Structural Analysis of Pathogenic Variations and Their Prevalence Across Protein-Protein Interface Regions of Human BTK Protein

Master's Thesis
Olumide Ademola, OLUFUWA
Institute of Biomedical Technology
University of Tampere
June, 2013

DEDICATION

This work is dedicated foremost to GOD ALMIGHTY, my ever present help and the sole source of my wisdom; to all curious minds that intellectually maraud the hidden mysteries of mother nature for the purpose of ameliorating mankind; to humanity.

ACKNOWLEDGEMENT

This work was performed to fulfill the partial requirement of Master's Degree Programme in Bioinformatics, Institute of Biomedical Technology, University of Tampere, Finland.

Firstly, I am indeed grateful to God almighty for the endless supply of the much needed wisdom and understanding He bestowed upon me during the course of my work; to my beloved parents for their immeasurable love, encouragement and support that they have enriched me with through the past years.

Likewise, I sincerely express my profound gratitude to my supervisors- Prof. Mauno Vihinen and Ayodeji Olatunbosun who provided me with the support and motivation during the seemingly endless period of this thesis work. In addition, I would like to thank the entire bioinformatics group at Institute of Biomedical Technology (IBT), University of Tampere, for their selfless intellectual contribution to the success of this work.

I would also like to appreciate the encouragement and support from my family, friends, colleagues and well wishers that persistently nudged me to the very point of completing this thesis work, without you all, this would not have been possible at this time.

Masters Degree Programme in Bioinformatics

OLUFUWA, OLUMIDE:

Master of Science Thesis,

June 2013

Major subject: Bioinformatics

Supervisors: Prof. Mauno Vihinen, Ayodeji E. Olatunbosun

Bruton tyrosine kinase (BTK) belong to the Tec family of non receptor tyrosine kinases that is a predominant cytoplasmic protein found in cells of humans that is responsible for B-lymphocyte development, differentiation, activation and signaling. Having 659 amino acids, the BTK protein contains five active domains: Pleckstrin homology (PH), Tec homology (TH), Src homology 3 (SH3), Src homology 2 (SH2) and Tyrosine kinase domain (TK). Pathogenic genetic variations, such as single nucleotide polymorphisms (SNP), in human BTK gene leads to a deficient immune condition known as X-linked agammaglobulinemia (XLA).

Pathogenic variations leading to XLA are traceable to the defected functional sites of the BTK protein. Thus, grounding our knowledge of the structural interactions occurring at the protein-protein interface regions of the BTK protein would assist in understanding the disease mechanism as well as determining the targets for therapeutic agents.

Six protein-protein interface predictors (cons-PPISP, PredUS, SPPIDER, Meta-PPISP, PPI-Pred and PIER) were selected to analyze the interfacial regions of PH, SH2 and kinase domains of human BTK protein. In addition, dataset on known pathogenic SNPs in BTK that causes XLA were collected from BTKbase and mapped to regions of the predicted protein interface based on the prevalence of the variations. The results showed that residues V64 (PH); W281, K284, R288 (SH2); G409, T410, Q412, G414, V415, Y617 (TK) are integral constituents of interface regions in their respective domains. Statistical analysis of amino acid abundance across interface region highlighted arginine as the most abundant across the interface regions of PH, SH2 and kinase domain. In addition, arginine also had the highest count of pathogenic SNPs known to cause XLA.

ABBREVIATIONS

AA	Amino acid
BLNK	B-cell linker gene
BP-135/TFII-I	Transcription factor deleted in Williams-Beuren syndrome
BRDG1	BCR Downstream signaling 1
CD28	Cluster of Differentiation 28
cDNA	Complementary deoxyribonucleic acid
c-Cbl	Chromosomal Casitas b-lineage lymphoma
Dok1	Docking protein 1
F-actin	Type of actin found in cytoskeleton
FAK	Focal Adhesion Kinase
Fas	A member of tumor necrosis factor (TNF) receptor family
GRB10	Gene coding for growth factor receptor-bound protein 10
mRNA	Messenger ribonucleic acid
NMR	Nuclear Magnetic Resonance
nsSNP	Non-synonymous Single Nucleotide Polymorphisms
PDB	Protein Database
PH	Pleckstrin Homology
PI3K	Phosphatidylinositol 3-kinases
PIP3	Phosphatidylinositol (3,4,5)-triphospate (PtdIns(3,4,5)P3)
PKC	Protein Kinase C
PTPD1	Protein Tyrosine Phosphatase 1D
RCSB	Research Collaboratory for Structural Bioinformatics
RSA	Relative Solvent Accessibility

SH2	Src Homology 2
SH3	Src Homology 3
SNP	Single Nucleotide Polymorphism
STAT	Signal Transducer and Activator of Transcription
TK	Tyrosine Kinase
Vav	A cell signaling protein family
WASP	Wiskott-Aldrich Syndrome Protein
XLA	X-Linked Agammaglobulinemia
ZAP-70	Zeta-chain associated protein kinase 70

CONTENTS

1 INTRODUCTION

Till date, continuous effort is being channeled in several fields of biology, proteomic science, bioinformatics and other related fields to understand the fundamental relationship between the structure of a protein in relation to its functions. Earlier studies have been able to ascertain that conserved regions of protein structure and specific coverage of sequences or sequence profiles are directly linked to functional/active sites of a protein (Chen and Zhou, 2005).

Bruton tyrosine kinase (BTK) is a cytoplasmic protein found in certain cells in the human body; it is responsible for B-lymphocyte development, differentiation, activation and signaling (Yu *et al.* 2006). Although BTK is predominantly expressed in B lymphocytes, other subcellular locations of the BTK protein in humans include: cytoplasm, nucleus, plasma membrane and peripheral membrane protein; it is expressed in numerous tissues and the most abundant of these are the lymph, tonsil, blood, lymph nodes, bone marrow, spleen and connective tissues (Várnai *et al.* 1999; Nore *et al.* 2000; Mohammed *et al.* 2000; Vargas *et al* 2002, Mohamed *et al.* 2009).

The BTK protein contains five functional domains: Pleckstrin homology, Src homology 3, Src homology 2, Tyrosine kinase and Tec homology domains also written as PH, SH3, SH2, TK and TH respectively (Gustafsson *et al.* 2012). These domain regions have been studied to exhibit high evolutionary conservation on both sequence level and structural level. Genetic variations in conserved residues are mostly responsible for various disease conditions (Chen and Zhou, 2005).

Genetic variations such as Single Nucleotide Polymorphisms (SNPs) have been observed to exhibit pathogenic phenotypes (Lindvall *et al.* 2005). A studied example of a pathogenic phenotype as a result of SNP in human BTK protein produces a disease condition known as X-linked agammaglobulinemia (XLA), or Bruton agammaglobulinemia.

According to Kang *et al.* (2001), 80% of XLA cases are usually as a result of a pathogenic variation in the human BTK gene which has been located at the long arm of the X chromosome at band Xq21.3 to Xq22. An on-going effort by Vihinen *et al.* (1998) and Väliaho *et al.* (2006) entails the meticulous curation of a registry specifically for XLA related variations which are as a result of pathogenic SNPs associated with human BTK. This registry is publicly accessible as a database which can be found at: http://bioinf.uta.fi/BTKbase.

Quite a number of biological activities are controlled or performed through interactions between proteins. Functionally important residues located at the surface of the protein serve as "interfaces" crucial for protein function. The interfacial contact or the characteristic of a protein interface has been defined by Chen and Zhou (2005) as "a pair of heavy atoms from two sides of a protein surface (interface) that are within 5 Å".

The focus of this thesis work is to identify potential correlations between the results of independent research studies and web tools (on protein interface prediction) which have converging conclusions. Amongst these include: a meta-analysis of the human BTK protein surface using interface predictors to determine amino acid residues that are most likely situated in the interface region of the protein structure; identifying the secondary structures pertinent to be located at interfaces; to deduce the potential effects of known pathogenic SNPs that have been implicated in those interfaces.

