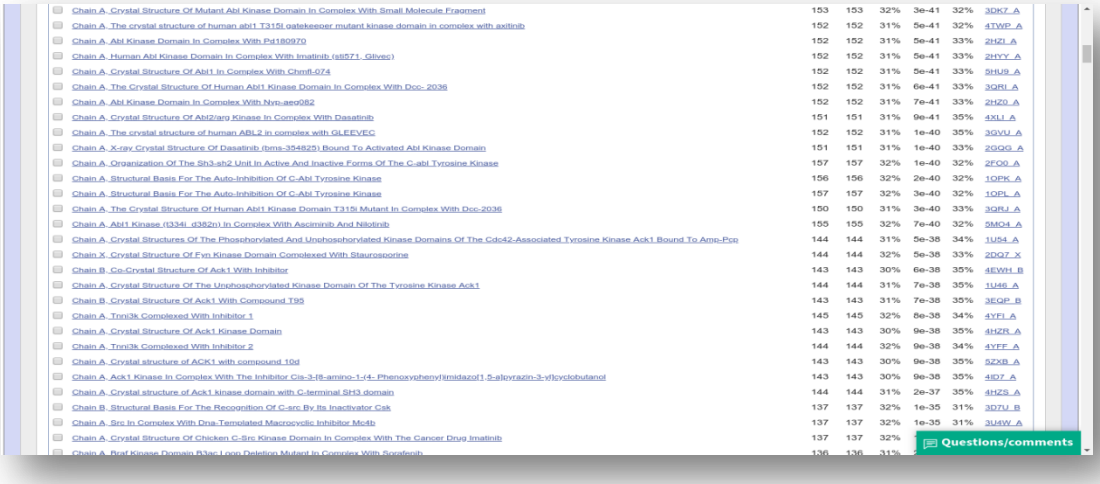
CHAPTER 3

RESULTS AND DISCUSSION

BLAST

Figure 3.1: BLAST hits





Chain ID Separation: Separated chain ids and chains using python script.

