Characterization of Pathogenic mutations in Kinase domain

Safinaj Arju Ara Alam

Master of Science Thesis in Bioinformatics

Institute of Biomedical Technology (IBT)

University of Tampere, Finland

February 2013

DEDICATION

This thesis work is dedicated to my parents, Rahima Alam and Safiqul Alam

ACKNOWLEDGEMENT

This thesis work was done to fulfill the partial requirement of Master's Degree Program in Bioinformatics, Institute of Biomedical Technology, University of Tampere lead by Professor Mauno Vihinen.

I would like to thank Professor Vihinen for giving me the opportunity and facilities to work in his group. You have always been a strong motivator for me. Your moral support, encouragement and advices have helped me a lot to carry out this research work.

A very special and big gratitude goes to Adjunct Prof. Csaba Peter Ortutay whose kind discussion and suggestions made my thesis work possible to be well-wrapped. It's a real great pleasure for such a kind person as my supervisor. I am forever appreciating your intellectual supports, helpful attitude and discussions of every difficulties and guidance throughout the project. I have learnt a lot from you.

I am also grateful to Jouni Väliaho for your overall support during this thesis work. Your assistance helped me a lot to overcome every obstacle. Special thanks to lecturer Martti Tolvanen, who helped me all along the Master's program. Your valuable suggestions and recommendations have helped me to manage everything nicely during this degree study. I am also remembering and would like to thank Late Ayodeji E. Olatubosun who gave me moral support while I started my thesis work. Special thanks to Harlan barker for his overall support. Thanks to all the members of bioinformatics group for your cooperation and friendly environment.

My cordial gratitude goes to my parents and all of my family members specially my sister Shahin Afroz Orin and my brother Azizul Hakim Abir for their paramount encouragement, support and motivation which help me to move forward and finish my study, even staying far from them in abroad.

I would also like to thank all of my friends, my near and dear ones, relatives and well-wishers for their mental support through this journey.

Safinaj Arju Ara Alam

Tampere, Finland

February, 2013

MASTER'S THESIS

Place	UNIVERSITY OF TAMPERE
	Institute of Biomedical Technology (IBT)
Author	Safinaj Arju Ara Alam
Title	Characterization of pathogenic mutations in kinase domain
Pages	44 pages
Supervisors	Adjunct Professor CsabaOrtutay, PhD; Professor MaunoVihinen
Reviewers	Adjunct Professor CsabaOrtutay; Professor MaunoVihinen
Time	February 2013

Abstract

Background and aims: Protein Kinases having significant effects on signal transduction as well as most of the cellular processes, frequently causes diseases. A large number of disease-causing mutations have been recognized from several protein kinases. Therefore, KinMutBase database, a comprehensive database for diseases causing mutations in Protein Kinase domain was established previously, where all the mutation related information was positioned. Protein kinases are related with numerous life threatening diseases including cancers. Hence to analyze pathogenic mutations in kinase domain level to figure out the characterization of mutation pattern was the crucial concern of this thesis work.

Methods: KinMutBase were thoroughly updated with new mutations as well as new gene entries where mutations are happening in their kinase domain range. Gene and respective protein sequences reported in the dataset were downloaded to perform Multiple Sequence Alignment. Conserving all those sequence to rearrange the mutation positions, BioEdit, software for assembling conserved sequence was used. By this mean, every mutation had given to new positions in the sequence file. In conclusion, statistical analyses were performed with *R* to examine mutation pattern. In addition, mutation pattern and mutation rate of Protein Serine/ Threonine kinase group as well as Protein Tyrosine kinase group in comparison with background data set (Faisal I., 2012) was also calculated for individual amino acids and for different amino acid groups.

Results: From this study it can be proposed that several genes like as ACVRL1, STK11, and BTK could be easily mutated in any amino acid positions between their kinase ranges. Yet again, in kinase range some amino acid like Arginine (R) can be easily and frequently mutated. Leucine (L) has also a good chance for mutation readily. Certain position in MSA files which also have chance where mutations can happen quiet frequently has been figure out.

Conclusion: Throughout the kinase region, several proteins, some certain amino acids as well as some specific positions are identified for causing mutations in genomic level. Specially, Non polar aliphatic amino acid group has the higher tendency for causing mutations which easily leads to diseases.

Contents

ABBREVIATIONS

AA	Amino acid
BLAST	Basic Local Alignment Search Tool
DNA	Deoxyribonucleic acid
IBT	Institute of Biomedical Technology
ML	Maximum-Likelihood
MSA	Multiple Sequence Alignment
SNP	Single nucleotide polymorphism
PSK	Protein Seine/Threonine kinase
PTK	Protein Tyrosine Kinase
PM	Pathogenic Mutation

1. INTRODUCTION

Protein kinase one of the largest superfamily of eukaryotes regulates most of the cellular processes of eukaryotic cells including proliferation, gene expression, metabolism, motility, membrane transport and apoptosis by a major mechanism referred as phosphorylation. Being involvement of almost every activity of cells, key regulators of biological control and signaling cascades, protein kinases are more susceptible for dieses causing mutations (Hanks *et al.*, 2003).

Variation of nucleotide inside a gene is the change of any nucleotide (A, T, C, or G) in DNA sequence of any species. Variations appear for different factors and replication errors during DNA replication stage of meiosis cell division. Variation could be harmful, neutral and as well as sometimes it could be beneficial. When the variation is harmful and leads to diseases, then it could be mentioned as Pathogenic mutations (Petrov *et al.*, 2005; Sawyer *et al.*, 2007). Therefore, Pathogenic Mutations or Disease-causing mutations may be referred as an error or fault in the gene that causes damage of the production of a protein and hence create certain clinical symptoms (Gang *et al.*, 2010).

According to the cancer genome project, human genome encodes some 518 catalytically active protein kinases among them 180 protein kinases are responsible for diseases causing mutations and hence produce life threatening diseases. Thereafter it was required to generate and preserved a comprehensive database enthusiastic for protein kinase related mutations data, diseases, patient name, sequences, structures and so on. For this point of view, KinMutBase were established and maintained by IBT, University of Tampere (Ortutay *et al.*, 2005).

Since defects in kinases frequently cause diseases, mutation data is valuable for researchers from several fields. Most of the diseases that are associated with protein kinases are numerous cancers. Therefore, inhibiting the protein kinases, which means inhibiting of phosphorylation, can treat these diseases (Cohen P.*et al.*, 2002). Hence, protein kinase inhibitors could also be used as drugs. From those above discussion it is quite clear that the KinMutBase database is fundamental for understanding the characterization of pathogenic mutations. KinMutBase datasets describe mutations on the genomic, RNA, and protein level, and contain information about mutations such as the number of patients, number of unrelated families, and the number of patients homozygous for a mutation and so on. KinMutBase was updated whenever it needed to keep the dataset frequent and rationalized.

In conclusion, the aim of this thesis work was to update the databases, add new mutations as well as related diseases, patients and families. It was also concern to find new genes where mutations are happening in kinase region. Establishing a pattern how mutation are happening in genome level and to figure out which amino acids are frequently causes diseases causing mutations, which gene are most vulnerable for mutations was also a prime concern.

2. THEORETICAL BACKGROUNDS

2.1 Protein Kinases

Protein Kinases are enzymes that modify localization and overall function of many proteins by attaching phosphate groups to them at specific sites by a process called Phosphorylation. Protein Kinases are key elements in intracellular signaling pathways that control many physiological processes. It constitute one of the largest 'super families' of eukaryotic proteins (approximately 2% of the human genes) and hence play many significant roles in biology and disease (Hanks *et al.*, 2003). Protein kinases organize the activity of almost all cellular processes including proliferation, gene expression, metabolism, motility, membrane transport, apoptosis and particularly prominent in intracellular signaling transduction (Calvez *et al.*, 2002). Recent analyses suggest that the human genome encodes some 518 catalytically active protein kinases among them 478 belong to a single superfamily whose catalytic domains are related in sequence (Bignell *et al.*, 2006).

2.1.1 Classification of Protein Kinases

According to Hanks, the eukaryotic protein kinases make up a large superfamily of homologous proteins which are related by virtue of their kinase domain or catalytic domain consist of approximately 250-300 amino acid residues and contain 12 conserved sub domains that fold into a common catalytic core structure, as revealed by the 3-dimensional structures of several protein-Serine kinases (Hanks *et al.*, 1995). Therefore, Protein kinases fall into three broad classes characterized by kinase domain phylogeny with respect to related substrate specificities and modes of regulation (Hanks *et al.*, 1988).

There are two main subdivisions within the superfamily: the protein-Serine/Threonine kinases and the protein-Tyrosine kinases (The eukaryotic protein kinase super family: kinase (catalytic) domain structure and classification (Hanks *et al.*, 1995)).

Eukaryotic protein kinases are enzymes that belong to a very extensive family of proteins which share a conserved catalytic core common with both Serine/Threonine and Tyrosine protein kinases. There are a number of conserved regions in the catalytic domain of protein kinases. In the N-terminal extremity of the catalytic domain there is a glycine-rich stretch of residues in the vicinity of a lysine residue, which has been shown to be involved in ATP binding. In the central part of the

catalytic domain there is a conserved aspartic acid residue, which is important for the catalytic activity of the enzyme (Knighton *et al.*, 1991). Hence the classification appears like as follows:

- Serine/Threonine -protein kinases
- Tyrosine-protein kinases
- Dual specificity protein kinases (e.g. MEK - phosphorylates both Thr and Tyr on target proteins). Histidine and Arginine kinases also exist in the protein kinase family.

2.1.2 Functions of Protein Kinases

Protein phosphorylation is one of the most important signal transduction mechanisms by which intercellular signals control central intracellular processes such as ion transport, cellular proliferation, and hormone responses. Therefore protein kinases can serve following physiological functions-

- Cell cycle regulation
- Signal Transduction
- Angiogenesis
- Immune Circumvention/ Evasion
- Proliferation
- Apoptosis
- Growth regulation
- Tissue remodeling

(Steinberg *et al.*, 2004; Diaz –Moralli *et al.*, 2012:Kim S *et al.*, 2002)

2.1.3 Serine/Threonine protein kinases

A Serine/Threonine protein kinase (EC 2.7.11.1) also referred as Serine kinase or Threonine kinase is a kinase enzyme that phosphorylates the OH group of Serine or Threonine which have similar side chains. Protein Serine/Threonine kinases (PSKs) play a principal role in cellular homeostasis and signaling through their ability to phosphorylate transcription factors, cell cycle regulators, and a vast array of cytoplasmic and nuclear effectors (Krebs *et al.*, 1987). Therefore, Serine/Threonine kinase receptors play a role in the regulation of cell proliferation, programmed cell death

(apoptosis), cell differentiation, and embryonic development. PSK have been associated in numerous human cancers.

2.1.4 Tyrosine Protein Kinases

A Tyrosine kinase (EC 2.7.10.2) is an enzyme that can transfer a phosphate group from ATP to a protein in a cell. It has the purposes of operating an "on" or "off" switch in many cellular functions. Protein-Tyrosine kinases (PTKs) are important regulators of intracellular signal-transduction pathways facilitating development and multi-cellular communication in Eukaryotes (Hunter *et al.*, 2001). Their activity is normally tightly controlled and regulated. Tyrosine kinases activate numerous signaling pathways, leading to cell proliferation, differentiation, migration, and metabolic changes (Schlessinger et al., 1992). Moreover, enhanced Tyrosine kinase activity is the assurance of a significant element of cancers as well as other proliferative diseases.

2.1.5 Mutations in diseases

Being key regulators of most cellular pathways, in the meantime protein kinases are frequently associated with diseases, either as causative agents or as therapeutic intervention points. The Protein kinase family's significant function in signal transduction for all organisms makes it a very attractive target class for therapeutic interventions in many disease states such as cancer, diabetes, inflammation, and arthritis (Calvez *et al.*, 2004). Furthermore, it has been established that the activity of protein kinases are altered in several human diseases such as cancer and autoimmune disorders. Kinase-diseases associations have summarized 180 kinases which are linked with numerous diseases. According to Kinase-diseases associations over 180 of the 518 human kinases are known to be mutated or imperfectly controlled in various diseases.

2.2 Hanks Subdomains

The eukaryotic protein kinases are related by virtue of their kinase domains also known as catalytic domains, that consists of around 250-300 amino acid residues. These kinase groups contain 12 kinase sub domains which can fold into a common catalytic core structure (Hanks *et al.*, 1995).

2.2.1 The Homologous Kinase Domains

The kinase domains of eukaryotic protein kinases connect according to their catalytic activity. The kinase domains are further subdivided into 12 smaller sub domains indicated by Roman numerals. According to Hanks these twelve regions recognized being invariant or nearly invariant throughout the eukaryotic domain of life.

SL No	Subdomain Name	Subdomain Region	Name of Amino acids within this subdomain range
1	Hanks_I	397-423	I,N,P,K,D,L,T,F,E,G,Q,V,Y,W, and R
2	Hanks_II	424-436	Q,Y,D,V,A,I,K,M,E,G, and S
3	Hanks_III	437-450	M,S,E,D,F,I,A,K,V, and N
4	Hanks_IV	452-466	L,S,H,E,K,V,Q,Y,G,C, and T
5	Hanks_V	467-491	Q,R,P,I,F,T,E,Y,M,A,N,G,C,L, and H
6	Hanks_V1A	492-514	R,F,Q,T,L,E,M,C,K,D,V,A,Y, and S
7	Hanks_VIB_cat	515-532	K,Q,F,L,H,R,D,A,N,C, and V
8	Hanks_VII	533-548	G,V,K,S,D,F,L,R, and Y
9	Hanks_VIII_A_	549-568	D,E,Y,T,S,V,G,K,F,P,R, and W
10	Hanks_IX	569-599	L,M,Y,S,K,F,D,I,W,A,G,V,E, and P
11	Hanks_X	600-620	R,F,T,N,S,E,A,H,I,Q,G,L,Y, and P
12	Hanks_X1	621-646	L,A,S,E,K,V,Y,T,I,M,C,W,H,D, R,P, and F

Table 2.1: Hanks Protein kinase Sub domain region

During this thesis work Hanks twelve sub domains have been used to identify kinase specific motifs.