2 THEORETICAL BACKGROUND

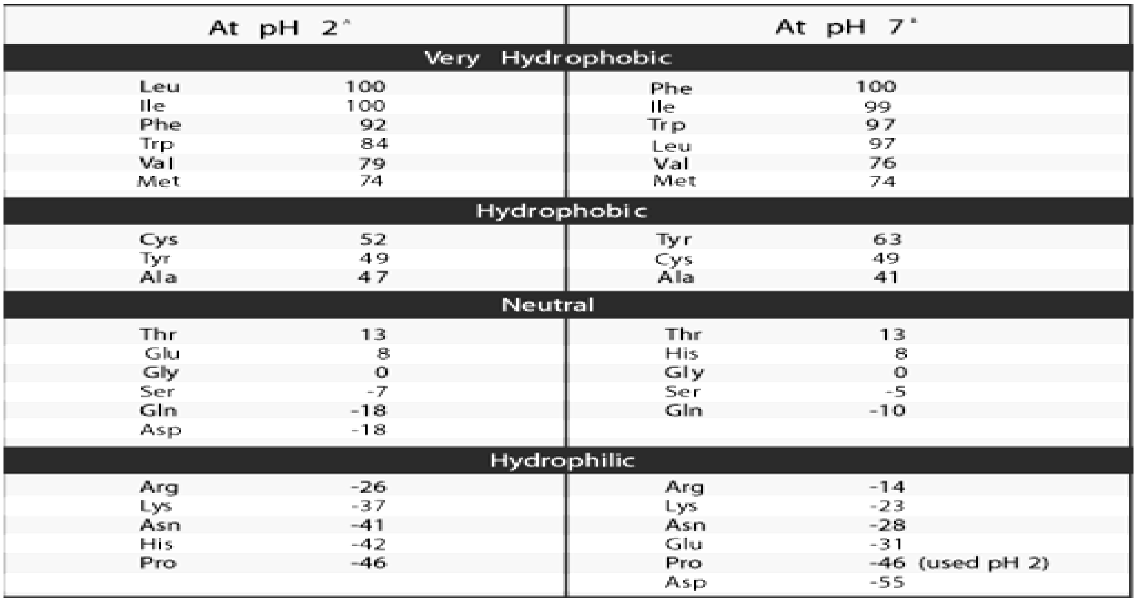
2.1 Amino acid and Proteins

Of all the biological molecules found in the human body proteins, in conjunction with other biomolecules, have been firmly accredited to be one of the most important of them all. Amino acids are sub units of proteins that ultimately form macro molecular structures through a series a biological processes. Structurally, these proteins are meticulously organized and fashioned in a step-wise manner through a series of electromagnetic interaction e.g. peptide bonds, disulphide and hydrogen and bonds.

Proteomics and other molecular biosciences are continuously inventing a vast array of novel techniques to study amino acids, which serve as building blocks of proteins. Amino acids are divided into standard and non-standard; the standard amino acids comprise of twenty abundant residues with each having unique chemical structure and property (Anfinsen *et al.* 1972).

2.1.1 Amino acid: composition and property

The standard twenty amino acids and their physiochemical properties that are of utmost interest in this particular work are shown in Figure 2.1.



At pH 2*		At pH 7*	
Very Hydrophobic			
Leu	100	Phe	100
Ile	100	Ile	99
Phe	92	Trp	97
Trp	84	Leu	97
Val	79	Val	76
Met	74	Met	74
Hydrophobic			
Cys	52	Tyr	63
Tyr	49	Cys	49
Ala	47	Ala	41
Neutral			
Thr	13	Thr	13
Glu	8	His	8
Gly	0	Gly	0
Ser	-7	Ser	-5
Gln	-18	Gln	-10
Asp	-18		
Hydrophilic			
Arg	-26	Arg	-14
Lys	-37	Lys	-23
Asn	-41	Asn	-28
His	-42	Glu	-31
Pro	-46	Pro	-46 (used pH 2)
		Asp	-55

Figure 2.1 Hydrophobic chart of the standard amino acids (Monera *et al.* 1995)

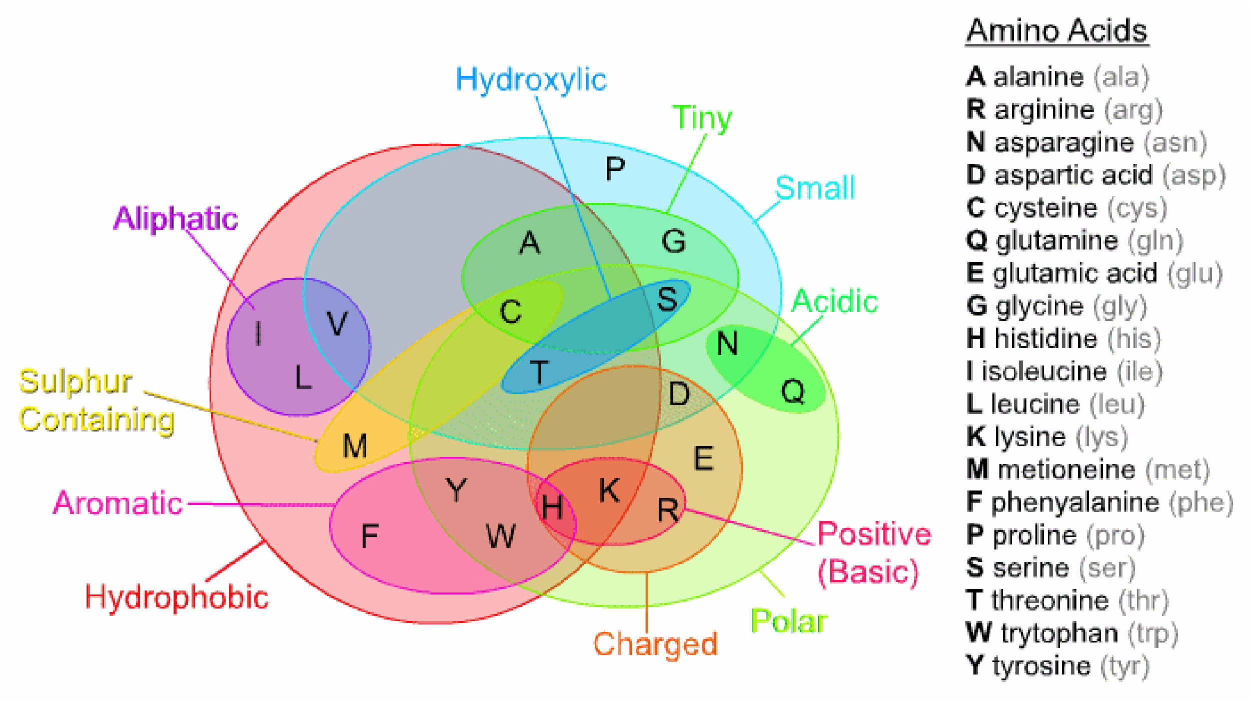


Figure 2.2 Venn diagram classifying physical and chemical properties of standard amino acids (Taylor 1986; Luu *et al.* 2012)

Table 2.1 The twenty standard amino acids with their corresponding chemical properties (Betts and Russell, 2003)

Amino Acid	3-Letter code	1-Letter code	Side-chain Polarity	Side-chain charge	Hydropathy Index
Alanine	Ala	A	Nonpolar	Neutral	1.8
Arginine	Arg	R	Polar	Positive	-4.5
Asparagine	Asn	N	Polar	Neutral	-3.5
Aspartic acid	Asp	D	Polar	Negative	-3.5
Cysteine	Cys	C	Nonpolar	Neutral	2.5
Glutamic acid	Glu	E	Polar	Negative	-3.5
Glycine	Gly	G	Nonpolar	Neutral	-0.4
Glutamine	Gln	Q	Polar	Neutral	-3.5
Histidine	His	H	Polar	Positive	-3.2
Isoleucine	Ile	I	Nonpolar	Neutral	4.5
Leucine	Leu	L	Nonpolar	Neutral	3.8
Lysine	Lys	K	Polar	Positive	-3.9

Methionine	Met	M	Nonpolar	Neutral	1.9
Phenylalanine	Phe	F	Nonpolar	Neutral	2.8
Proline	Pro	P	Nonpolar	Neutral	-1.6
Serine	Ser	S	Polar	Neutral	-0.8
Threonine	Thr	T	Polar	Neutral	-0.7
Tryptophan	Trp	W	Nonpolar	Neutral	0.9
Tyrosine	Tyr	Y	Polar	Neutral	-1.3
Valine	Val	V	Nonpolar	Neutral	4.2

2.2 Protein Structure Organization

Several levels of organization occur in proteins, and these are usually comprised of amino acid subunits conglomerated by chemical bonds. As shown in Figure 2.2, Anfinsen *et al.* (1972) also deduced that the net physiological property of proteins are a function of the constituting amino acids.