<u>Figure 3.2</u> Sorted chains with respect to length

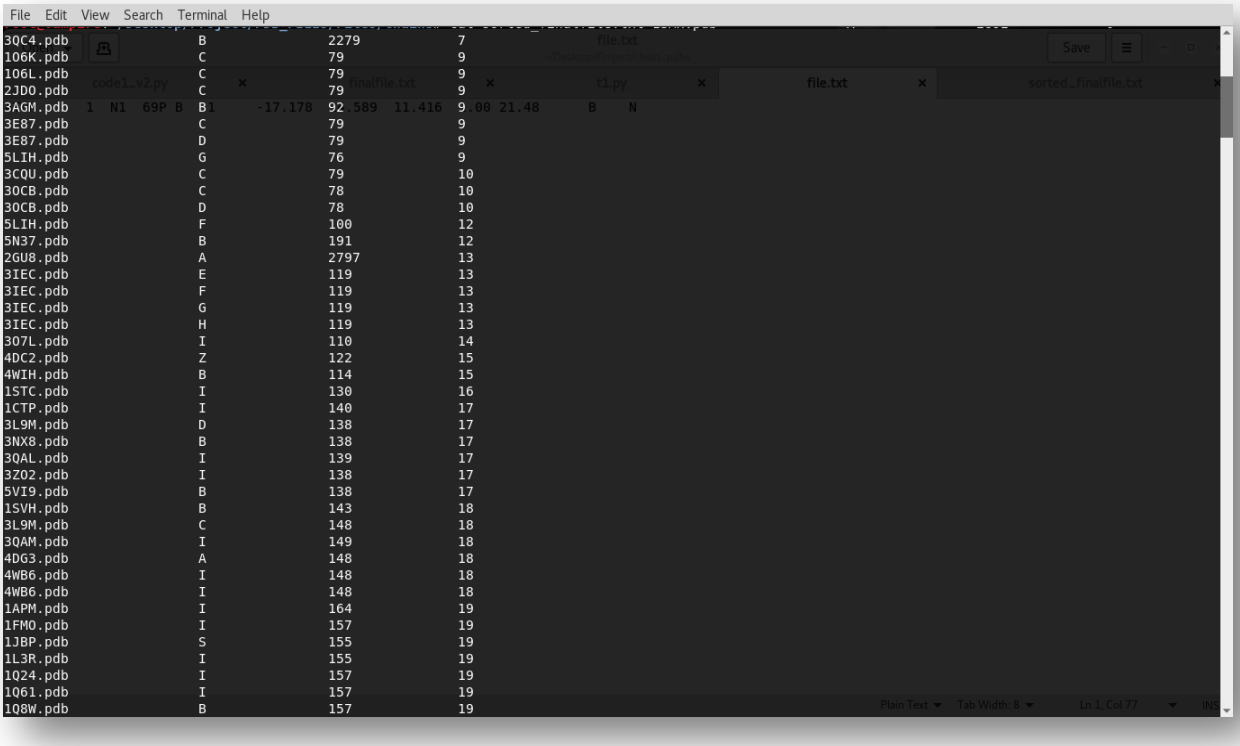


Table 3.1Selected PDB ids length less than 10

Pdb id	chain	length	residues length
2JAM.pdb	E	31	4
4AZE.pdb	E	25	4
4AZE.pdb	F	25	4
4AZE.pdb	G	25	4
1ZYS.pdb	B	30	5
4O27.pdb	C	47	5
2JAM.pdb	D	42	6
4JDJ.pdb	B	66	6
4NM5.pdb	C	44	6
5LW1.pdb	L	48	6
1QMZ.pdb	E	58	7
1QMZ.pdb	F	58	7
2PHK.pdb	B	65	7
4FIF.pdb	C	57	7
4FIF.pdb	D	57	7
2Y8O.pdb	B	56	8
3O71.pdb	B	64	8
4UX9.pdb	G	59	8
5ETF.pdb	B	59	8
3AGM.pdb	B	92	9
4UX9.pdb	I	77	9
5DE2.pdb	D	76	9
5LW1.pdb	C	72	9
5LW1.pdb	F	72	9
1LEW.pdb	B	79	10
1O6K.pdb	C	79	10
1O6L.pdb	C	79	10
1UKH.pdb	B	84	10
2G01.pdb	F	73	10
2G01.pdb	G	73	10
2JDO.pdb	C	79	10
3CQU.pdb	C	79	10
3E87.pdb	C	79	10
3E87.pdb	D	79	10
3O17.pdb	F	83	10
3O17.pdb	G	83	10
3OCB.pdb	C	78	10
3OCB.pdb	D	78	10
3OXI.pdb	J	83	10
3PTG.pdb	J	84	10
3QHR.pdb	J	79	10
3QHR.pdb	L	79	10
3QHR.pdb	M	79	10
3VUD.pdb	F	84	10
3VUG.pdb	F	84	10
3VUH.pdb	F	84	10
3VUI.pdb	F	84	10
3VUK.pdb	F	84	10
3VUL.pdb	F	84	10
4UX9.pdb	F	77	10
5LIH.pdb	G	76	10

- **Binding of Substrate with the protein kinase:**

  Separated chains from the PDB files. Poly-peptides and protein kinase are obtained. Now further using this data the analysis of where Substrate is binding to the protein kinase was done and observed in Pymol.

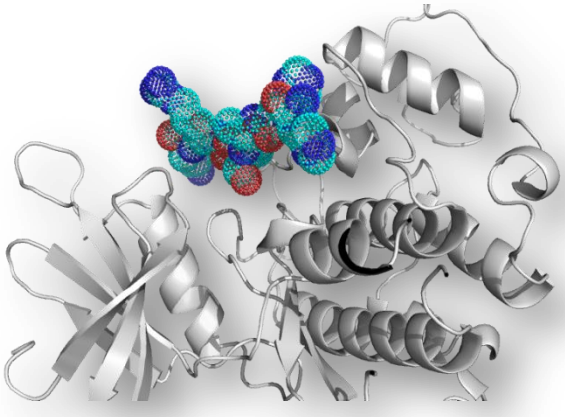


Figure 3.3(a) Substrate (chain E) binding with the protein kinase (chain A) in 1QMZ.