2.3 Amino acids

Amino acids are physiological macromolecules and building blocks of all biological proteins having a common structure. Amino acids, as the name implies, have two functional groups, an amino group ($–NH_2$) and a carboxyl group (–COOH) and a side chain that is exclusive for each amino acid. Amino acids link together via peptide bonds in a particular order as defined by genes. **C** (carbon), **H** (hydrogen), **O** (oxygen), and **N** (nitrogen) are vital elements of an amino acid. Amino acids play central roles both as building blocks of proteins and as intermediates in metabolism. The 20 amino acids that are found within proteins convey a vast array of chemical versatility. While amino acids are fundamental macromolecules of protein formation, they also play vital role in many physiological activities. Humans can produce 10 of the 20 amino acids which are referred as non-essential amino acids. However, other amino acids which are referred as essential amino acids are not synthesized automatically by cells and hence must be supplied by means of nutrition, protein containing foods, protein diets and so on (Champe *et al*., 2004; Nelson *et al*., 2005; Kyte *et al*., 1982; Ambrogelly *et al*., 2007; Meierhenrich *et al*., 2008; Erives *et al*., 2011).

2.3.1 Classification of amino acids

Functional and structural properties of amino acids vary a lot depending on the side chains. Depending upon their polarity, acidic, basic or neutral property, aromatic property or aliphatic property amino acids are classified differently. Classification of amino acids depending on their properties are given below-

- A, C, G, I, L, M, F, P, W, and V are in the group of non-polar amino acids.
- R, N, D, E, Q, H, K, S, T, and Y amino acids indicate polar properties.
- R, H and K are positively charged (basic) amino acids.
- D and E are acidic amino acids comprising negative charge.
- A, N, C, Q, G, I, L, M, F, P, S, T, W, Y, and V are neutral (uncharged) amino acids.
- F, W (Non polar), and Y (Polar) having aromatic ring in their structure.
- G, A, V, L, and I are aliphatic non-polar amino acid.

Therefore, amino acids vary a lot in properties. Some of them show mixed properties and therefore might be considered member of more than one group (Koolman *et al*., 1994; Champe *et al*., 2004;

Nelson *et al.*, 2005; Kyte *et al.*, 1982; Ambrogelly *et al.*, 2007; Meierhenrich *et al.*, 2008; Erives *et al.*, 2011).

Amino acids were divided in five groups during this thesis work in order to compare variability and mutability of different groups. These five groups were acidic (polar negatively charged), basic (polar positively charged), neutral (no charge), aliphatic (non-polar) and amino acids containing aromatic ring. Phenylalanine (F), Tyrosine (Y) and tryptophan (W) are the amino acids which contain aromatic ring in their structure therefore comprises aromatic group. Only aspartic acid (D) and glutamic acid (E) are acidic in nature.

2.4 Pathogenic Mutations

Mutations are alterations in the genetic sequence, and they are a main cause of diversity among entities. These variations occur at many different levels, and they can have widely differing consequences. Although various types of molecular changes exist, the word "mutation" typically refers to a change that affects the nucleic acids (Loewe, 2008). Therefore, pathogenic mutation or variation is the change of nucleotide (A, T, C, or G) in DNA sequence of any species. Pathogenic Mutations or Disease-causing mutations may be referred as an error or fault in the gene that causes damage of the production of a protein and hence certain clinical symptoms can occur. If mutations occur in non-germline cells, then these changes can be categorized as somatic mutations. From an evolutionary perspective, somatic mutations are uninteresting, unless they occur systematically and change some fundamental property of an individual--such as the capacity for survival. For example, cancer is a potent somatic mutation that will affect a single organism's survival (Loewe, 2008). Disease-causing mutations occur often inside the protein (buried) and at hydrogen-bonding residues (Gong *et al.*, 2010). According to Wang and Moult, several characteristics such as high conservation between species, presence in several haplotypes, heteroplasmic state and good correlation between heteroplasmic level and clinical symptoms are notable criteria for a pathogenic mutation (Wang *et al.*, 2001). Certain additional tools have also been suggested to understand pathogenic mutations, including the effects of single nucleotide polymorphisms in the coding region on protein function, protein stability, ligand binding, catalysis, allosteric regulation and post-translational modification (Wang *et al.*, 2001).

The smallest mutations are point mutations, in which only a single base pair is changed into another base pair. Another kind of mutations called nonsynonymous mutation, in which an amino acid sequence is changed. Such mutations lead to either the production of a different protein or the premature termination of a protein. As opposed to nonsynonymous mutations, synonymous mutations do not change an amino acid sequence, although they occur, by definition, only in sequences that code for amino acids. Synonymous mutations exist because many amino acids are encoded by multiple codons. Base pairs can also have diverse regulating properties if they are located in introns, intergenic regions, or even within the coding sequence of genes.

Mutations may also take the form of insertions or deletions, which are together known as indels. Indels can have a wide variety of lengths. At the short end of the spectrum, indels of one or two base pairs within coding sequences have the greatest effect, because they will inevitably cause a frameshift (only the addition of one or more three-base-pair codons will keep a protein approximately intact). At the intermediate level, indels can affect parts of a gene or whole groups of genes.

In conclusion, Mutational effects can be beneficial, harmful, or neutral, depending on their context or location. Most non-neutral mutations are deleterious. In general, the more base pairs that are affected by a mutation, the larger the effect of the mutation, and the larger the mutation's probability of being deleterious.

3. OBJECTIVES

Since defects in kinases frequently cause diseases .The key objectives of this thesis work were –

- Established the context that kinase domain is more likely to be mutated and Causing different diseases including various types of cancers.
- Characterization of pathogenic mutations in kinase domain.
- To figure out new mutations and new genes which cause mutations in kinase domain.
- Remapping of mutation positions to find out which position is more likely to be mutated.
- To establish the fact that certain genes (like as STK11, BTK, and ACVRL1) have shown the tendency to be mutated easily.
- Statistical analysis to understand which amino acid is most frequent for diseases causing within kinase range.
- Discovering abundances and relative mutability of individual amino acids and amino acid groups for pathogenicity in comparison with PSK domain group, PTK domain group and Background dataset.
- Discover the protein structure features in terms of mutations in that respective protein.

4. MATERIALS AND METHODS

4.1 Materials

4.1.1 KinMutBase mutation dataset

There were previously available KinMutBase (http://bioinf.uta.fi/KinMutBase/), a database for diseases causing mutations in kinase domain created and maintained by Institute of Biomedical Technology (IBT) of University of Tampere. This dataset describe mutations on the genomic, RNA, and protein level, and contains information on existence of mutations such as the number of patients, number of unrelated families, and the number of patients homozygous for a mutation (Ortutay *et al.*, 2005).

Previously available KinMutBase mutation dataset contains 41 genes among them, RPS6KA3 and JAK3 has two different kinase domain ranges. 17 genes are in Serine/Threonine protein kinase as well as 24 genes are in Tyrosine protein kinase.

The Previous version of KinMutBase (Ortutay *et al.*, 2005) contains 709 mutations among them 232 entries for Serine and 477 entries for Tyrosine.

4.1.2 Updated version of KinMutBase mutation dataset

During this thesis work, The KinMutBase dataset was systematically updated. The new dataset currently comprises 19 more new genes. Among them 17 are in Serine/Threonine protein kinase and rest of two in Tyrosine Kinase. The newer version of KinMutBase currently contains 710 new mutations among them 540 entries for Serine and 170 entries for Tyrosine.

However, 1419 pathogenic mutations from 60 reported protein or gene sequences which have mutations in their kinase domain both in Serine/Threonine kinase and Tyrosine kinase were selected for further analysis in this thesis work.

4.1.3 Background dataset

The neutralized data for 20 amino acids derived from 5,958 pathogenic mutations of VariBench (http://bioinf.uta.fi/VariBench/), a benchmark database for human protein variations created and

maintained by Institute of Biomedical Technology (IBT) of University of Tampere was used as background dataset during this thesis work (Nair *et al.*, 2012) which was used as Imrul's master's thesis also (Faisul I, 2012).

4.1.4 Sequences

According to the kinase domain range, selected proteins sequences for all reported Serine/Threonine kinase and their corresponding gene sequences as well as Tyrosine kinase and their gene sequences were downloaded from Ensembl database in FASTA format. Multiple sequence alignments (MSA) were made for Serine/Threonine kinase file and Tyrosine kinase file separately.

4.1.5 BioEdit- Tool for assembling conserved sequences

BioEdit, a biological sequence alignment editor was used (Hall *et al.*, 1997) to access the MSA files both of the groups and make the sequences more preserved to each other. Hence, rearranging the mutation positions were done by virtue of BioEdit.

4.1.6 Software for statistical analysis

R (http://www.r-project.org/**)** was selected here for both data management and statistical analysis. Various plots have also been generated by using R during this thesis work.

4.1.7 Software for programming

Programming was also a significant concern during this thesis work. Programming was required to cut all protein sequences according to their kinase range, find the proportion of amino acid in the sequences and it was also very helpful for file management and other computational management like as data manipulation. **Python** (www.python.org) programming software or programming language was used here for programming issues especially for picking up kinase region as well as data manipulation.

4.1.8 Software for Protein Visualization

UCSF CHIMERA (http://www.cgl.ucsf.edu/chimera/**),** an extensible molecular modeling system was used for molecular visualization of mutations in kinase domain of a selected Protein. Mutations were also labeled by using CHIMERA during this thesis work.

4.2 Methods

4.2.1 Preparation of datasets

4.2.1.1 Apprising the database

By the help of MUTbase software, KinMutBase was thoroughly updated at the beginning of this thesis work. MUTbase software (Riikonen *et al.*, 1999), which was formerly established to construct and maintain locus-specific mutation databases, was also modified to conserve data concurrently for several kinases and to accept mutations only in the kinase domain in KinMutBase dataset. During the updating procedure of the dataset a previously available Perl-cgi submission form has been used which is capable of inserting new mutation data in the database which is publicly accessible on the Internet. The data were handled by Perl scripts; therefore they are capable of interpreting the most common mutation types such as point mutations, insertions and deletions. Other cases are implicated manually.

As a result of the work performed in this thesis the most recently available KinMutBase contains these newly identified mutations (Mainly missense mutation, some nonsense mutation and very few expressive polymorphism has also been included) that occurred in Kinase domain. 19 new genes and the analogous 208 new mutations of these genes have been supplementary in the dataset. However, altogether 60 new proteins, their corresponding genes and 1419 pathogenic mutations have been selected for the further analysis.

4.2.1.2 Downloading Protein Sequences

Two groups were designed conferring to Serine/Threonine Kinase and Tyrosine kinase. Separate alignments are presented for Serine/Threonine kinases (PSKs) and Tyrosine kinases (PTKs), together with mutations. Serine/Threonine kinase contains 34 genes including RPS6KA3 has two kinase domain ranges. On contrary, Tyrosine contains 26 genes consuming JAK3 has two different

kinase ranges. The protein sequences of these two diverse groups were collected in FASTA format and stored for making multiple sequence alignments. Multiple sequence alignments (MSA) were made for each file containing the sequences of all genes of those groups together. These sequence alignments were further required for conserving as well as mapping for the reposition of all these mutations.

4.2.1.3 Conservation of the sequences

To generate conservation of the Serine/Threonine sequence file and Tyrosine sequence file, BioEdit (Hall *et al.*, 1997), a biological sequence alignment editor which basically practiced ClustalW alignment sequence was used during this thesis work. BioEdit can easily access MSA data files. It has the feature "Automated ClustalW alignment" and "Edit" options for making it more preserved. Therefore, it was very handy to keep a record how many characters have been cut off to make all those sequences conserved to each other. Using BioEdit, it was possible to specify the total number of the characters that was removed to make the sequences more conserved. From the BioEdit software two different files both for Serine/Threonine and Tyrosine were kept for further mutation mapping.

The alignments were further visualized with MultiDisp (http://bioinf.uta.fi/cgi-bin/MultiDisp.cgi) (P. Riikonen and M. Vihinen). In below Figure-4.1 shows how it looked like in MultiDisp for Serine/Threonine kinase file.