Protein structures exist in four levels of organization namely: primary, secondary, tertiary, and quaternary. Firstly, proteins are synthesized as a primary sequence (by the bonding of individual amino acid residues into a chain via peptide bonds) and then fold into secondary structures (alpha (α) helices and the beta (β) strands). Furthermore, the secondary structures fold into more complex form of polypeptides to form tertiary structures (including loops) and finally into quaternary structures- two or more polypeptides arranged in space in a specific orientation (Branden *et al.* 1997).

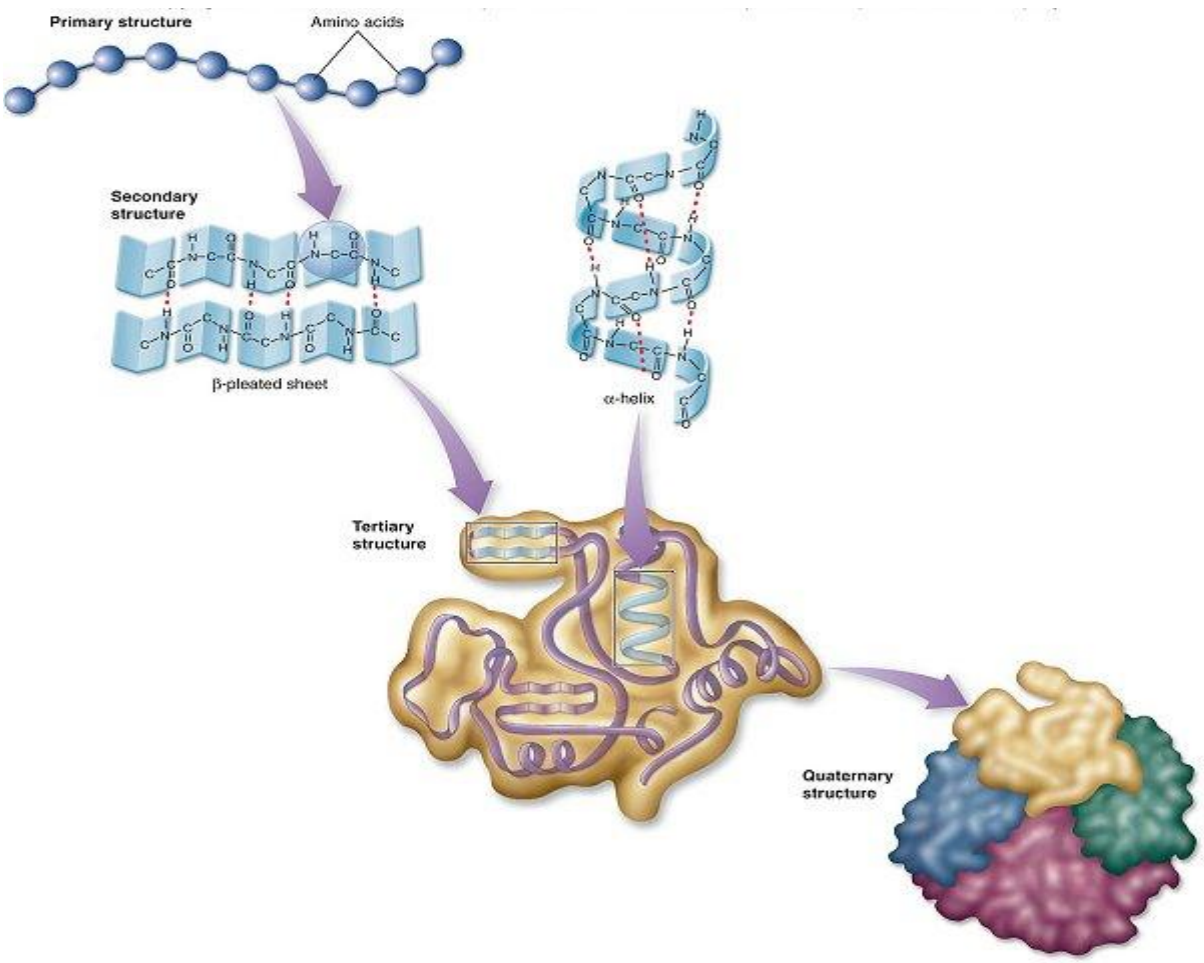


Figure 2.3 Image showing the hierarchical organization of protein structures (College of Siskiyous moodle: http://www.yellowtang.org/images/levels_of_protein_s_c_la_784.jpg, 2012)

2.2.1 Factors Influencing Structures of Proteins

Brooker *et al.* (2008) identified some factors that are pertinent to the structure of proteins. These factors consequently affect the folding of the protein structure as well as the stability: both have been shown to be major consequences of disease-causing amino acid substitution variation (Bross *et al.*, 1999; Wang and Moult, 2001; Ferrer-Costa *et al.,* 2002; Yue *et al*., 2005).

Earlier studies by Thomas and John (1987) suggested that "the charge and charge distribution of proteins are likely to influence surface activity because it is known that most of the charged amino acids reside at the exterior of protein molecules". Below are factors influencing protein structures according to Brooker *et al.* (2008).

2.2.1.1 Hydrogen bond

A high percentage of hydrogen bonds exist in protein: hydrogen bonding is the attraction of a hydrogen atom to an electronegative atom. Examples of electronegative atoms found in proteins include: nitrogen (N) and oxygen (O). Hydrogen bonds can occur in intramolecular or intermolecular forms. A collection of these bonds form substantial forces within the protein molecule.

2.2.1.2 Disulfide bridges

Disulfide bridges are formed when sulfhydryl groups from different amino acids bind to form a connection between the two amino acids. A sulfhydryl group consists of a sulfur atom combined with a hydrogen atom (–SH). According to Khan and Vihinen (2010), disulphide bonds contribute to the protein stability- they are usually formed by the bonding activity between sulfhydryl groups in cysteine molecules.

2.2.1.3 Polar interactions

Polar interactions are the attraction of negative side chains to positive side chains. Ionic bonding is a type of polar interaction. Ions are attracted to other ions of the opposite charge so the binding of the ions can result in a net charge of zero. Various amino acids have a net polarity (see Table 1) which is as a result of their side chains i.e. their R-group.

2.2.1.4 Van der Waals forces

Van der Waals forces are the weak attractions and repulsions of atoms. If an atom is within a certain distance of another atom, they will experience an attraction. If the atoms are too close, there will be a repulsion.

2.2.1.5 Hydrophobicity

Hydrophobic effect is the repulsion of nonpolar molecules from water, which is a polar substance. The nonpolar chains of a polypeptide move into the center of the structure away from the water and towards other nonpolar chains- this type of adaptive structures are called zwitterions. As shown above (see Figure 1), the level of hydrophobicity of each amino acids is peculiar to that amino acid.

Hydrophobic residues at interfaces tend to form clusters in comparison to non-hydrophobic residues (Neuvirth *et al.* 2004). In addition, interfacial sites tend to comprise of both hydrophobic residues and polar residues. These clusters account for the synergistic network of binding forces occurring at interfaces: both transient and obligatory interfaces.

2.3 Protein Interfaces

2.3.1 Protein Structural Organization and Conservation

Amino acids are the building blocks of proteins (see **Figure 2.3**), several studies from various fields of proteomics, structural genomics and bioinformatics have published substantial scientific evidences of how protein structure (through conservation of amino acids) correlates strongly to protein function (Tsai *et al*. 1996; Vihinen *et* al. 2000; Neuvirth *et al.* 2004; Keskin and Nussinov 2005; Liang *et al.* 2006; Porollo and Meller 2007; Khan and Vihinen, 2010).