Figure 3.3(b) Substrate binding with kinase protein in different PDB ID 3E87, 3O17.

Table 3.2 Pair Potential matrix

	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	1.35	0.76	0.40	0.72	0.48	0.74	0.48	0.54	0.40	0.57	0.81	0.74	0.51	0.68	0.48	0.35	0.44	0.24	0.74	0.81
R	0.76	1.12	0.44	0.72	0.00	0.60	0.35	0.54	0.48	0.63	0.76	0.54	0.35	0.44	0.40	0.44	0.40	-0.60	0.57	0.63
N	0.40	0.30	1.07	0.48	0.10	0.35	0.35	0.44	0.48	0.40	0.83	0.72	-0.12	0.44	0.18	0.48	0.18	-0.30	0.24	0.57
D	0.70	0.65	0.51	1.18	0.10	0.44	0.48	0.57	0.18	0.51	0.90	0.78	0.10	0.40	0.35	0.74	0.30	0.24	0.54	0.68
C	0.30	-0.30	0.00	0.00	0.68	0.10	0.10	0.30	-0.30	0.40	0.51	0.35	0.18	-0.30	0.00	0.10	-0.30	-0.60	-0.12	0.30
E	0.74	0.54	0.40	0.40	0.10	1.31	0.40	0.57	0.40	0.86	0.93	0.85	0.57	0.70	0.51	0.65	0.48	0.00	0.60	0.65
Q	0.48	0.35	0.40	0.48	0.10	0.35	1.04	0.35	0.00	0.60	0.85	0.54	-0.30	0.40	0.10	0.18	0.00	-0.60	0.44	0.68
G	0.54	0.57	0.48	0.65	0.30	0.65	0.40	1.15	0.00	0.78	0.74	0.48	0.44	0.40	0.24	0.51	0.30	0.00	0.40	0.57
H	0.35	0.48	0.48	0.18	-0.12	0.40	0.10	0.00	0.86	0.48	0.44	0.54	0.30	0.30	0.18	0.44	0.00	-0.30	0.00	0.00
I	0.60	0.68	0.40	0.51	0.48	0.86	0.57	0.80	0.51	1.37	0.93	0.92	0.35	0.48	0.57	0.60	0.24	0.18	0.63	0.80
L	0.85	0.76	0.83	1.00	0.70	0.97	0.83	0.80	0.40	0.90	1.53	1.06	0.80	0.78	0.40	0.95	0.35	0.40	0.68	1.05
K	0.74	0.54	0.72	0.80	0.35	0.85	0.54	0.48	0.54	0.92	1.06	1.43	0.51	0.81	0.44	0.68	0.40	0.18	0.54	0.92
M	0.51	0.35	-0.12	0.10	0.18	0.57	-0.30	0.44	0.30	0.35	0.80	0.51	0.93	0.10	0.00	0.24	-0.30	-0.60	0.24	0.70
F	0.65	0.44	0.44	0.40	0.24	0.60	0.40	0.54	0.30	0.48	0.80	0.81	0.10	1.14	0.54	0.57	0.40	0.00	0.40	0.60
P	0.48	0.35	0.18	0.40	0.00	0.51	0.10	0.24	0.18	0.57	0.40	0.44	0.00	0.54	1.15	0.44	0.24	0.00	0.35	0.40
S	0.35	0.44	0.48	0.72	0.24	0.70	0.18	0.63	0.44	0.54	0.78	0.63	0.24	0.51	0.48	1.06	0.48	0.00	0.51	0.40
T	0.60	0.35	0.24	0.80	0.48	0.51	0.00	0.35	0.10	0.24	0.24	0.30	-0.60	0.40	0.18	0.51	1.02	0.00	0.44	0.60
W	0.24	-0.60	-0.30	0.24	-0.60	0.00	-0.60	0.18	-0.30	0.18	0.40	0.18	-0.60	0.00	0.00	-0.12	0.10	0.54	-0.30	0.40
Y	0.74	0.57	0.24	0.57	0.00	0.60	0.44	0.40	0.00	0.63	0.68	0.54	0.24	0.40	0.35	0.57	0.44	-0.30	1.16	0.78
V	0.81	0.63	0.57	0.68	0.35	0.74	0.63	0.57	0.00	0.80	1.04	0.92	0.70	0.57	0.40	0.35	0.60	0.35	0.74	1.29

In pair potential matrix negative values indicates that the score is less than expected so it is unlikely to interact at the interface, zero value indicates by random match they may be favorable. Positive values indicate that the interactions are favorable.