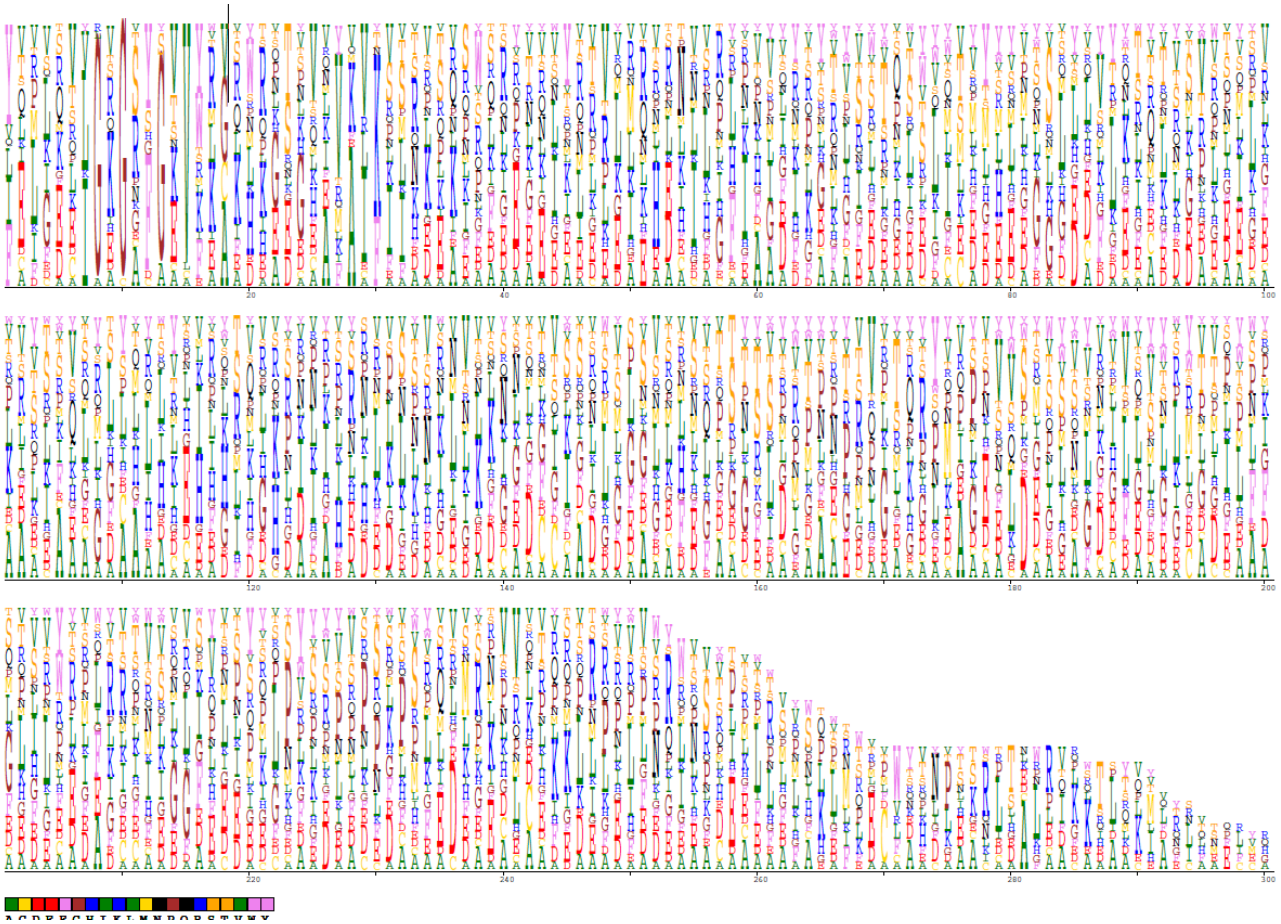

Figure 4.1 Alignments of Serine/Threonine protein kinase produced by MultiDisp

4.2.2 Mutation Mapping

The KinMutBase original data file comprises the record of every individual mutation named after dissimilar accession numbers. During this thesis work, to find out if there was any correspondence between the pattern of how mutations occur and which position has greatest frequency for a certain mutation, mutation remapping was obligatory and hence completed very carefully. For doing so, first of all for each and every accession number of KinMutBase data file, mutation position in KinMutBase was noted. Thereafter, that particular accession number lies on which group (Serine/Threonine protein kinase or Tyrosine protein kinase) was monitored. Formerly, two separated MSA file for Serine/Threonine protein kinase and Tyrosine protein kinase was viewed by using BioEdit software in accordance with that respected gene in that particular accession number. In Table 4.2 below, the picture shows how the new position of every mutation looks like in the MSA files. Therefore, every mutation was documented and had a new position in MSA file.

During conserving the MSA file some positions were cut off which have more gaps, a very few mutations (only 12 among those 1419 mutations) reclined between those cut off regions.

Table 4.2 is the sample of 15 selected mutations according to their accession number after repositioning them in MSA file.

SL NO	**Accession Number**	**Gene Name**	**Kinase**	**SwissProt ID**	**Mutation**	**Position in Kinase Range**	**Original Amino Acid**	**Change AA**	**Position in MSA files**
1	K00001	TGFBR2	Serine/Threonine	P37173	R537P	537	R	P	**357**
2	K00002				E526Q	526	E	Q	**341**
3	K00003				T458A	458	T	A	**263**
4	K00004				T315M	315	T	M	**88**
5	K00005				D405G	405	D	G	**198**
6	K00561				V250A	250	V	A	**8**
7	K00562				Y448C	448	Y	C	**253**
8	K00563				K488E	488	K	E	**302**
9	K00564				M373I	373	M	I	**155**
10	K00565				S401F	401	S	F	**194**
11	K00739				E480X	480	E	X	**290**
12	K00740				T516K	516	T	K	**331**
13	K00741				V513E	513	V	E	**328**
14	K00742				R528H	528	R	H	**343**
15	K00743				D524N	524	D	N	**339**

Table 4.2: Repositioning numbers of fifteen selected mutations among 1419 mutations were collected from MSA file.

4.2.3 Patterning Amino acids in Fasta file

During remapping of the mutation positions it was also marked which amino acids were changed and thus occurs a kinase mutation in a different file. Accordingly, anyone interested about this thesis work could perceive how mutations are trendy in Kinase domain part of a protein. Also it could be also traceable which genes have the highest chance to be mutated easily. The Figure- 4.3 indicates the fact that certain genes like as ACVRL1, BTK, STK11 were more frequently mutated.

Labeling of Mutations happening FASTA format of kinase domain of the proteins

Gene	Kinase Domain	PDB ID	PSSM ID	UniProt ID	Kinase Range	FASTA format of Kinase domain range of protein
ACVRL1	Serine/threonine -protein kinase receptor R3	3MY0	cd00180 1F3M_C	P37023	202-492	VALVECVGKGRYGEVWRGLWHGESVAVKIFSSRDEQSWFRETEIYNTVLLRHDNILGF IASDMTSRNSSTQLWLITHYHEHGSLYDFLQRQTLEPHLALRLAVSAACGLAHLHVEIF GTQGKPAIAHRDFKSRNVLVKSNLQCCIADLGLAVMHSQGSDYLDIGNNPRVGTKRY MAPEVLDEQIRTDCFESYKWTDIWAFGLVLWEIARRTIVNGIVEDYRPPFYDVVPNDPS FEDMKKVVCVDQQTPTIPNRLAADPVLSGLAQMMRECWYPNPSARLTALRIKKTLQKI
BMPR2	serine/threonine kinase	3G2F	cd00180	Q13873	203-504	LKLLELIGRGRYGAVYKGSLDERPVAVKVFSFANRQNFINEKNIYRVPLMEHDNIARFIV GDERVTADGRMEYLLVMEYYPNGSLCKYLSLHTSDWVSSCRLAHSVTRGLAYLHTEL PRGDHYKPAISHRDLNSRNVLVKNDGTCVISDFGLSMRLTGNRLVRPGEEDNAAISEVG TIRYMAPEVLEGAVNLRDCESALKQVDMYALGLIYWEIFMRCTDLFPGESVPEYQMAF QTEVGNHPTFEDMQVLVSREKQRPKFPEAWKENSLAVRSLKETIEDCWDQDAEARLTA QCAEERMAEL
BTK	Bruton tyrosine kinase	3P08	cd05113	Q06187	402-655	LTFLKELGTGQFGVVKYGKWRGQYDVAIKMIKEGSMSEDEFIEEAKVMMNLSHEKLV QLYGVCTKQRPIFIITEYMANGCLLNYLREMRHRFQTQQLLEMCKDVCEAMEYLESKQ FLHRDLAARNCLVNDQGVVKVSDFGLSRYVLDDEYTSSVGSKFPVRWSPPEVLMYSK FSSKSDIWAFGVLMWEIYSLGKMPYERFTNSETAEHIAQGLRLYRPHLASEKVYTIMY SCWHEKADERPTFKILLSNILDVM
RET	Tyrosine protein kinase	2IVT	cd05045	P07949	724 – 1016	LVLGKTLGEGEFGKVVKATAFHLKGRAGYTTVAVKMLKENASPSELRDLLSEFNVLKQ VNHPHVIKLYGACSQDGPLLLIVEYAKYGSLRGFLRESRKVGPGYLGSGGSRNSSSLDHP DERALTMGDLISFAWQISQGMQYLAEMKLVHRDLAARNILVAEGRKMKISDFGLSRDV YEEDSYVKRSQGRIPVKWMAIESLFDHIYTTQSDVWSFGVLLWEIVTLGGNPYPGIPPER LFNLLKTGHRMERPDNCSEEMYRLMLQCWKQEPDKRPVFADISKDLEKMMVKRRDYL
STK11	Serine/threonine -protein kinase	2WTK	cd00180	Q15831	49 – 309	YLMGDLLGEGSYGKVKEVLDSETLCRRAVKILKKKKLRRIPNGEANVKKEIQLLRRLR HKNVIQLVDVLYNEEKQKMYMVMEYCVCGMQEMLDSVPEKRFPVCQAHGYFCQLID GLEYLHSQGIVHKDIKPGNLLLTTGGTLKISDLGVAEALHPFAADDTCRTSQGSPAFQP PEIANGLDTFSGFKVDIWSAGVTLYNITTGLYPFEGDNIYKLFENIGKGSYAIPGDCGPP LSDLLKGMLEYEPAKRFSIRQIRQHSWF
TGFBR2	Serine/threonine -protein kinase	1PLO	cd00180	P37173	244 – 544	IELDTLVGKGRFAEVYKAKLKQNTSEQFETVAVKIFPYEEYASWKTEKDIFSDINLKHEN ILQFLTAEERKTELGKQYWLITAFHAKGNLQEYLTRHVISWEDLRKLGSSLARGIAHLHS DHTPCGRPKMPIVHRDLKSSNILVKNDLTCCLCDFGLSLRLDPTLSVDDLANSGQVGTA RYMAPEVLESRMNLENVESFKQTDVYSMALVLWEMTSRCNAVGEVKDYEPPFGSKVR EHPCVESMKDNVLRDRGRPEIPSFWLNHQGIQMVCETLTECWDHDPEARLTAQCVAE RFSELEHL
EGFR	Tyrosine protein kinase	3UG1 2EB2 2EB3	cl09925	P00533	712 – 979	FKKIKVLGSGAFGTVYKGLWIPEGEKVKIPVAIKELREATSPKANKEILDEAYVMASVDN PHVCRLLGICLTSTVQLITQLMPFGCLLDYVREHKDNIGSQYLLNWCVQIAKGMNYLED RRLVHRDLAARNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALESILH RIYTHQSDVWSYGVTVWELMTFGSKPYDGIPASEISSILEKGERLPQPPICTIDVYMIMVK CWMIDADSRPKFRELIIEFSKMARDPQRYL

Figure 4.3: Labeling of mutations happening in FASTA format of kinase domain of the proteins

4.2.4 Calculating the new position and normalization of the data

BioEdit was very beneficial while designing the new position. During the thesis work every single mutation was repositioned and gave a new positioning number with the help of BioEdit. Whereas working with the mutations data file it was required to normalize the data due to eliminating redundant data and data dependency. Data normalization was done for every amino acid in comparison with their proportion in sequence file both of PSK and PTK dataset.

4.2.5 Workflow for overall procedure

Therefore, the overall workflow expressions should be similar the followings-

Overall Working Procedure:

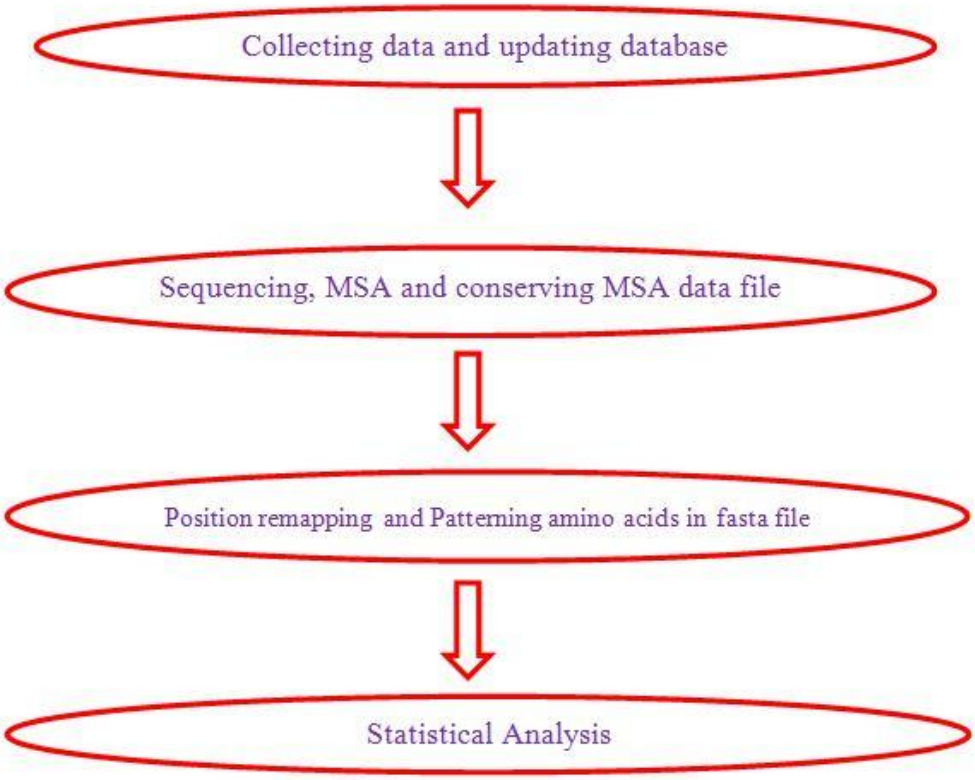

Figure 4.4: workflow of overall procedure

4.2.6 Statistical analysis

The new mutation pattern and position was analyzed statistically both for Serine/Threonine and Tyrosine file. For instance, distributions of the amino acids were plotted in graphs to visualize and compare. Several box plots were drawn to show mutation patterns. Statistical analysis was required also to compare PSK and PTK group with background dataset. These figures were made not only to compare overall distribution of two different types of kinase mutations but also to observe the differences between each amino acid of both types. More details of this aspect are verified in results and discussion parts.