Only some parts of the entire protein surface are actively involved in biological processes; they are known to be functionally active sites by interacting with other molecules including other proteins. Evolutionarily important residues are most probable to be found at interfaces, and these have a high potential to be conserved as well (Chen and Zhou, 2005). In previous studies (Armon *et al*. 2001; Landgraf *et al.* 2001; Lichtarge and Sowa, 2002), conservation of residues at interfaces was a key criteria in predicting "hot spot residues" that are involved in protein-protein interactions. These residues were observed to be more conserved than other surface residues (Liang *et al.* 2006).

2.3.2 Protein Surface and Protein Interface

Protein interactions are pivotal to the optimal function of every protein- a concept that is continuously being researched especially with the exponential technological advancement in fields such as structural genomics (Zhou and Qin, 2007). Several research groups and institutes have developed bioanalytical web servers and/or stand-alone software installations that assist in predicting interface residues for any given protein (Chen and Zhou, 2005; Liang *et* al., 2006; Porollo and Meller, 2007; Qin and Zhou, 2007)

Following the first successful automated method by Zhou and Shan (2001) that was tailored to predict interface residues in protein-protein complexes, immense efforts has been channeled in exploring the possibilities such methods provide. In 2004, Neuvirth *et al.* suggested that protein interaction does not occur throughout the entire surface of a protein but only in a few parts, and these few interacting surfaces of the protein share common physiochemical properties.

Porollo and Meller (2007) further developed a method to predict interface residues using relative solvent accessibility (RSA) as the basis of the algorithm. The RSA-based prediction method further differentiated between interacting and non-interacting residues on protein surfaces.

2.3.3 Definitions for Interaction Site / Interface

The definition for an "interaction site" stems from the method of prediction specified in the algorithm used for that prediction. Porollo and Meller's (2007) method of RSA defined an interaction site as "the difference between unbound and bound (complex) structure of an individual chain". This can be simplified as: computing the difference between the values of (a) the exposure of an amino acid residue before binding with another molecule and (b) its exposure after it has form a complex upon binding. Noticeably, the former value should be higher than the latter for those residues predicted to be found in interfaces.

Using a consensus neural network based on sequence profiles and solvent accessibility, Chen and Zhou (2005) designed a method for the prediction of interface residues between protein-protein complexes; they established a definition for interfacial contact between protein complexes derived from PDB to be: "a pair of heavy atoms from two sides of an interface that are within 5 Å".

Generally, protein-protein interface predictors have one or more of the following definitions for interfaces which are based upon:

1. The relative distance between the van der Waals surfaces of the atoms in partner chains

2. The relative distances between the centers of the atoms in different protein chains
3. The relative distance between alpha carbon atoms in protein chains (Jordan *et al*. 2012)

2.3.4 Types of Protein-Protein Interfaces

An extensive analysis carried out at the Columbia University, NY by Ofran and Rust (2003) enumerated six significant types of protein-protein interfaces using just sequence features of the proteins. They were classified in the following category:

- Intra-domain: interfaces within one structural domain
- Domain–domain: interfaces between different domains within one chain
- Homo-obligomer: interfaces between permanently interacting identical chains
- Homo-complex: interfaces between transiently interacting identical protein chains
- Hetero-obligomer: interfaces between permanently interacting different protein chains
- Hetero-complex: interfaces between different transiently interacting protein chains

The term "obligomer" refers to "interfaces between residues from two chains that are obligatory" (Ofran and Rust, 2003).

2.3.5 Characteristics of Protein Interface Residues

Irrespective of the methods used for prediction of interface residues, the result of most method predict similar "hot spot" residues for the same protein. This unison-like results from various prediction methods are mostly due to the physiochemical properties common to interface residues (Zhou and Qin, 2007) or at the interface itself.

Currently, emerging prediction methods combine various prediction algorithms in other to increase the accuracy of prediction (Qin and Zhou, 2007). The most eminent characteristics of protein interfaces and/ interface residues include:

.3.5.1 Solvent accessibility:

Porollo and Meller (2007); Chen and Zhou (2005) ascertained that interface residues have higher solvent accessibility than non-interface residues. Since non-interface residues do not form complexes with other external molecules, they tend to be buried further in the protein structure where they exert more intra-molecular interactions. Thus, having lesser accessibility to the surface of the protein (Jones and Thornton, 1996; Zhou and Qin, 2007;).

2.3.5.2 Sequence Conservation:

In comparison to non-interface residues, residues found at interfaces are evolutionarily conserved (Liang *et al.*, 2007). Some of the prediction methods utilize this property when designing their algorithm or during data training of neural networks to set certain parameters e.g. sequence coverage (Chen and Zhou, 2005).

2.3.5.3 Distribution of amino acids:

Certain studies have observed that some amino acids, especially arginine are more abundant at protein interfaces than others (Zhou and Shan, 2001). Crowley and Golovin (2005) attributed this richness in arginine at protein interfaces to cation-π interactions.

2.3.5.4 Chemical Properties:

Protein interfaces have been observed to be predominantly hydrophobic (see Figure 1) - as a result of abundant hydrophobic amino acid residues located at interfaces (Neuvirth *et* al. 2004). In addition, the evolutionarily active residue positions "hot spots", are oftentimes conserved and conflicting reports suggest that they are composed of polar and/or non polar residues (see Figure 2) (Zhou and Shan, 2001; Hu *et al.,* 2000; Glaser *et al.,* 2001; De Lano, 2002;) except for arginine (Zhou and Shan, 2001). Also, the analysis of Chen and Zhou (2005) showed that non polar as well as charged residues were more favored in interfaces. The non polar residues observed to be abundant include: leucine, isoleucine, phenylalanine, tyrosine valine and methionine.

2.3.5.5 Structural Properties:

Structurally significant observations regarding interfaces suggests that they frequently appear in between domains, in large proteins (Jones and Thornton, 1997; Lo Conte *et al*., 1999; Ma *et* *al*., 2002). Interfaces are usually circular (Kleanthous, 2000) and loops are often present at the edges of interfaces, contributing about 40% of interfacial contacts (Miller, 1989).

Additionally, Neuvirth *et al.* (2004) discovered from their analysis that β-strands are more favored than α-helices at interfaces. It was suggested that the flat surfaces of β-strands appears to form a favorable three-dimensional binding opportunity at interfaces in comparison to the cylindrical surface of α-helices.

2.3.5.6 Forces at interfaces:

Many hydrophobic and electrostatic interactions at interfaces are known to occur by the means of hydrogen bonds; due to the nature of transient complexes, hydrogen bonds are suggested to be more abundant, accounting for the occurrence of a weak affinity (Jones and Thornton 1995; Lijnzaad and Argos, 1997).

Salt and disulphide bridges have been observed to be present at interfaces. Even though they are rare, disulphide bonds appear to have a large stabilizing effect when they occur at interfaces (Neuvirth *et al*., 2004).

2.3.5.7 Interface residue-energy distribution:

Previous efforts to identify interface residues involved the analysis of entropic activity of interface residues. Researchers such as Elcock (2001), predicted interface residues by analyzing their high electrostatic energy. Others explored the differences in free-energy of amino acids to identify interface residues (Cheng *et al.* 2005).

2.3.6 Types of Complexes at Interaction Sites

In 2007, Porollo and Meller suggested that various complexes of proteins maybe involved in several interactions that occur at interface regions. Earlier studies in structural genomics proposed the various types of complexes formed by protein interaction. They include but are not limited to: transient versus obligatory complexes, homodimers versus heterodimers, enzyme binding complexes versus other complexes, (Jones and Thornton, 1996; Jones and

Thornton, 1997; Hu *et* al., 2000; Glaser *et* al., 2001; Zhou and Shan, 2001; Ofran and Rost, 2003).

2.4 Human Bruton Tyrosine Kinase (BTK) Protein

The BTK protein has been identified in about sixty species, and it is most predominant in mammals (Hubbard *et al.*, 2007). In humans, BTK -an enzymatic cytoplasmic protein, is usually expressed in cells of the immune system. BTK in humans is responsible for B-lymphocyte development, differentiation, activation and signaling (Yu *et al.* 2006).