Table 3.3(a): Score of PAK group binding peptide using Pair-potential matrix.

RETNLDSLPLVDT	9.9
ENTLQSFRQDVDN	10.4
LQSFRQDVDNASL	11.2
LDSLPLVDTHSKR	11.3
DVSKPDLTAALRD	13.3
QDSVDFSLADAIN	13.4
NFSSLNLRETNLD	13.5
DLTAALRDVRQQY	13.5
QLTNDKARVEVER	13.6
NASLARLDLERKV	14.2
TRTNEKVELQELN	14.7
TRSVSSSSYRRMF	14.8
RSSVPGVRLLQDS	14.8
YESVAAKNLQEAE	15.4
QDTIGRLQDEIQN	15.5
PSTSRSLYASSPG	15.5
MSTRSVSSSSYRR	15.7
VDTHSKRTLLIKT	16.6
SRSYVTTSTRTYS	16.7
VETRDGQVINETS	16.7
TRSSAVRLRSSVP	16.8
QESTEYRRQVQSL	17
DLSEAANRNNDAL	17.2
TASRPSSSRSYVT	17.3
IATYRKLLEGEES	17.3
TYSLGSALRPSTS	17.5
RSSAVRLRSSVPG	17.6
VTTSTRTYSLGSA	17.6
KRTLLIKTVETRD	17.7
TRTYSLGSALRPS	17.8
KNTRTNEKVELQE	17.8
THSKRTLLIKTVE	17.9
NESLERQMREMEE	17.9
RISLPLPNFSSLN	17.9
ESTEYRRQVQSLT	18
SLTCEVDALKGTN	18
RPSSSRSYVTTST	18.3
SRSLYASSPGGVY	18.3
VSSSSYRRMFGGP	18.4
YVTTSTRTYSLGS	18.5
TSTRTYSLGSALR	18.6
SSSRSYVTTSTRT	18.7
ASSPGGVYATRSS	18.8
GKSRLGDLYEEEM	18.9
LRSSVPGVRLLQD	19.3

YASSPGGVYATRS	20
SVSSSSYRRMFGG	20
LGSALRPSTSRSL	20
INTEFKNTRTNEK	20.1
VESLQEEIAFLKK	20.4
EESRISLPLPNFS	20.6
TTSTRTYSLGSAL	20.7
RPSTSRSLYASSP	21.1
STSRSLYASSPGG	21.5
SSSSYRRMFGGPG	23.5

Table 3.3(b): Score of PAK group binding peptide using Betancourt Thirumalai matrix

VETRDGQVINETS	-3.8
GKSRLGDLYEEEM	-3.6
SSSRSYVTTSTRT	-3.4
PSSSRSYVTTSTR	-3.1
KNTRTNEKVELQE	-2.7
THSKRTLLIKTVE	-2.5
TRTNEKVELQELN	-2.4
YATRSSAVRLRSS	-2.4
ANRNNDALRQAKQ	-2.3
PSTSRSLYASSPG	-2.1
IATYRKLLEGEES	-2
EESRISLPLPNFS	-2
SRSYVTTSTRTYS	-1.9
DVSKPDLTAALRD	-1.9
RPSSSRSYVTTST	-1.4
KGTNESLERQMRE	-1.1
ASSPGGVYATRSS	-1.1
TTSTRTYSLGSAL	-0.8
STSRSLYASSPGG	-0.7
YASSPGGVYATRS	-0.6
SSSYRRMFGGPGT	-0.5
RPSTSRSLYASSP	-0.3
YKSKFADLSEAAN	-0.2
MSTRSVSSSSYRR	0.5
NFSSLNLRETNLD	0.6
TSTRTYSLGSALR	0.6
PGTASRPSSSRSY	0.6
LRSSVPGVRLLQD	0.7
TASRPSSSRSYVT	0.8
VDTHSKRTLLIKT	0.9
RETNLDSLPLVDT	1
YVTTSTRTYSLGS	1
RSSAVRLRSSVPG	1.2