In addition, statistical tests for distribution comparison of both two groups' kinase mutations were performed in comparison with also background data set. During this thesis work , 5,958 pathogenic mutations mentioned in VariBench dataset which was used for calculating **Ka/Ks ratio** was used as background dataset to compare and contrast correlation of mutation pattern (Faisal I., 2012).

5. RESULTS

5.1 New data entries

5.1.1 New Mutations both for PTK kinase and PSK kinase

The previous version of KinMutBase contained overall 709 different mutations in 25 PTK domains and in 17 PSK domains including RPS6KA3 and JAK3 has two different kinase domain ranges. Among them, 232 mutations were presented in PSK domains and 477 mutations were in PTK domains. During this thesis work the database was updated systematically. The updated version of KinMutBase currently contains 540 newly available mutations in PSK kinase domains and 170 new mutations in PTK domains which are showed in Figure-5.1. However, all-together 1419 pathogenic mutations both in PSK kinase and PTK kinase domain are currently available in the KinMutBase dataset were selected. Surprisingly newly available data shows PSK domain has more entries than that of PTK domain.

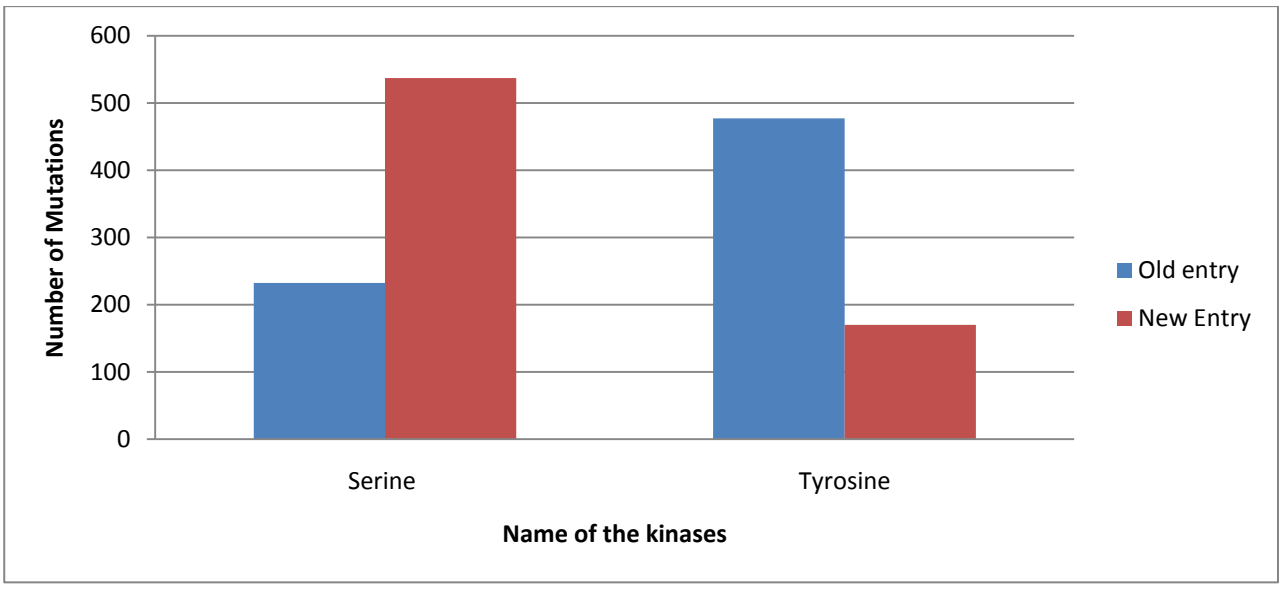

Figure 5.1: Comparison of PSK and PTK domain according to their previous entry and new entry.

5.1.2 New Genes both for PTK kinase and PSK kinase

The new dataset contains 19 new Genes. Among them 17 are in Serine/Threonine protein kinase and the remaining two in Tyrosine kinase. Consequently all-together 60 new gene entries and their corresponding 208 new mutations were available in the dataset including JAK3 and RPS6KA3 has two kinase domains. 174 entries were listed for new PSK kinase and 34 entries were for PTK kinase. Among all these new 19 genes, the PSK kinase domain region of gene PINK1 has the highest numbers of mutations (60). On the contrary in the PTK kinase domain region of gene EGFR has the highest numbers of mutations (33) which shows in Figure –5.2.

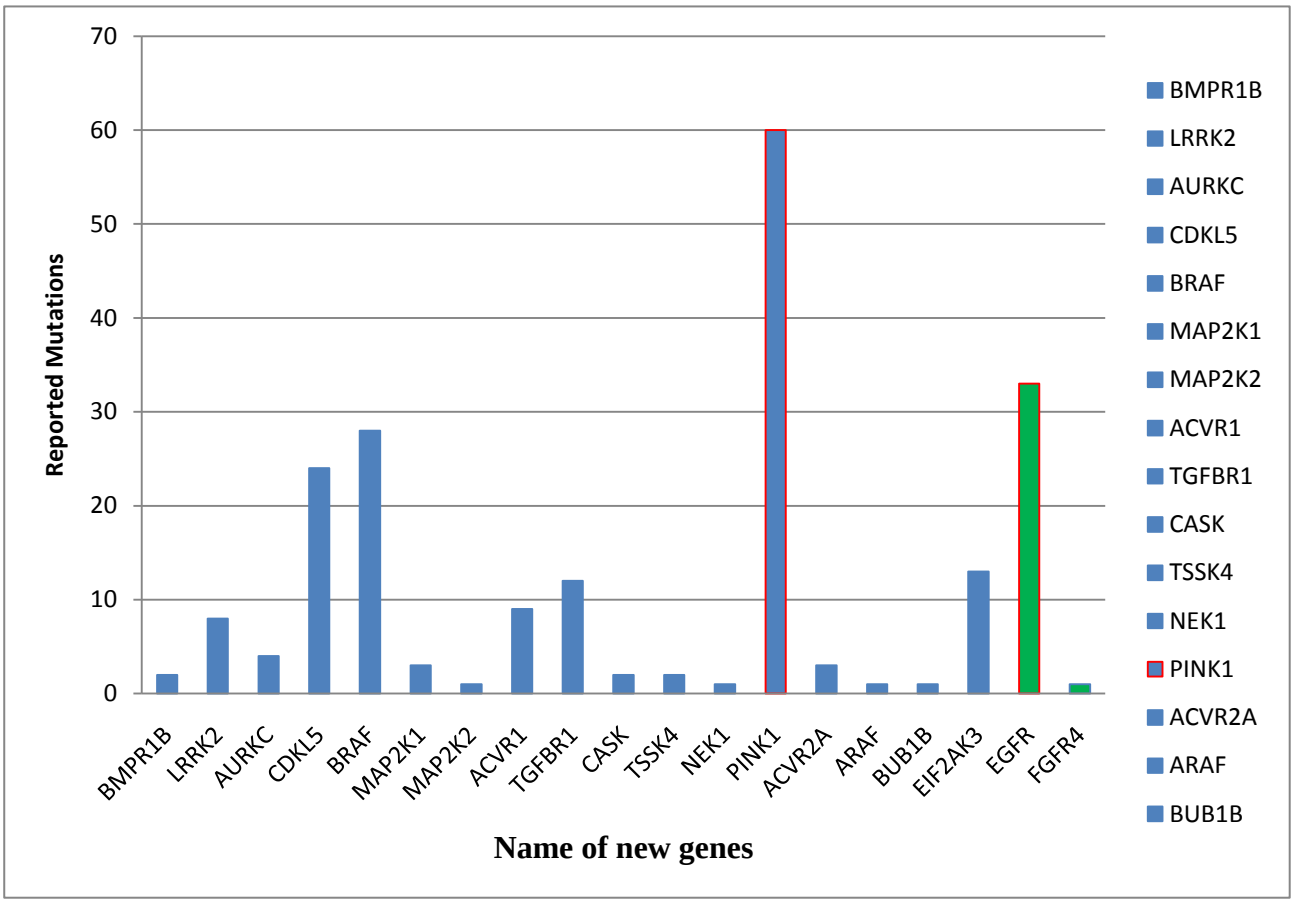

Figure 5.2: Reported new mutations as well as new genes in KinMutBase. Genes as well as reported mutations in Tyrosine kinases shows in green color and rest of the genes and their corresponding mutations are in blue. The red box mark around the bar shows the highest frequency of mutation in both PSK group and PTK group.

5.1.3 Distribution of Mutation in PSK kinase and PTK kinase

Distribution of KinMutBase mutations according to multiple sequence alignment after remapping the position performed by BioEdit software in Serine/Threonine kinases (Figure-5.3) and Tyrosine kinase (Figure-5.4) are shown in following images. The number of missense mutations is displayed in the bar charts for both of those two records.

While performing the repositioning of all those mutations in MSA file, it was noticed that position 343 in MSA file has the highest frequencies for PSK kinase. Total 14 mutations accommodate in position 343. In addition, 11 mutations were countable for position 353 in PSK domain.

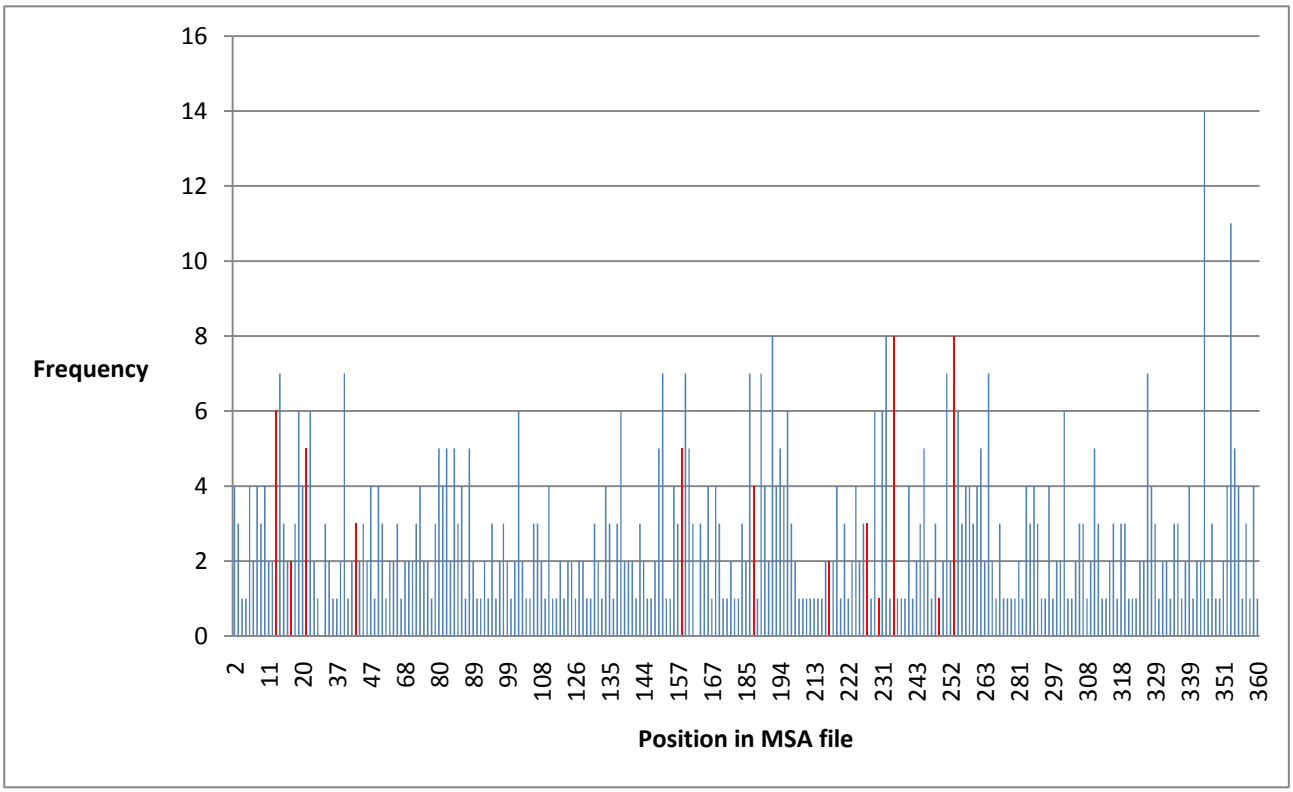

Figure5.3: Distribution of KinMutBase Mutations according to MSA in Serine/Threonine kinases. Locations of kinase-specific motifs and sub domains are presented as red bars.

For PTK kinase domain position 163 and 185 in MSA file has the highest frequencies. Overall 13 mutations had been held for position 163 and 185. Furthermore, 11 mutations are observed for position 186,189 and 198. Likewise, 12 mutations are noticed for position 158.