The subcellular location of the BTK protein in humans include: cytoplasm, nucleus, plasma membrane and peripheral membrane protein; it is expressed in numerous tissues and the most abundant of these are the lymph, tonsil, blood, lymph nodes, bone marrow, spleen and connective tissues (Várnai *et al.* 1999; Nore *et al.* 2000; Mohammed *et al.* 2000; Vargas *et al* 2002). In addition, BTK is known to be expressed in hematopoietic cells, excluding T lymphocytes and terminally differentiated plasma cells (Smith *et al.*, 1995).

The gene coding for BTK in humans is known to span over 36 kb and is composed of 19 exons (Sideras *et al.*, 1994). Vetrie *et al.* (1993) described the complete nucleotide sequence of the mRNA (cDNA) encoded by human BTK to be having 659 amino acids.

2.4.1 BTK Domains and Interaction

BTK protein has five functional domains: The end terminal Pleckstrin homology, Src homology 3, Src homology 2, Tyrosine kinase and Tec homology domains also known as PH, SH3, SH2, TK and TH respectively (Vihinen *et al.* 1994; Smith *et al.* 1994; Gustafsson *et al.* 2012). These domain regions have been studied to exhibit high evolutionary conservation on both sequence level and structural level (Thusberg and Vihinen, 2009).

2.4.1.1 PH Domain

(Pleckstrin homology domain): PH domains are the 11th most common domain in the human genome (Lemmon, 2007). In reference to Haslam *et* al. (1993) and Mayer *et al*. (1993), it derives its name as a result of the sequence elements found to be repeated in both the N-terminal and C-terminals of pleckstrin.

"Pleckstrin", the protein where the PH domain was first detected, is a major substrate for protein kinase C in platelets and leukocytes (Jackson *et al.* 2011). PH domains consist of approximately 120 residues (Vihinen *et* al., 1994) and although there is little conservation among PH protein families, their tertiary structure is quite similar consisting of a conserved of a β-barrel composed of two perpendicular anti-parallel β-sheets followed by a C-terminal amphipathic helix (Ferguson *et al*. 1995; Saraste and Hyvönen, 1995).

Functionally, PH domain is known to aid in the activation of the catalytic transphosphorylation of the BTK protein (Li *et al*., 1997) by the binding to end products of phosphatidylinositol 3-kinase (PI 3-kinase) family which includes phosphatidylinositol 3,4,5-trisphosphate (Salim *et al.,* 1996; Rameh *et al.*, 1997) and inositol 3-phosphates (Fakuda *et al*., 1996).

Earlier studies (Ferguson *et al.,* 1995; Fakuda *et al*., 1996; Salim *et al*., 1996; Rameh *et al*. 1997) using computer-generated models suggested that residues located in protein-protein interfaces which are essential for the binding of these lipid molecules within the BTK PH domain includes L12, F25, and R28.

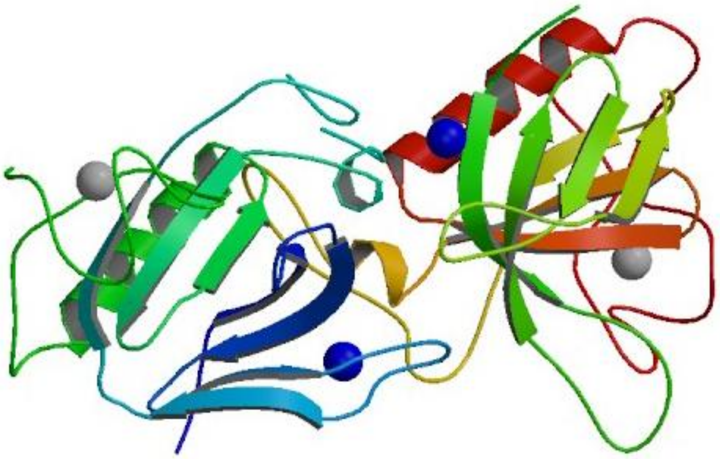


Figure 2.4 A structure from PDB (1BTK) showing PH domain and BTK motif from human Bruton tyrosine kinase protein; included is the position of the mutant arginine residue (R28C), associated with XLA in humans (Hyvönen and Saraste, 1997). The ball representations in the structure denotes the positions of the mutant residue R28C in both chains of the molecule.

2.4.1.2 SH2 and SH3 Domains

The Src homology (SH), a region of homology between two tyrosine kinases that lay outside the catalytic domain, was discovered by Pawson's group in 1986 and termed SH2 and SH3 based on the variety of proteins they contain and their respective catalytic activity (Mayer and Baltimore, 1993). Cytoplasmic proteins containing SH domains are known to be involved in signal transduction (Mayer, 2001) and they also appear to mediate controlled protein-protein interactions (Mayer and Baltimore, 1993).

SH3

In BTK protein, sequence annotation suggests that the SH3 domain proceeds the SH2, and this order is also reflected in the catalytic activity of both domains (Mayer and Baltimore, 1993). SH3 is a small domain of about 60 amino acid residues that is present in a large number of eukaryotic proteins (Musacchio *et al.* 1992); they have a moderate affinity and specificity for proline-rich ligands that affects myriads of biological processes ranging from regulation of enzymes by intra-molecular interactions, increasing the local concentration or altering the subcellular localization of components of signaling pathways to mediating the assembly of large multi-protein complexes (Mayer, 2001). SH3 domain belong to the class of Pro-rich binding domains (Gushchina *et al.* 2011).

Structurally, the basic fold of many SH3 domains contains five anti-parallel beta-stands that are closely packed to form two perpendicular beta-sheets. The ligand-binding site consists of a hydrophobic patch that contains a cluster of conserved aromatic residues and is surrounded by two charged and variable loops; SH3 domains bind to Pro-rich peptides that form a left-handed polyPro type II helix, with the minimal consensus Pro-X-X-Pro (Gushchina *et al.,* 2011). Each Pro is usually preceded by an aliphatic residue. Each of these aliphatic-Pro pairs binds to a hydrophobic pocket on the SH3 domain (Nguyen *et al.*, 1998; Gushchina *et al.,* 2011).

Okoh and Vihinen (2002) established an association between the SH3 and TH domains in BTK which occurs in a regulative fashion. In another study by Patel *et al* (1997), a point amino acid variation that led to the deletion of 14 amino acids at the terminal end of the SH3 domain was observed in a patient with XLA. They suggested that the consequence of such variation led to reduction in the binding activity between SH3 and TH domains. Thus, a probable etiology for the XLA in the patient being studied.

SH2

In respect to domain class, SH2 belongs to the Phospho-Tyr binding domains that function as regulatory modules of intracellular signaling cascades by interacting with phosphotyrosine-containing peptides and proteins, and are usually composed of approximately 100 amino acids (Bibbins *et al*., 1993). Vihinen *et al*., (1996); Vihinen *et al*., (1997); Vihinen *et al*., (1999) showed that pathogenic variations in the SH2 domain of BTK are associated with impaired B cell function and this may result in the XLA immunodeficiency in humans, as a result of a disruption in the phosphotyrosine binding sites. In one study by Tzeng *et al*., (2000) the BTK SH2 domain was found to be essential for phospholipase C-γ phosphorylation by the interaction of B-cell linker protein (BLNK), and variations in the SH2 domain were shown to cause XLA.

Structure-wise, SH2 domains typically contain a central anti-parallel beta sheet that is surrounded by two α-helices; the loop between the second and third β-strands provides many of the binding interactions with the phosphate group of its phosphopeptide ligand (Pawson *et* al., 2002). According to studies by Waksman *et al.*, (1992) "conserved residues contribute to the hydrophobic core or are involved in pY (phosphorylated Tyrosine) recognition while more variable residues contribute to specific recognition of C-terminal residues". Furthermore, "an invariant arginine residue in the SH2 domain coordinates the phosphate oxygens of pY and is essential for high affinity phosphopeptide binding".