SSSSYRRMFGGPG	1.2
QESTEYRRQVQSL	1.2
KRTLLIKTVETRD	1.8
ESTEYRRQVQSLT	1.8
RSSVPGVRLLQDS	2.1
LGSALRPSTSRSL	2.1
VSSSSYRRMFGGP	2.3
INTEFKNTRTNEK	2.3
NESLERQMREMEE	2.3
TYSLGSALRPSTS	2.4
SVSSSSYRRMFGG	2.4
ENTLQSFRQDVDN	2.6
YESVAAKNLQEAE	2.6
TRTYSLGSALRPS	2.7
NASLARLDLERKV	3
TRSSAVRLRSSVP	3
DFSLADAINTEFK	3.1
TRSVSSSSYRRMF	3.6
DLTAALRDVRQQY	4.3
IKTVETRDGQVIN	4.3
DLSEAANRNNDAL	4.6
LQSFRQDVDNASL	5.1
VESLQEEIAFLKK	5.8
QDSVDFSLADAIN	6.1
SRSLYASSPGGVY	6.4
LDSLPLVDTHSKR	6.5
FSSLNLRETNLDS	6.6
VQSLTCEVDALKG	8.2
RISLPLPNFSSLN	8.4
LLQDSVDFSLADA	8.6
SLTCEVDALKGTN	10.3

This is the score of PAK group binding peptide. The benchmarking of known PAK group is done by default matrix as well the pair potential matrix which was prepared by us. Further, work is to refine pair-potential matrix on the larger dataset.

CHAPTER 4

CONCLUSION

Protein kinases constitute the largest family in the eukaryotes. It modifies other proteins by adding phosphate group and 30% of all proteins may be phosphorylated by the protein kinases. In this project we analyzed the crystal structure which are in complex with protein kinases substrate which were retrieved by performing BLAST against the protein databank. Identified the interactions between the peptide and the protein kinases. The short peptides with length less than 15 were only used. Thus, protein and Substrate complex was formed using pymol. After the Substrate and protein complex was formed, all possible amino acid - amino acid contacts from the crystal structures of protein kinase- substrate peptide complexes were identified and binding preferences of all amino acids were calculated as the log ratios of observed / expected frequencies. Thus, a pair potential matrix was prepared which was specific to the protein kinase and the substrate interface. This work has been done for the first time. **There are other pair potential matrix also which are generalized to certain principles but our matrix is specific to the protein peptide interactions.** Therefore we calculated the pair potential matrix which is specific to substrate which can be further used for the benchmarking.

APPENDIX

CODE 1:

Python code to retrieve chain ID, Fasta sequences, PDB file chains separation, removal of water molecules and ligands, and shorts peptides.

```
from Bio.PDB.PDBParser import PDBParser
from Bio.PDB.Substrate import three_to_one
from Bio.PDB.Substrate import is_aa
from Bio.Alphabet import IUPAC
from Bio.Seq import Seq
from Bio.SeqRecord import SeqRecord
import sys
import MDAnalysis

with open("file.txt", 'r') as fh:
for pdbFile in fh:
pdbFile= pdbFile.rstrip("\n")
print(pdbFile)
p = PDBParser(PERMISSIVE=1)
structure = p.get_structure(pdbFile, pdbFile)
for model in structure:
i=0
for chain in model:
seq= list()
chainID= chain.get_id()
for residue in chain:

if is_aa(residue.get_resname(), standard=True):
seq.append(three_to_one(residue.get_resname()))
else:
seq.append("X")
print ">Chain_" + chainID+ "\n" + str("".join(seq))
print "ChainID"+ chainID+ " "
print i
u = MDAnalysis.Universe(pdbFile, permissive=False)
A = u.select_atoms('segid '+ chainID)
filename = chainID+ '_' + pdbFile
```

```
A.write("/root/Desktop/Project/PDB_FILES/Files/chains/"+filename)
pdb= open("/root/Desktop/Project/PDB_FILES/Files/chains/"+filename, "r")
file = open("/root/Desktop/Project/PDB_FILES/Files/atoms/Atom_"+filename, "w")
for line in pdb:
if line[:4] == 'ATOM' and line[17:20] !='HOH' and line[23:26] != '  0':
i+=1
maximum = line[23:26]
file.write(line)

file0 = open("/root/Desktop/Project/PDB_FILES/Files/atoms/Atom_"+filename, "r")
data = file0.readline()
minimum = int(data[23:26])
if minimum == 1:
minimum = 0
file0.close()
res = abs(minimum - int(maximum))
finalfi= open("/root/Desktop/Project/PDB_FILES/Files/chains/finalfile.txt", "a+")
data = str(pdbFile+ "           " + chainID+ "           " + str(i) + "           "
+ str(res) + "    " +"\n")
finalfi.write(data)i=0
```