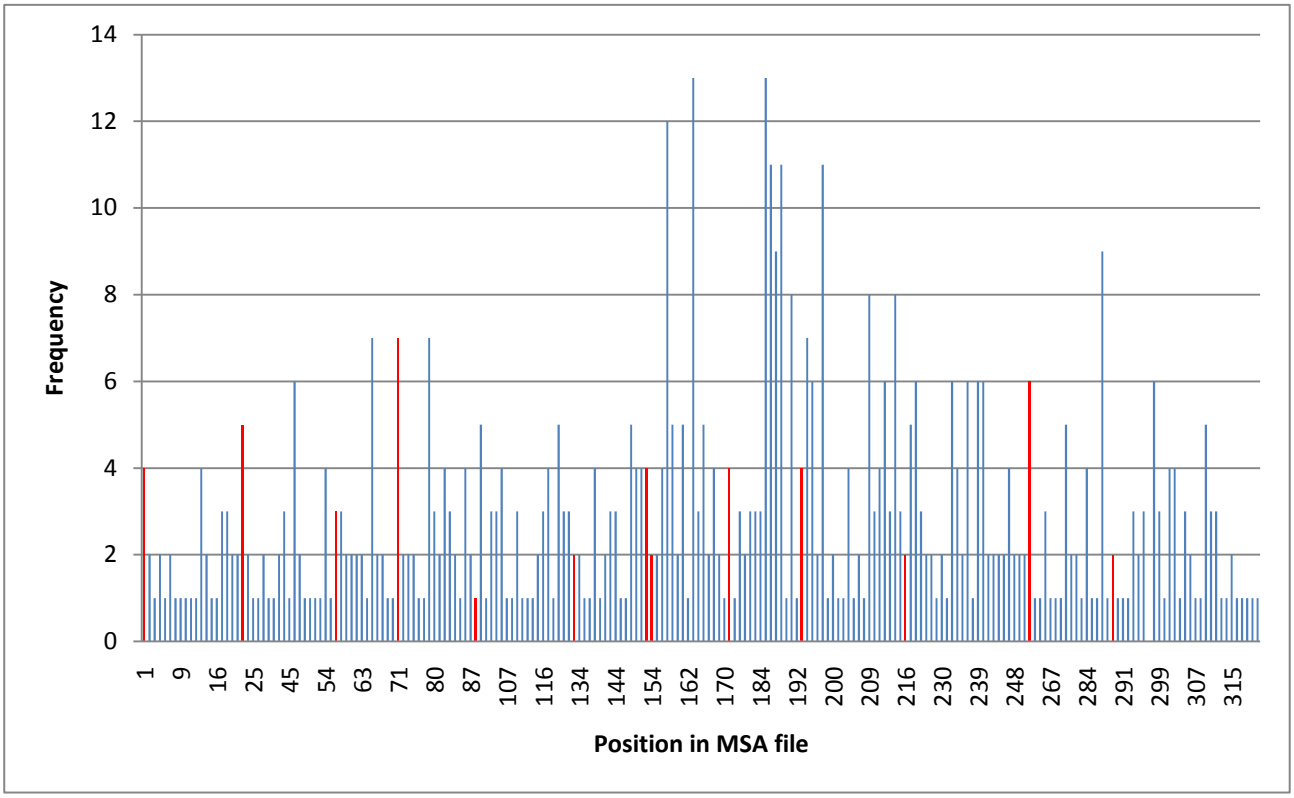

Figure5.4: Distribution of KinMutBase Mutations according to MSA in Tyrosine kinases. Locations of kinase-specific motifs and sub domains are presented as red bars.

The distribution of mutations within the domains is visualized. A large proportion of the mutations affect the conserved kinase motifs of Hanks (Hanks *et al.*, 2003). Many of the mutations are in sub domains VIB and VIII, which are responsible for substrate recognition (Johnson *et al.*, 1998; Taylor *et al.*, 1995).

It was also difficult to determine which amino acid has the greatest tendency to be mutated. From the 1419 overall mutations it appeared that Serine kinase is more likely to be mutated than Tyrosine kinase. Of the 17 recently added new genes and 537 newly available mutations in PSK kinase domains also provided support of this idea. In below Figure-5.5 demonstrates this concept.

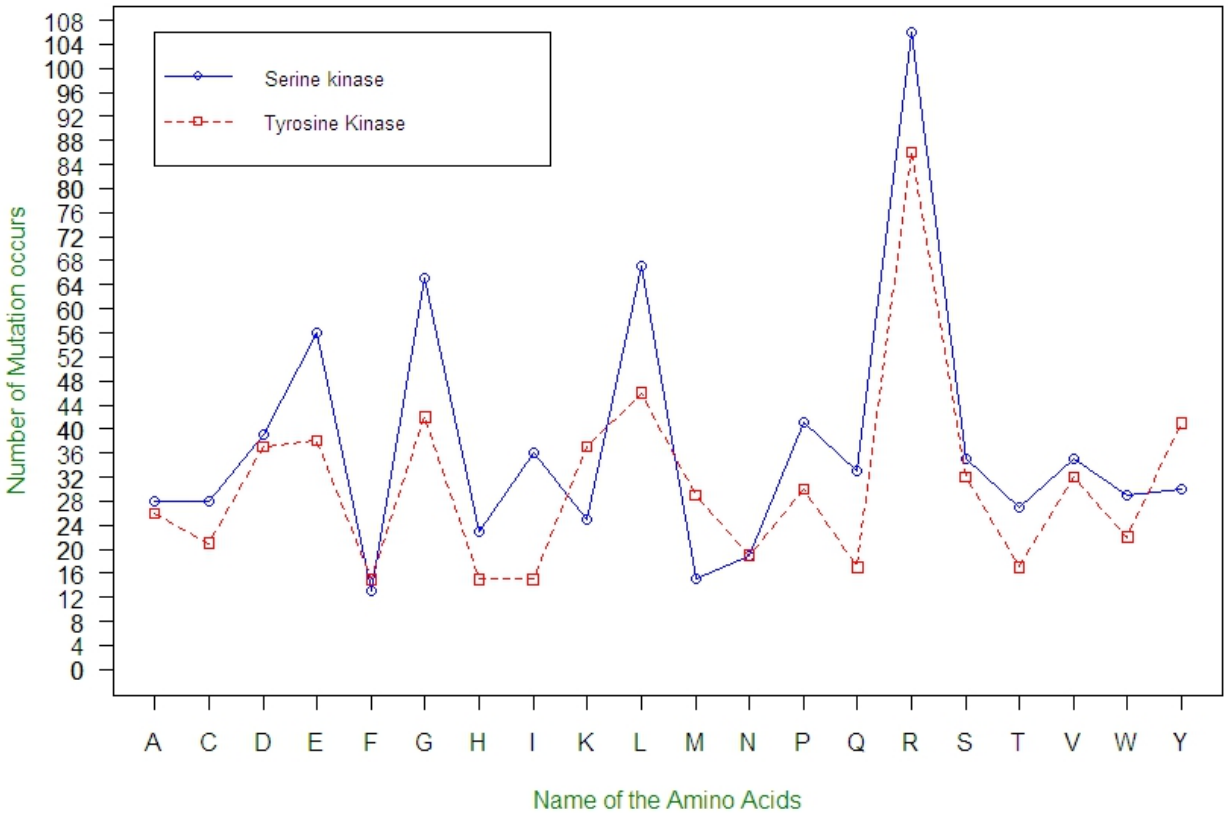

Figure5.5: Comparison of original Amino acid changes for both Serine and Tyrosine kinases

5.2 Mutability Comparison

5.2.1 Normalization of data

In the mutation data file, arginine (R) had the highest frequency of triggering mutation for both PSK domain and PTK domain. Hence, amino acid frequencies were normalized to minimize redundancy and dependency in relationship with sequence data file. Figure-5.6 shows data for PSK domain, R has the highest rate even though R was not that frequently in sequence data file for Serine. Leucine (L) has peak proportion in overall sequence file for Serine.

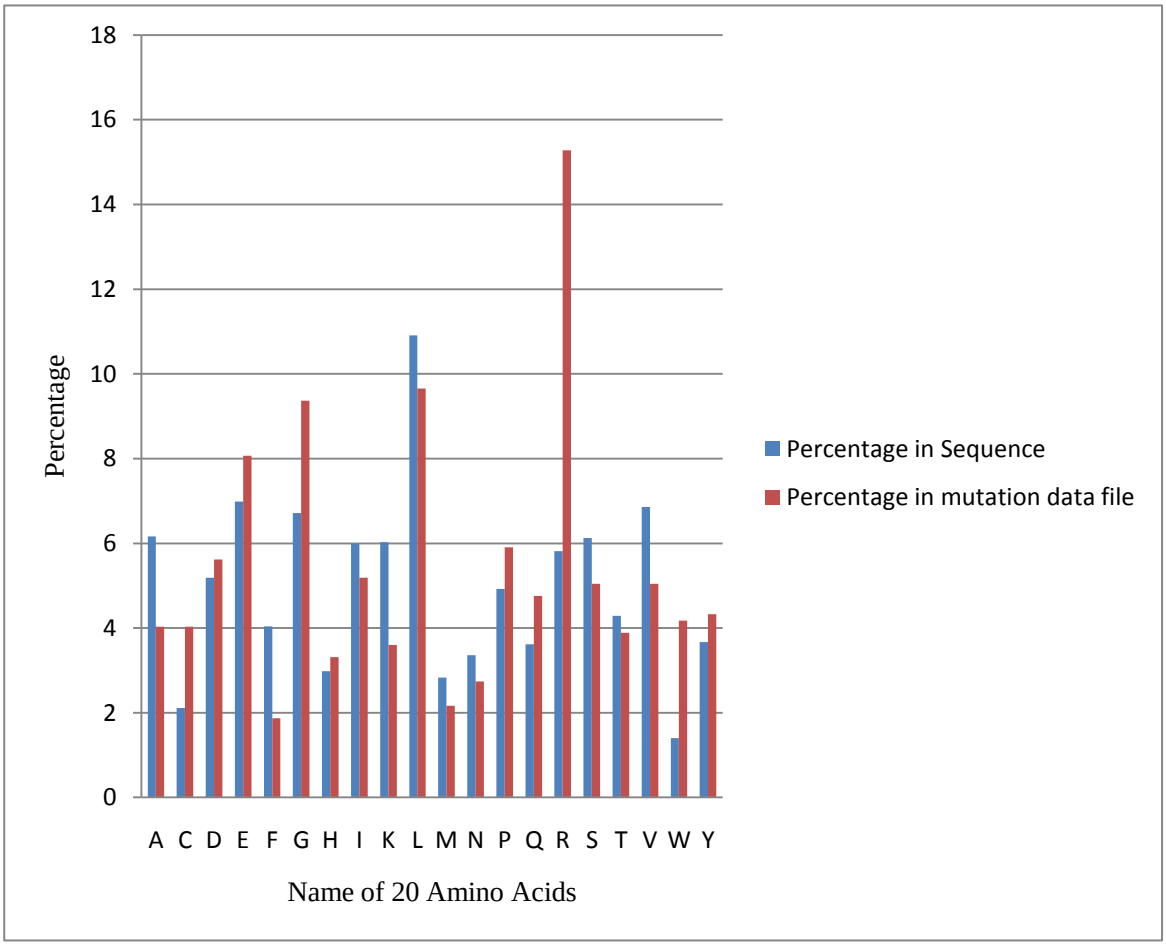

Figure 5.6: Comparison of Percentage of Serine in sequence data files in accordance with mutation data file

Similarly, in Figure-5.7 for PTK domain, R has the highest rate even though R was not frequently in sequence data file for Serine. Leucine shows a peak proportion in overall sequence file for Tyrosine also.

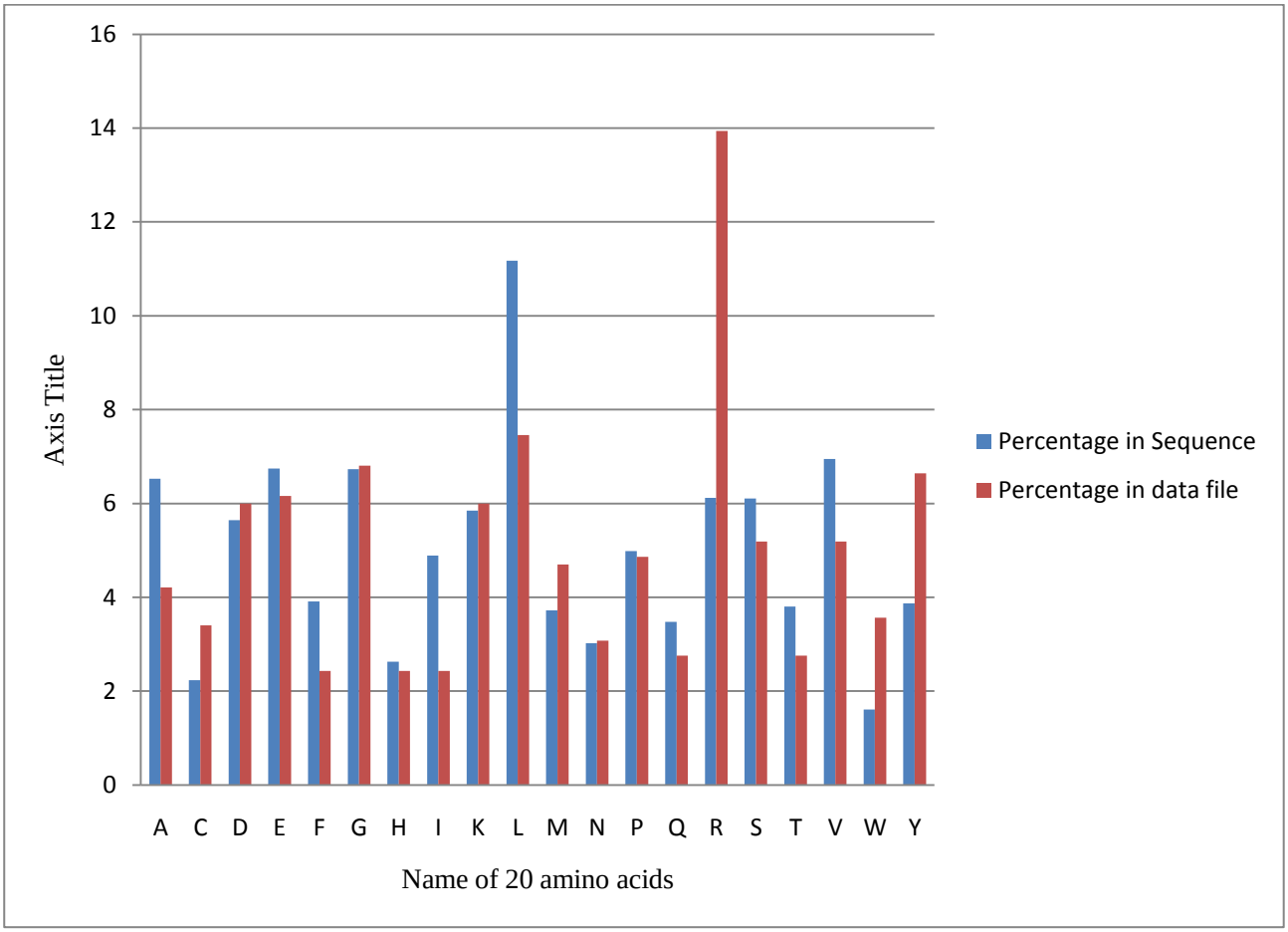

Figure 5.7: Comparison of Percentage of Tyrosine in sequence data files in accordance with mutation data file

5.2.2 Comparison of PSK and PTK normalized dataset

Normalization of both the PSK and PTK datasets was done during this thesis work. Figure-5.8 indicates how it looks like after normalizing both PSK and PTK datasets. It can be perceived that after normalizing both datasets frequencies of tryptophan are greater than that of other amino acids. Arginine also follows the same trend for both cases. Cysteine also shows the increasing affinity. On the contrary alanine (A) and phenylalanine (F) shows opposite trend.