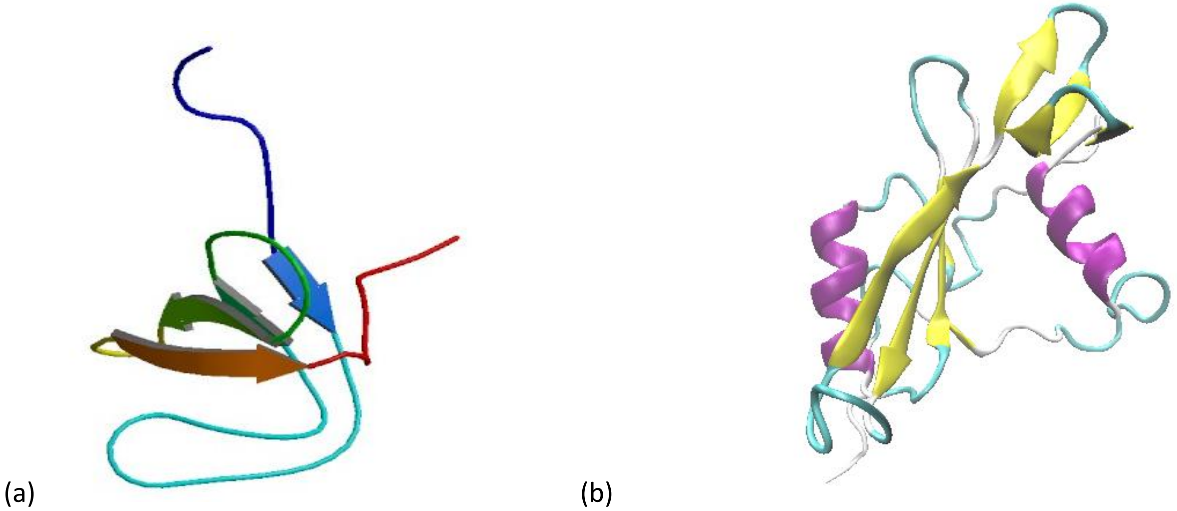


Figure 2.5 (a) Cartoon structure (PDB:1AWX) by Hansson *et al.* (1998) showing the anti-parallel beta-stands of SH3 domain from human BTK protein. Patel *et al* (1997) suggested that the C-terminal end of the domain is responsible for its binding activity with TH domain within the same BTK protein. (b) SH2 domain structure (Huang *et al,* 2006) of human BTK protein showing anti-parallel beta-stands flanked by two α-helices. Its side chains are responsible for interactions with phosphotyrosine-containing peptides.

The sequence annotation of the human BTK protein and available structures have been tabularized below. Some of the domains were observed to have more than one solution structure in the PDB database.

Table 2.2 Sequence annotation of human BTK protein showing various regions (The UniProt Consortium, 2012). Representative structures derived from RCSB PDB (Berman *et al.*2000) are also included in the table.

Regions	Residues	Length	Description	Pdb_id
Domain	3 – 133	131	PH	1BTK (1.60)
Domain	214 – 274	61	SH3	1AWX (NMR)
Domain	281 – 377	97	SH2	2GE9
Domain	402 – 655	254	Tyrosine kinase	3P08 (1.50),
Zinc finger	135 – 171	37	Btk-type	
Nucleotide binding	408 – 416	9	ATP	
Region	12 – 24	13	Inositol-(1,3,4,5)-tetrakisphosphate 1-binding	
Region	474 – 479	6	Inhibitor-binding	
Motif	581 – 588	8	CAV1-binding	

Table 2.3 Domain wise categorization of BTK family kinases protein interactions

DOMAIN	DOMAIN DESCRIPTION	INTERACTING PATNERS
PH and TH	• Multifunctional domain assumed to bind to phospholipids • The interaction during signal transduction is likely to be transient and dynamic • Linked to XLA (Rawlings *et al.*, 1993; Thomas *et al.*, 1993)	1. Phosphoinositides 2. heterotrimeric G-protein subunits 3. F-actin 4. PKC isoforms 5. BP-135/TFII-I 6. STATs 7. FAK 8. Fas

		9. PTPD1
SH3	• It recognizes proline-rich motifs in many proteins • Regulates BTK via binding to regulatory proteins or internal folding • So far, no pathogenic SNPs leading to XLA has been found	1. CD 28 2. c-Cbl 3. WASP 4. ZAP-70
SH2	• Recognizes phosphotyrosine containing peptides and proteins • Linked to XLA (Vihinen *et al.* 1999)	1. Vav 2. BLNK 3. Dok-1
Kinase	• Highly conserved domain (Qiu *et al.* 1998b) • Involved in regulation	1. BRDG1 2. PI3K 3. GRB10

Other examples of interacting partners with human BTK include:
CD4: T-cell surface antigen T4/Leu3 (Wang, 2004)
CD34: Hematopoietic progenitor cell antigen CD34 precursor (Meffre *et al.*, 1997)
HCPH: Hematopoietic cell protein-tyrosine phosphatase (KEGG pathway: hsa04662)
LYN: Tyrosine-protein kinase (KEGG pathway: hsa04662, hsa04664)
BAD: Bcl2 antagonist of cell death (Ottilie*et al.*, 1997; Danial, 2003)
HCK: Hemopoietic cell kinase (Cheng *et al.*, 1994)
RAG2: Recombination-activating protein 2 (Kouro *et al.*, 2001)
RAL: Ral-A precursor (de Gorter *et al*., 2008)
LCP2: Lymphocyte cytosolic protein 2 (Liu *et al.*, 2006)
CD22: B-cell receptor CD22 precursor (Moschese *et al.*, 2004)
GTF2I: General transcription factor II-I (Sacristán *et al.*, 2004)
DRP2: Dihydropyrimidinase-related protein 2 (Sedivá *et al.*, 2007)
IGB: Immunoglobulin beta (Lougaris *et al.,* 2008)
GP6: Glycoprotein VI (platelet precursor) (Yi *et al.*, 2005)
BCM: B-cell maturation protein (Jin *et al.*, 2008)
Sab: SH3-domain binding protein 5 (BTK associated) (Yamadori *et al.*, 1999)
Ly49: Killer cell lectin-like receptor subfamily A, member 1 (Mestas and Hughes, 2004)
PLCG1: Phospholipase C-gamma-1 (KEGG pathway: hsa04664)
SPNS1: Sphingolipid transporter

FKBP4: FK506 binding protein 4
CD164: Putative mucin protein precursor 24 (Kneidinger *et al.*, 2008)
PLCG2: Phospholipase C-gamma-2 (KEGG pathway: hsa04662)
ACTL6B: Actin-like protein 6B
PRDM1: PR domain zinc finger protein 1 (Takatsu *et al.*, 2004)
PIK3R5: Phosphoinositide 3-kinase regulatory subunit 5 (KEGG pathway: hsa04664)
RPS6KB2: Ribosomal protein S6 kinase, 70kDa, polypeptide 2
KHDRBS1: KH domain-containing, RNA-binding, signal transduction-associated protein 1 (Guinmard *et al.*, 1997)
TNFRSF10D: Tumor necrosis factor receptor superfamily member 10D precursor (Bouralexis *et al.*, 2003)

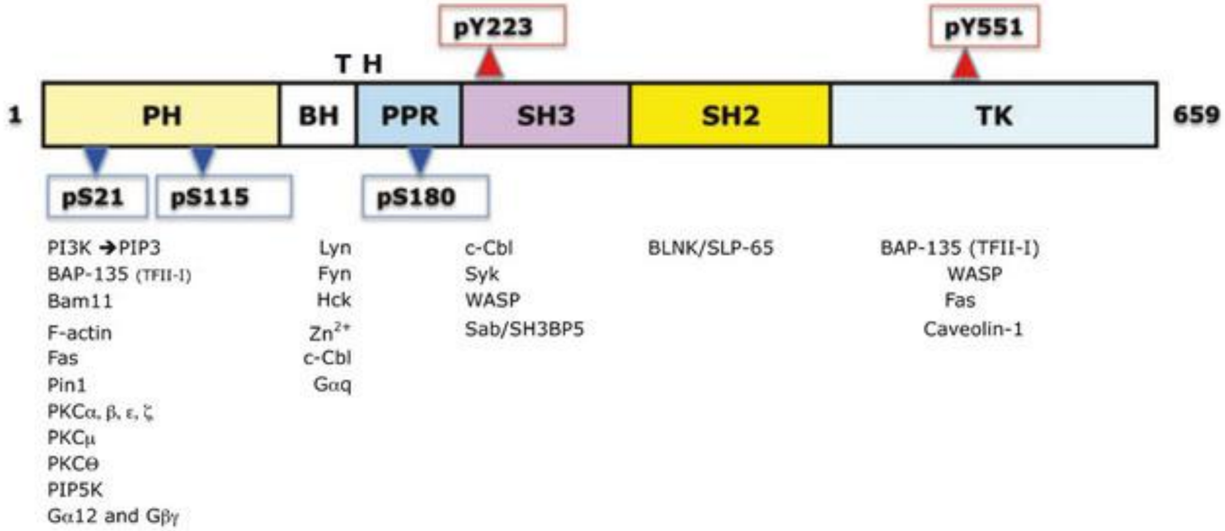


Figure 2.5 Annotated domains of the BTK and interacting partners (Mohamed *et al.*, 2009)

2.5 Human BTK and X-linked agammaglobulinemia (XLA)

BTK is an highly conserved cellular protein, variations in functionally known conserved residues are mostly responsible for disease conditions as observed in XLA (Chen and Zhou, 2005). It belongs to the Tec family of protein tyrosine kinases (PTKs) and it is involved with many crucial cellular processes; pathogenic variations affecting BTK results in developmental deformities in the maturation of B-lymphocytes.