CODE 2:

Title: Python code to generate multiple links to download PDB file.

```
with open('script.sh', 'r') as f:
    lines = f.readlines()
lines  = [ line.replace('wget  https://files.rcsb.org/download/',  '')  for  line  in
lines]
with open('script.sh', 'w') as f:
f.writelines(lines)
```

CODE 3:

Python code for the calculation of distances
s

```
from Bio.PDB import PDBParser

# create parser
parser = PDBParser()
```

```
# read structure from file
with open("peptideFilee.txt", 'r') as fh, open("Proteinfilee.txt", 'r') as
rf:
    for pdbFile, pdbFile2 in zip(fh,rf):
        pdbFile = pdbFile.rstrip("\n")
        pdbFile2 = pdbFile2.rstrip("\n")
       pdbFile = pdbFile.rstrip("\r")
        pdbFile2 = pdbFile2.rstrip("\r")
        print(pdbFile)
        print(pdbFile2)
       alpha = str(pdbFile[5:6])
       beta = str(pdbFile2[5:6])
       print(alpha)
        print(beta)
       structure = parser.get_structure('P', pdbFile)
       structure1 = parser.get_structure('P', pdbFile2)
       model = structure[0]
       model1 = structure1[0]
       chain = model[alpha]
       chain1 = model1[beta]

       # this example uses only the first residue of a single chain.
       # it is easy to extend this to multiple chains and residues.
       for residue1 in chain:
           for residue2 in chain1:
               file = open("Residue_info/Data_" + pdbFile + pdbFile2,
"a+")
               #if residue1 != residue2
               try:
                   distance = residue1['CA'] - residue2['CA']
               except KeyError:
                   continue

               if distance <= 6:
                   print(residue1, residue2, distance)
                   print ("\n")
                   res = str(residue1)
                   res1 = str(residue2)
                   file.write(res[9:13] + "  " + res[19:29]+ "  " + alpha
+ "  " + "--------->" +  "   " +  res1[9:13] + "  " + res1[19:29] + "  "+
beta + "  " + str(distance) + "\n")
```

REFERENCES

1. Benjamin E turk, ”Understanding and exploiting substrate recognition by protein kinases”: Current opinion in chemical biology ,4-10(2008).

2. Roskoski R Jr “A historical overview of protein kinases and their targeted small molecule inhibitors” :pharmacological research 100 ,1-23(2015).

3. Paulo Sérgio L**.** de Oliveira , Felipe Augusto N. Ferraz, Darlene A. Pena, Dimitrius T. Pramio, Felipe A. Morais, and Deborah Schechtman “Revisiting protein kinase–substrate interactions” Science Signalling, vol 9: 420(2016).

4. Holger Dinke, Claudia Chica, Allegra Via, Cathryn M. Gould, Lars J. Jensen, Toby J. Gibson, and Francesca Diella. “Phospho.ELM database for phosphorylation sites” : Nucleic Acid Research, D261-D267(2011).

5. Debasisa Mohanty , Narendra Kumar “MODPROPEP: a program for knowledge-based modeling of protein – peptide complexes” : Nucleic Acid Research,W549-W555(2007).