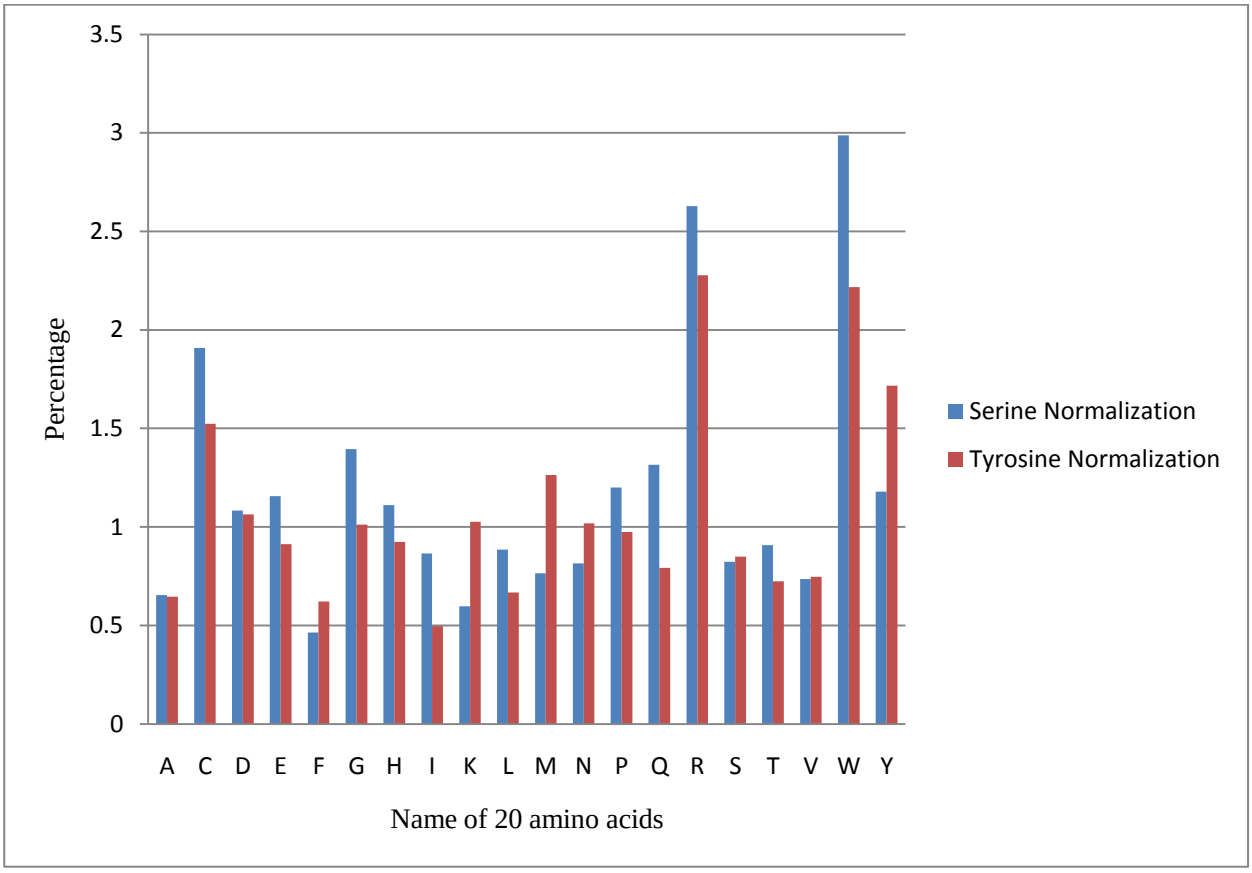

Figure 5.8: Normalized data for Serine/Threonine kinase and Tyrosine kinases which shows mutation patterns.

5.2.3 Evaluation of PSK domain and background dataset:

Mutability or Pathogenicity of individual amino acids was calculated in order to examine their mutation rate. Some amino acids have revealed a higher trend of conversion to pathogenic mutations while others displayed lower tendency to be mutated.

Thus, the evolution of the PSK domain and background dataset shown in Figure-5.9 where amino acids associated with mutation percentage is shown. In comparison with the background dataset G, L and R. have scored highest in pathogenic property whereas W, N, M etc. have shown the lowest mutation percentage. P, S, V, D etc. can be considered as almost medium in this contest.

It sharpens the contrast when you take into consideration that even though R has lowest frequency in percentage data file, R has the highest value for both PSK and background dataset. Glycine also

follows the same pattern whereas leucine has more or less similar frequency in its percentage data file as well as mutation data both for PSK and background dataset.

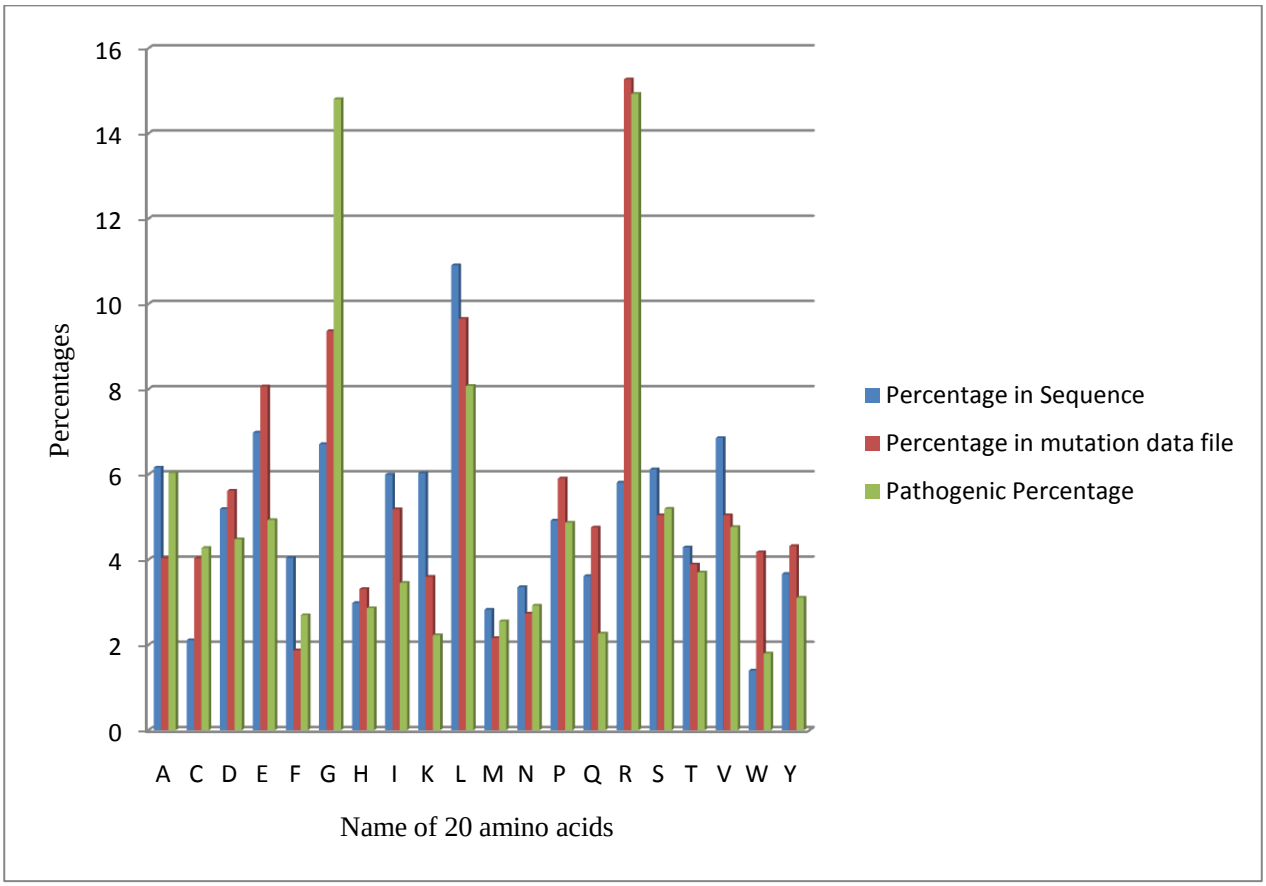

Figure 5.9: Percentage of amino acids in both mutation data and sequence data in Serine kinase comparison with background pathogenic dataset

5.2.4 Evaluation of PTK domain and background dataset:

Thereafter, the evolution of PTK domain and background dataset is shown in Figure 5.10 where amino acids associated with mutation percentage are revealed. In comparison with background dataset G, L and R have scored highest in pathogenic property even though R has smaller frequency in data file. Again W, N, M, H etc. have shown lowest mutation percentage whereas P, S, V, D etc. occupy a middle ground. Hence both PSK and PTK dataset the mutation frequency followed analogous pattern for every amino acid.

Additionally even though R has lowest frequency in percentage data file, R and Y has the highest value for both dataset. Glycine also follows the same pattern whereas L has more or less similar frequency in their percentage data file as well as mutation data both for PTK and background dataset. It is also noticeable that G has higher frequencies in background dataset than that of PSK and PTK.

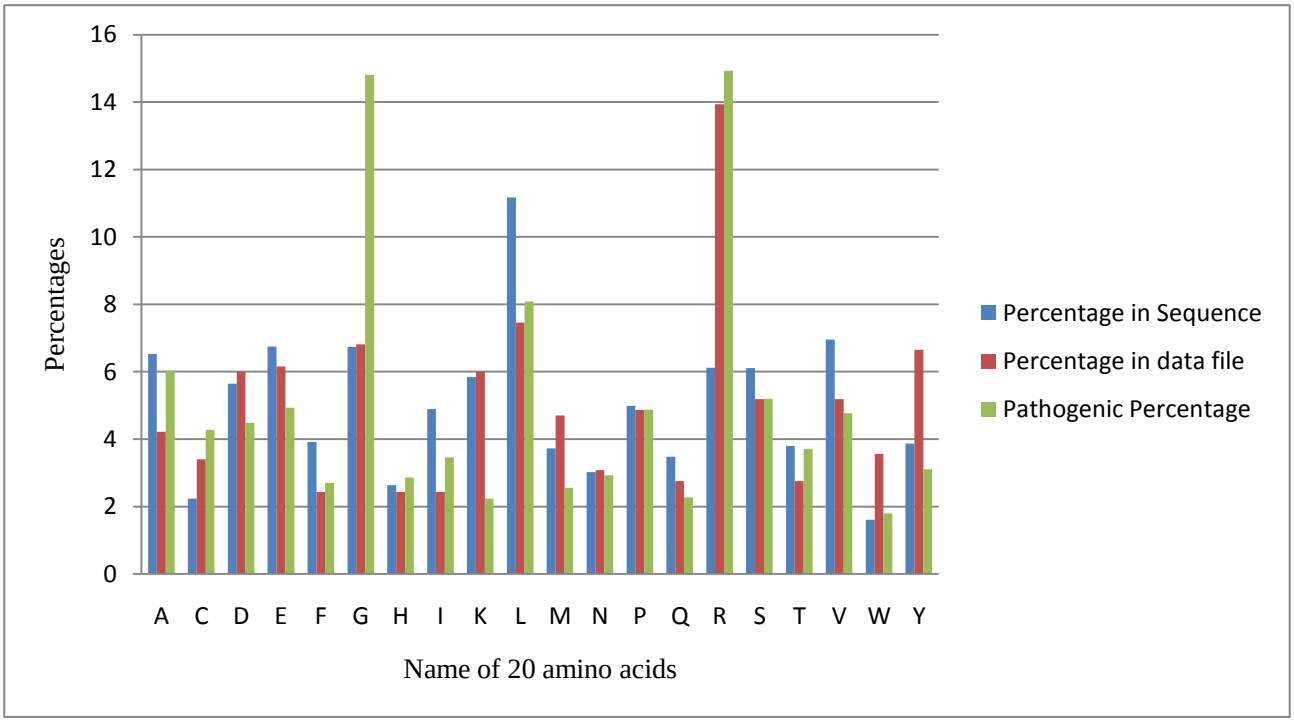

Figure 5.10: Percentage of amino acids in both mutation data and sequence data in Tyrosine kinase comparison with background pathogenic dataset

5.2.5 Evaluation of normalized data for group PSK, PTK and background pathogenic dataset

After comparison of normalized data for Serine, Tyrosine and background dataset, it was noticeable that in all these three cases R has the highest frequency of mutation. Again tryptophan (W) also follows the same trend for Serine and Tyrosine but in background dataset it shows moderate level of mutation. Cysteine (C), tyrosine (Y) and glycine (G) have also higher tendency to be mutated. On contrary, some amino acids like F, I and A trend in the opposite direction which shows in Figure-5.11.