X chromosome-linked (alternatively X-linked) agammaglobulinemia, generally known as Bruton agammaglobulinemia (Bruton, 1952), is a severe disease condition characterized by failure in normal maturation of the B lymphocytes. It is associated with pathogenic variations in the gene coding for BTK (Lindvall *et al.* 2005). Statistically, XLA has an average frequency of 1/150,000 male births in humans (Smith *et al*., 1995).

From the serology outlook, individuals with XLA generally have low serum immunoglobulin and (Campana *et al.*, 1990) are unable to produce adequately crucial components of the immune system- the immunoglobulins (Conley *et al.*, 1986). Consequently, such individuals with this type of compromised immune system have been observed to having increased susceptibility to certain microbial infections: both bacterial (Lederman and Winkelstein, 1985) and *Enterovirus* species infections (McKinney *et al.*, 1987).

2.6 Genetic Variation: Polymorphism in BTK

Genetic variations such as Single Nucleotide Polymorphisms (SNPs) result into non-synonymous SNPS (nsSNPs) pathogenic SNPs, otherwise known as disease-causing variations (Cargill *et al.*, 1999; Halushka *et al*., 1999). The occurrence of pathogenic SNPs has been observed to affect the function of a protein after post-translational modifications (Lindvall *et al.* 2005; Thusberg and Vihinen, 2009).

The prevalence of pathogenic SNPs during protein transcriptional processes such as the formation of pre m-RNA, mRNA splicing, mRNA stability and alternative splicing (Thusberg and Vihinen, 2009), have been observed to account for less than 10% of all nsSNPs cases (Stenson *et al*., 2003).

Although the precise mechanism of action of some of these pathogenic SNPs have not been fully understood, amongst many, XLA (X chromosome-linked agammaglobulinemia) is as a result of pathogenic SNPs occurring in the gene coding for human BTK protein (Lindvall *et al.* 2005).

2.7 Application of Protein-Protein Interface Predictions

Protein interactions form the core of understanding the diverse functions of several biomolecules, and various methodologies of predicting interfacial regions of proteins have various applications in the field of molecular science, especially proteomics. Knowledge from protein interface prediction can be applied to solve problems in certain areas such as binding and docking site predictability, protein-protein interface regions etc.

2.7.1 Functional Sites

The binding activity occurring at interaction sites of proteins mostly suggest that they are the functional parts of the molecules. Algorithms that are used in the prediction of functional sites mostly emphasize evolutionary relations as seen in phylogenetic trees (Landau *et al.*, 2005). In addition of evolutionary relations, the physiochemical and structural properties of residues amplifies the accuracy of functional site prediction (Zhou and Qin, 2007). As seen in BTK, pathogenic variations are known to occur in protein interfaces (Zhou, 2004; Brautigam *et al.*, 2006) and computational prediction of interfaces can help understand the disease mechanisms and subsequent design of therapeutic agents (Zhou and Qin, 2007).

2.7.2 Docking Sites

Qin and Zhou (2007) defined docking as "the procedure by which the structure of a protein complex is built from the unbound structures of the subunits". The process of docking presents a complication in identifying which sets of subunits fits and how to rank accurately the lists of probable sets. Protein-protein interface prediction helps to alleviate the problems arising from conformational changes that occur during the docking process: commonly referred to as the front end use (Zhou and Qin, 2007; Qin and Zhou, 2007b). Protein-protein interface predictions are first used to determine if two protein molecules interact, subsequently, docking predictions are employed to determine which regions of the molecules interact what type of interaction takes place- obligate or non-obligate; transient or obligatory (Li and Kihara, 2012).

2.7.3 DNA-binding Sites

The underlying rationale that guides the prediction of protein-protein interfaces has been utilized to predict DNA-binding sites on proteins that interact with DNA (Zhou and Qin, 2007). Earlier studies have utilized e.g. the neural network method (Ahmad *et al.*, 2004; Tjong and Zhou 2007) to address the prediction of DNA-binding sites on proteins with much success. Although similar in rationale, the non-DNA contacting regions of a protein can be distinguished from DNA-binding sites by peculiar characteristics- DNA-binding sites are enriched in positively charged Arg (R) and Lys (K) residues and depleted in negatively charged Asp (D) and Glu (E) residues (Zhou and Qin, 2007).

2.7.4 Drug discovery and design

Experimental methods have been used in times past to identify therapeutic substances from various sources including microorganisms e.g. penicillin from *Penicillium* fungi. The use of protein-protein interface prediction technologies enables the specification of vital components of organism's interactome that requires further research (Fuentes *et al.*, 2009). The knowledge of protein-protein interface networks empowers pharmaceutical proteomists to distinguish which protein-protein interaction network an upcoming therapeutic agent should be targeted; the optimal design such agent would best deliver the intended therapeutic effect (Grosdidier *et al.*, 2009).

2.7.5 Etiology of Diseases

Several diseases of genetic origin are consequences of miscommunication between protein-protein interactions at genomic level. In a recent article by Hosur *et al.* (2011), a protein-protein interface predictor (iWRAP) was designed and tested on set of yeast genes that were related to cancer. This dataset of disease genes were derived from the CYGD by Güldener *et al.*, (2005). iWRAP successfully identified peculiar genes that were subjected to further experimental study (Hosur *et al.*, 2011)

3 STUDY OBJECTIVES

The meta-analysis of the interface region of human BTK protein with respect to known disease-causing SNPs was performed with the following aims and objectives:

1. To classify the RSA of pathogenic SNPs associated with human BTK using various single and meta-web servers .

2. To assess variation "hotspots" in human BTK domains.

3. To map predicted pathogenic SNPs in relation to their interfacial locations on the domains of human BTK protein.

4. To analyze the possible implications of pathogenic variations situated at the BTK protein interface.

5. To statistically compute the abundance of each of the twenty amino acids across the predicted interface regions of human BTK domains.

4 MATERIALS AND METHODS

4.1 Dataset of Missense Variations (SNPs)

Approximately 80% of patients suffering from agammaglobulinemia have a pathogenic variation in the gene coding for human BTK protein (Vihinen *et al.*, 1999). A set of data containing pathogenic amino acid substitutions in human BTK protein and their corresponding domains was obtained from BTKbase: http://bioinf.uta.fi/BTKbase/?content=pub/IDbases (August, 2012).

The BTKbase is a publicly available biological database that continuously curates clinical cases of agammaglobulinemia (Vihinen *et al,* 1998; Vihinen *et al.* 1999; Väliaho *et al.* 2006) and is continuously maintained (Version 8.52, last update: 16th June, 2011) by the bioinformatics group at the Institute of Biomedical Technology (IBT), University of Tampere, Finland. The webpage is available at: http://bioinf.uta.fi/BTKbase/

The dataset considered for analysis contained 560 amino acid substitutions (pathogenic SNPs) which included: 105 entries for PH domain, 11 entries for the Zinc finger region, 106 entries for SH2 and 338 entries for the kinase domain. Apart from entries for zinc finger region, all other entries were used in the domain-wise analysis of protein interfaces in the human BTK. As of the time of this analysis (August 2012), there is no known pathogenic amino acid substitution that causes XLA in the SH3 domain of human BTK protein.