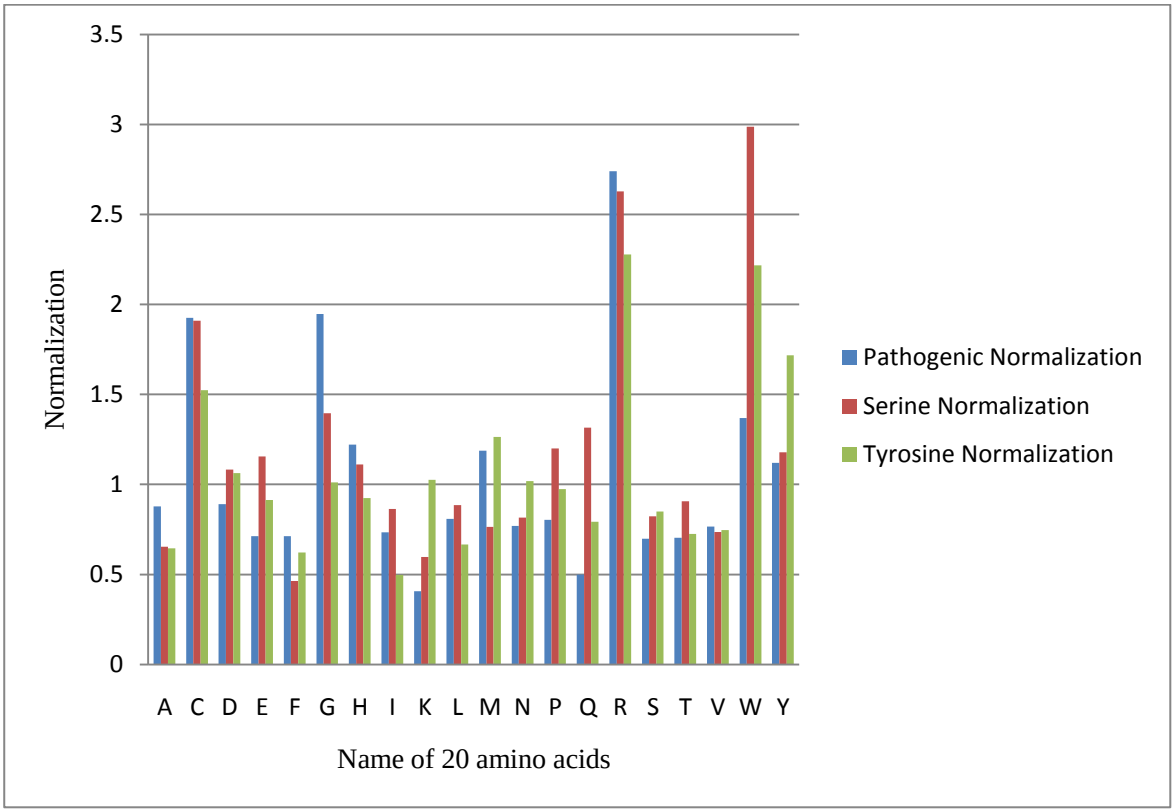

Figure 5.11: Comparison of normalized data for Serine, Tyrosine and background dataset

5.2.6 Group wise evaluation of amino acid mutability between PTK, PSK and background data set

Upon reviewing individual amino acid mutations, it was also a major concern in this work to observe group wise comparison of amino acid mutability rate in comparison with background dataset. The central purpose of this evaluation was to see mutability characteristic for different groups of amino acids.

Group wise comparisons of amino acid mutations for the respected three groups (PSK, PTK and background dataset) are shown in Figure-5.12 with their corresponding proportion in protein sequences.

Therefore, the mutability was compared according to normalization with original sequence proportions for PSK, PTK and the background dataset. Group wise amino acid in comparison with their respective quantities into protein sequences for all three mentioned groups are shown in Figure-5.12.

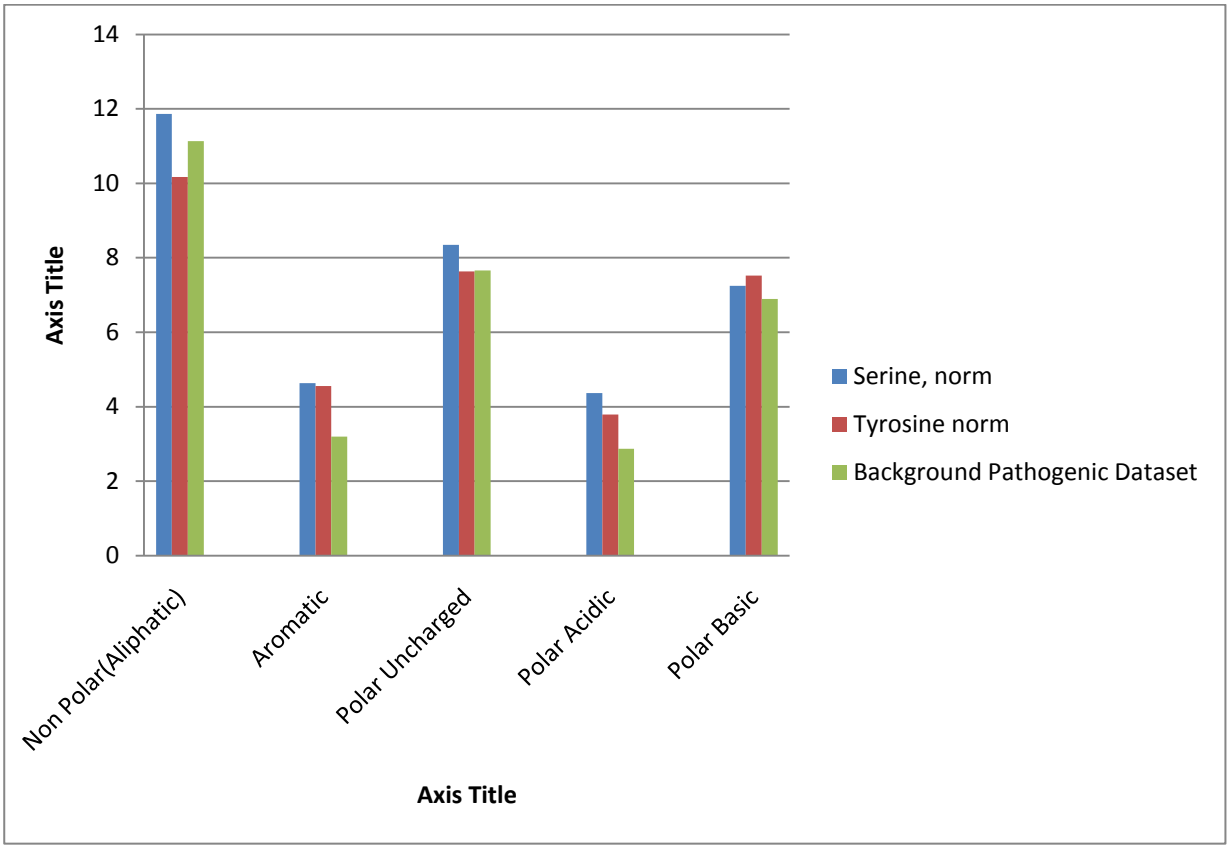

Figure 5.12: Group wise comparison of amino acids for PSK, PTK and background dataset

5.2.7 Co-relation coefficient:

To measure the co-relation between these three groups (PSK, PTK and background dataset) statistically co relation was measured and the result found it was statistically meaningful. All three datasets were positively correlated with each other, which indicate that their mutation pattern follows similar trends.

Correlation between Background dataset mutations and Serine/ Threonine dataset was: 0.73

As for example the result originated was:

```
Spearman's rank correlation rho
data:  a and b
S = 356.2676, p-value = 0.0002427
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.7321296
```

The correlation between background dataset mutations and Tyrosine dataset was: 0.71386

Similarly the correlation between Serine/Threonine data set and Tyrosine dataset was: 0.852997

The correlation between Serine/Threonine and Tyrosine kinase was quite good correlation. Therefore it could be appealed that the mutation pattern of both PSK and PTK is corresponding with most of the circumstances.

5.3 Visualization of mutations in Protein

The kinase domain structure of activin A receptor type II (PDB ID: 3MY0) was used to visualize the localization of mutations in PSKs (Figure - 5.13). For the disease-related PTKs, no structure has been determined. Location of frequently mutated residues in PSKs indicated in the ACVRL1 kinase domain structure (PDB code: 3MY0). The figure was created with **UCSF CHIMERA (**http://www.cgl.ucsf.edu/chimera/**),** an extensible molecular modeling system for molecular visualization. Mutations were labeled with red color. The sequence and structural studies as well as

this molecular visualization reveal that disease-causing mutations are widely distributed within the domain, indicating that kinase is vulnerable for alterations in many locations.

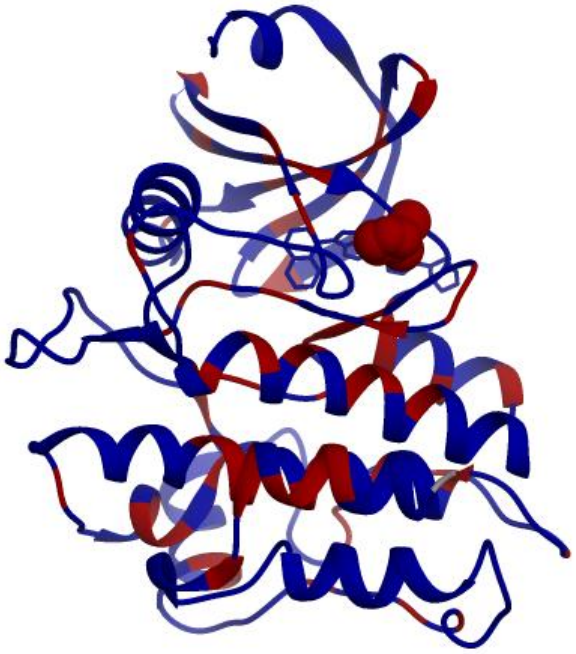

Figure 5.13: Molecular Visualization of ACVRL1 chain-A created by CHIMERA. Disease-causing mutations are widely distributed within the domain are marked as red color here.

6. DISCUSSION

Protein kinases are more likely to be mutated and thus concerning some life threatening diseases including various cancers. Since defects in kinases frequently causes diseases, mutation data is valuable for researchers from several fields. KinMutBase database was developed to integrate all the mutations together and related information in kinase domains.

The new version of KinMutBase currently contains 1419 new mutations. From the overall 1419 mutations, it appeared that Serine kinase is more likely to be mutated than Tyrosine kinase. The recently added 17 new genes and 537 newly available mutations in PSK kinase domains also provide support for this conclusion. Similarly Figure-5.2 and 5.5 demonstrate this concept respectively.

6.1 New genes and mutations feature

Mutations frequently occur in both Serine kinase and Tyrosine kinase. 17 new genes in Serine kinase as well as 208 reported mutations for those 17 Serine kinase focuses the fact that Serine kinase is more vulnerable for mutations than that of Tyrosine kinase. Additionally the 537 recently added new mutations in Serine kinase also highlighted the fact.

After normalizing the mutation data both for PSK kinase and PTK kinase it was noticeable that arginine (R) had the highest frequency of mutation. In the PSK sequence file, even though L shows the peak proportion, in reality arginine was mutated more repeatedly. This observation is also precise for Tyrosine kinases.

While performing the repositioning of all those mutations in MSA file, it was noticed that position 343 in MSA file has the highest frequencies for PSK kinase. Total 14 mutations accommodate in position 343. In addition, 11 mutations were counted for position 353 in PSK domain. This should be an important fact which will focus more insight inside mutation characterization.

Meanwhile, in case of PTK kinase domain, position 163 and 185 in MSA file have the highest frequency of mutation. Overall 13 mutations had been held for position 163 and 185. Furthermore, 11 mutations are observed for positions 186, 189 and 198. Likewise, 12 mutations are noticed for position 158.

6.2 Amino acid mutability features

6.2.1 Mutability of individual amino acids for Serine/Threonine kinase

To minimize redundancy and dependency, normalization of amino acid variability in respect to their proportion in protein sequences was accomplished**.** During the comparison of individual amino acid mutability for PSK, specifies that some amino acids (E, G, L, W and R) are very likely to be mutated (Figure-5.6). These amino acids have been found to have the higher affinity to be mutated while their proportions in protein sequences are relatively less (except L). Arginine and leucine are examples of this kind of amino acids where more variations have been noticed. Glycine has also shown same phenomenon similar with arginine and it is also more likely to be converted into diseases causing mutations. On contrary, some amino acids have been observed less frequent to become pathogenic in PSK group even though their proportions are higher in original protein sequences. A, F, K and V are the examples of this category. However, some amino acids like lysine and glutamine are unlikely to be changed.

6.2.2 Mutability of individual amino acids for Tyrosine kinase

During the comparison of individual amino acid mutability for PTK, identifies that some amino acids (E, Y) are very likely to be mutated (Figure-5.7) while their proportions in protein sequences are relatively less. On contrary, some amino acids have been observed less frequent to become pathogenic in PSK group even though their proportions are higher in original protein sequences (A, V, S). Few amino acids L, G, E and D in PTK have similar percentages in protein sequences as well as to be mutated.

6.2.3 Mutability comparison of normalized amino acid data for both PSK and PTK group

From the normalized data of amino acids of both PSK and PTK group it can be observed that for both these two groups R and W have higher possibilities to be frequently mutated. Cysteine and tyrosine also follow the similar pattern. Among these twenty amino acids, A, D, S, and V have almost similar frequencies to be mutated for both two groups. From the available normalized data, Figure-5.8 perceived that except M, Y and K rest of the amino acids in PSK group has the higher

tendency to be mutated than that of PTK group. On contrary, M, Y and K are more frequent in PTK group than that of PSK group.

6.2.4 Mutability comparison of amino acid data for PSK group and background dataset

In comparison with individual amino acid with both PSK and background dataset, Figure-5.9 demonstrates that twenty amino acids have almost the same trends of mutability for both of these two groups. Therefore it can be claimed that arginine (R) has the highest frequency to be easily mutated for both background dataset as well as PSK dataset group. However, G and L also follow the similar pattern. In PSK dataset some amino acids like W, K and F has more probability to be a variant mutation than that of background dataset.

6.2.5 Mutability comparison of amino acid data for PTK group and background dataset

While taking consideration of PTK and background dataset, it was correspondingly evident that R is more likely to be a disease causing mutations which was shown in Figure-5.10. But here in this case, G has the more frequency in background dataset than that of PTK. Other amino acids except Y in PTK follow more or less the similar pattern. Y in PTK dataset has the more frequency to be mutated than that of background dataset.

6.2.6 Mutability of individual normalized amino acids of PSK and PTK group in comparison with background dataset

Variability of amino acid groups have been more clearly observed upon normalization with their respective proportions in protein sequences in Figure-5.11. Few amino acids like R, C and W have scored highest frequencies whereas S, T, V have shown almost same lower frequency to be mutated. Again Y and G can be considered as almost medium in this contest. Therefore all those Figures-5.11 illustrates that there is a very strong correlation between the mutation patterns for individual amino acids between these three groups. Background dataset and PSK are positively correlated and the correlation between background dataset and PSK dataset is 0.73 which indicates quite a good correlation between these two dataset. Similarly, the positive correlation between

background dataset and Tyrosine is 0.71386 which also refers to a good association between these two dataset. Likewise, the correlation between PSK and PTK dataset is 0.86 which actually suggested a very good assembling between these two groups. Finally it can be mentioned that, all those figures as well as correlation values suggested that mutation pattern of twenty amino acid have more or less the similar outline.

6.2.7 Mutability of different amino acid groups for PSK, PTK and background dataset

During this thesis work, it was also carefully observed the mutability group wise amino acid in comparison with background data set with both PSK and PTK group. To find out the pattern of amino acid mutability among five different amino acid groups, Figure-5.12 describes outlines of amino acid mutability in different groups. Some group has displayed higher mutation occurrence, whereas some are less abundant to be mutated compared to their magnitudes in protein sequences. For example, in contrast with polar acidic and polar basic group, positively charged basic amino acids have expressed higher variability than their sequence proportions, but polar acidic group has shown opposite characteristics in this point of view. However, pathogenicity property upon variation can be compared nicely among the groups from Figure-5.12. Non polar aliphatic group was found most susceptible for causing mutations. Amino acids belong to this group are changing more frequently into pathogenic type.

Yet again Figure-5.12 nicely shows more variability of polar basic amino acids than that of aromatic group. Uncharged (neutral) and polar basic amino acid groups have discovered almost equivalent mutability. Aromatic amino acid group and polar acidic group show least mutation rate. Nevertheless, the non-polar aliphatic group has highest frequency for overall other groups. This group has been found maximum probability to be mutated easily and therefore disease causing mutations are happening here. In this point of view, polar acidic and aromatic amino acid group was found less pathogenic than others. Pathogenicity of non-polar aliphatic group was found almost two times higher than polar acidic and aromatic group.

6.3 Future perspectives

This thesis work was intended to found new mutations in Kinase domain as well as found new genes where mutations are happening in Kinase domain range and hence established the mutation pattern. Within the protein kinase family Serine/Threonine kinases have received comparatively lesser attention, in comparison with Tyrosine kinases (Capra et al., 2006). This thesis work also highlights the idea that the increasing amount of mutations in PSK which focus PSK should have more attention for further research. From this study it can also be proposed that several genes like as ACVRL1, STK11, and BTK could be easily mutated in any amino acid positions between their kinase ranges. Yet again, in kinase range some amino acid like arginine (R) can be easily and frequently mutated. As this study has given wonderful results, it will help in further bioinformatics research for disease causing mutations happening in Kinase domain. It will definitely inspire larger scale research with other disease causing mutations to find more genes and more frequent positions where mutations are happening would be known for each protein coding site of human genome.

As all these mutations in kinase domain are related with life threatening diseases like cancer, the updated version of KinMutBase database will also help the cancer researchers for analysis. Sometimes KinMutBase is updated manually. Consequently, it could be a worthy possibility to build a system where KinMutBase would be able to accept and conserve the data automatically.

7. SUMMARY

Disease causing mutations are quite related with Kinase domain and causing life threatening diseases. Higher mutation rate as well as mutations within kinase domain range for new recorded genes in KinMutBase also established the circumstance that kinase domain is susceptible for mutations.

The overall objective of this study was to analyze missense mutation pattern for both PSK and PTK groups. Previously Serine kinase receives moderately lesser attention than Tyrosine kinase. But recently comprised higher mutation rate in Serine kinase indicates that mutations in PSK kinases have similar tendency to be happening like as Tyrosine kinase. Moreover, several genes like as ACVRL1, STK11, and BTK could be easily mutated in any amino acid positions between their kinase ranges. Therefore more specific study should be required to distinguish beside these genes which particular genes are also frequently vulnerable for diseases causing mutations. In addition, amino acid mutability studies exposed some motivating aspects for individual specific amino acids and different groups of amino acids. Arginine (R) has been found most abundant to be mutated and causes diseases, whereas M and N was opposite for both PSK and PTK group. In addition, L and G have been found to follow the trend like R for both of that group. However, F, S, T and V were least in this consideration.

Again certain positions in MSA files like 343, 353 for PSK and 158, 163, 185, 186, 189, and 198 are also have shown the chance to be mutated easily and quiet frequently. More study should be done about this fact why these positions are vulnerable for mutations and so on. Over again Non Polar aliphatic amino acids groups have been considered as most frequently to be mutated for both PSK and PTK group and this have shown two times more pathogenicity than polar acidic amino acids and aromatic groups. Polar basic amino acids and Polar uncharged have been found medium mutation pattern.

REFERENCES

Altschul S, Gish W, Miller W, Myers E, Lipman D. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3): 403–410.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. (2010). A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4): 248–449.

Blume-Jensen P, Hunter T. (2001). Oncogenic kinase signaling.*Nature*, 411(6835): 355–365.

Brábek J, Hanks SK. (2004). Assaying protein kinase activity.*Methods Mol. Biol.*, 284: 79-80

Bignell G *et al*. (2006). Sequence analysis of the protein kinase gene family in human testicular germ-cell tumors of adolescents and adults. *Genes Chromosomes Cancer*, 45(1): 42-46.

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat*, 30(8): 1237–1244.

Capra M, Nuciforo PG, S, Quarto M, Bianchi M, Nebuloni M, Boldorini R, Pallotti F, Viale G, Gishizky ML, Draetta GF, Di Fiore1 PP. (2006). Frequent Alterations in the Expression of Serine/Threonine Kinases in Human Cancers. *Cancer Res.*, 66(16): 8147-8154.

Capriotti E, Calabrese R, Casadio R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22(22): 2729–2734.

Champe PC, Harvey RA, Ferrier DR. (2004). Lippincott's Illustrated Reviews: Biochemistry. Lippincott Williams & Wilkins.

Cohen P. (2002). Protein kinases--the major drug targets of the twenty-first century?.*Nat. Rev. Drug Discov.*, 1(4): 309-315.

Cross TG, Scheel-Toellner D, Henriquez NV, Deacon E, Salmon M, Lord JM. (2000). Serine/Threonine protein kinases and apoptosis. *Exp. Cell Res.*, 256(1): 34-41.

Creighton, Thomas H. (1993). Chapter 1, Proteins: structures and molecular properties. San Francisco: W. H. Freeman.

Devlin TM. (1992). The Textbook of Biochemistry. 3rd Edition.NY: Wiley-LissInc.

DeGroot MH. (1991). Probability and Statistics.3rd ed. Reading.MA: Addison-Wesley.

Dixit A, Yi L, Gowthaman R, Torkamani A, SchorkNJ, Verkhivker GM. (2009). Sequence and structure signatures of cancer mutation hotspots in protein kinases.*PLoS One*, 4(10): e 7485.

Davies J, Shaffer Littlewood B. (1979). Elementary Biochemistry – An Introduction to the Chemistry of Living Cells.New Jersey: Prentice-Hall Inc.

Darwin C. (1859).The Origin of Species. United Kingdom: John Murray.

Diaz-Moralli S, Tarrado-Castellarnau M, Miranda A, Cascante M. (2013).Targeting Cell Cycle Regulation in Cancer Therapy.*Pharmacol.Ther.*,pii S0163-7258 (13): 00023-00025.

Edelman AM, Blumenthal DK, Krebs EG. (1987). Protein Serine/Threonine kinases. Annu. *Rev. Biochem.*, 56: 567-613.

Eric JV, Bruce TL. (2004). Positive selection on the human genome.*Hum. Mol. Genet.*, 13: 245-254.

Faisal I. (2012). Analysis of Evolutionary Pressure and Pathogenicity of Missense Variations. Thesis for the Master degree program in Bioinformatics, University of Tampere.

Flicek P *et al*. (2010). Ensembl's 10th year.*Nucleic Acids Res.*, 38: D557-D562.

Forsdyke DR. (2006). Evolutionary Bioinformatics. New York: Springer.

Gong S, Blundell TL. (2010). Structural and functional restraints on the occurrence of single Amino acid variations in human proteins. *PLoS One*, 5(2): e9186.

Hanks SK. (2003). Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. *Genome Biol.*, 4(5): 111.

Hanks SK, Quinn AM, Hunter T.(1988). The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science*, 241(4861): 42-52.

Hanks SK, Hunter T. (1995). Protein kinnases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB. J.*, 9(8): 576-596.

Hervé Le Calvez. (2004). Kinases in cancer. *AMS Biotechnology (Europe) ltd.*, 1(1).

Hunter T. (1991). Protein kinase classification.*Methods Enzymol.*, 200: 3-37.

Jukes TH, Cantor CR. (1969). Evolution of protein molecules. In Mammalian Protein Metabolism (ed. Munro, H.N.). pp. 21-123. New York, USA: Academic Press.

Koolman J, Rohm K-H. (1996). Colour Atlas of Biochemistry. Stuttgart: Thieme.

Kim S, Bakre M, Yin H, Varner JA. (2002). Inhibition of endothelial cell survival and angiogenesis by protein kinase A. *J. Clin.Invest.*, 110(7): 933-941.

Kyte J, Doolittle RF. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1): 105–132.

Knighton DR, Zheng JH, Ten Eyck LF, Ashford VA, Xuong NH, Taylor SS, Sowadski JM. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase.*Science*, 253(5018): 407-414.

Knighton DR, Zheng JH, Ten Eyck LF, Ashford VA, Xuong NH, Taylor SS, Sowadski JM. (1991). Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, 253(5018): 414-420.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007). Clustal W and Clastal X version 2.0.*Bioinformatics*, 23(21): 2947-2948.

Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, ZipurksyL SL, Darnell J. (2004). Molecular Cell Biology.5th edition. New York: WH Freeman and Company.

Loewe L. (2008). Genetic mutation. *Nature Education*, 1(1)

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 25(21): 2744–2750.

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. (2002). The protein kinase complement of the human genome.*Science*, 298(5600):1912-34.

Millward T, Cron P, Hemmings BA. (1995). Molecular cloning and characterization of a conserved nuclear Serine (threonine) protein kinase. *Proc. Natl. Acad. Sci. USA*, 92(11): 5022-5026.

Nei M, Kumar S. (2000). Molecular Evolution and Phylogenetics.USA: Oxford University Press.

Nelson DL, Cox MM. (2005). Lehninger's Principles of Biochemistry.4th edition.New York: W. H. Freeman and Company.

Ortutay C, Väliaho J, StenbergK, Vihinen M. (2005). KinMutBase: a registry of disease-causing mutations in protein kinase domains. *Hum Mutat.*, 25(5): 435-442.

R Development Core Team (2011). "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sasidharan Nair P, Vihinen M. (2013, Epub: 2012). VariBench: A Benchmark Database for Variations. *Hum Mutat.*, 34(1): 42-49.

Schlessinger J, Ullrich A. (1992). Growth factor signaling by receptor Tyrosine kinases.*Neuron*, 9(3): 383-391.

Sergei LKP, Simon DWF. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, 21: 25312533.

Steinberg SF. (2004). Distinctive activation mechanisms and functions for protein kinase Cdelta. *Biochem J.*, 384(Pt 3): 449-459.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*., 13: 2129-2141.

Thusberg J, Vihinen M. (2009). Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum Mutat.*, 30: 703–714.

Timberlake KC. (1992). Chemistry – 5th Edition. New York: Haper-Collins Publishers Inc.

Väliaho J, SmithCI, Vihinen M. (2006). BTKbase: the mutation database for X-linked agammaglobulinemia. *Hum Mutat.*, 27(12): 1209-1217.

Wang Z, Moult J. (2001). SNPs, protein structure, and disease. *Hum Mutat.*, 17(4): 263-270.