From the BTKbase data, it was observed that missense variations caused by single nucleotide substitutions signified that it contained a limited subset amino acid changes (i.e. given the physiochemical properties of amino acids, every residue at each position cannot change to all other 19 residues). There were however, multiple entries for few amino acid residue positions that subsequently suggests the possibility that those positions are 'hotspots'- amino acid positions more susceptible to variation. **Table 8.4** in the appendix shows the raw datasets grouped according to domain.

4.2 Sequences and Structures

In order to understand the level of variation in the dataset, reference sequences for the BTK domains and the human BTK protein were obtained from NCBI's RefSeq database (except for SH3 domain because currently, as at the time of this thesis work, there are no known amino acid substitution leading to Agammaglobulinemia) and EMBL-EBI's Uniprot database (The UniProt Consortium, 2012). The sequence ontology feature at Uniprot was used to understand the boarders of each domain and classify them accordingly- PH, SH2, Kinase domains. For the analysis and rendering of the results into structural presentation, the VMD (Visual Molecular Dynamics) standalone software by the Theoretical and Computational Biophysics Group at the University of Illinois, was used. According to Humphrey *et al*. (1996), VMD is "a molecular visualization program for displaying, animating, and analyzing large biomolecular systems using 3-D graphics and built-in scripting". It is a free computer software program that is available on major computer platforms. The VMD home page is available at: http://www.ks.uiuc.edu/Research/vmd/

Similarly, through the process of querying the RCSB PDB (Protein Data Bank) (Berman *et al.*, 2000) database with the molecule name- human BTK, structures of PH, SH2 and Kinase domain were obtained. Although there were more than one entry for some of the domains, careful selection was done to obtain the best candidate structure. The selection was done based on the following criteria:

4.2.1 The Experimental Source of the Structure

The methods used to obtain the protein tertiary 3D structure are namely: X-ray crystallography, NMR spectroscopy or electron microscopy (Berman *et al.*, 2000). At the time of this thesis work, there were over 70,000 protein structures solved by X-ray method and just under 10,000 structures solved by NMR method. The source of the structure was considered as a criteria for selection because a few of the protein interface predictors performed poorly (incomplete prediction e.g. PPI-Pred) with protein structures derived from NMR method when compared to those derived through X-ray method. This was as a result of incompatibility of the NMR data file with the algorithm of the web server.

4.2.2 The Resolution [Å] of the Structure

The X-ray method was preferred due to the favorable data structure of the pdb file. Also, with the X-ray data, the resolution [Å] of the candidate protein structure is indicated. The structures of PH (1BTK, 1.6 Å) and Kinase (1K2P, 2.10 Å) domains were derived from X-ray while that of SH2 (2GE9) was from NMR.

4.2.3 The Amount of Chains in the structure

The solution structure of the BTK domains are available in 3D the quaternary structure from PDB database. Since most of the interface predictors needed chain specificity, the nature (identical or not) of the chains and amount were used as another criterion for the selection of a candidate structure. The structures of PH and Kinase domain have two identical chains (A and B), in which chain A was used for the analysis. For the SH2 domain structure, it had only one chain which consequently reduced the ambiguity of chain selection.

4.3 Selection of Interface Predictors

With the emergence of various methods for protein interface predictions, a brief review by Zhou and Qin (2007) of some of the key aspects of protein interface prediction listed some elements one needs to consider while selecting an interface predictor. Some of them include: the characteristics of interface residues (such as sequence conservation, secondary structure, solvent accessibility etc.) and methods of interface prediction (scoring function and neural network, support vector machine etc.).

Six protein-protein interface predictors were selected for this thesis work using the above listed criteria in **section 4.**2. Additionally, the simplicity of the web server's interface; the web server's processing and downtime; the structure and format of the prediction results, were also used to for the selection.

The selected predictors include: **cons-PPISP** (Chen and Zhou, 2005), **PredUS** (Zhang *et al.* 2010), **SPPIDER** (Porollo and Meller, 2007), **Meta-PPISP** (Qin and Zhou, 2007), **PPI-Pred** (Bradford and Westhead, 2004) and **PIER** (Kufareva *et al.* 2007). In other to create congruence in the analysis, these chosen predictors use different algorithms of prediction.

4.3.1 Cons-PPISP (**Cons**ensus **P**rotein**-P**rotein **I**nteraction **S**ite **P**redictor)

Cons-PPISP is a consensus neural-network protein-protein interaction site predictor that was developed by the Zhou group at Florida State University (FSU). The input is the unbound structure of a protein, which is known to bind another protein. It utilizes a trained neural network algorithm to predict which residues will most likely form the binding site for another protein (Zhou and Shan, 2001; Chen and Zhou, 2005). This predictor was selected because the best available structure of the SH2 domain obtained from the PDB database was derived through NMR method. From an earlier research by Chen and Zhou (2005), they reported that cons-PPISP was a fitting predictor for analyzing structures with NMR data. The web server can be located at: http://pipe.scs.fsu.edu/ppisp/

4.3.2 PredUS (**Pred**iction of Protein Interfaces **U**sing **S**tructural Alignment)

PredUS is an interactive web server for prediction protein-protein interfaces through a support vector machine prediction method. According to Zhang *et al.* (2011), "potential interfacial residues for a query protein are identified by 'mapping' contacts from known interfaces of the query protein's structural neighbors to surface residues of the query". In other words, interfacial residues from all known complexes of the query protein and its structural neighbors are sequentially scored and mapped to the query structure. As an example, the image below is a query result generated from the analysis of the PH domain using the PredUS web server.

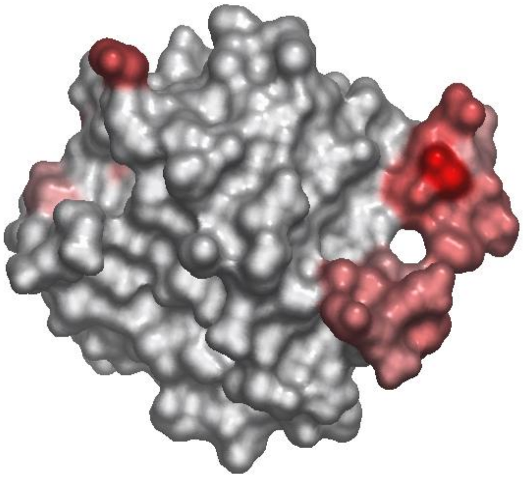


Figure 4.1 An image from PredUS result showing a contact (heat) map of predicted interfacial residues. The higher the residue (contact) score, the darker the shade of red.

Besides the interactive graphical "contact (heat) map" it produces as a result from the query, another reason it was chosen was because of the simplicity of the user interface that is available at: http://bhapp.c2b2.columbia.edu/PredUs/. As an additional feature, the possibility of refining the results of a prediction can be achieved if the user specifies which neighboring structures should be used for contact scoring of the residues in the query protein.

4.3.3 **SPPIDER** (**S**olvent accessibility-based **P**rotein-**P**rotein **I**nterface i**DE**ntification and **R**ecognition)

Jarek Meller's bioinformatics group, of the University of Cincinnati, maintains a web server-SPPIDER (http://sppider.cchmc.org/) that recognizes protein-protein interaction sites through a consensus classifier. The SPPIDER is available in two options: (a) prediction of potential protein-protein interfacial residues (using one protein chain as the query protein) and (b) identification of protein interface with protein-protein complex (The query structure is a complex structure containing more than one protein). The algorithm of the consensus classifier uses relative solvent accessibility (RSA)-based fingerprint to significantly distinguish between the interacting and non-interacting sites of the query protein (Porollo and Meller, 2007). This method claimed to dramatically improve the accuracy of prediction up to 70%.

4.3.4 Meta-PPISP (**Meta** Server for **P**rotein-**P**rotein **I**nteraction **S**ite **P**rediction)

Also from the Zhou group, meta-PPISP is a bioinformatics web tool designed as a meta server that is made up from the combination of three individual web servers: cons-PPISP, PINUP, and Promate. As one would expect, a cross validation of this meta server against the individual servers showed that it has an improved accuracy and coverage (Qin and Zhou, 2007).

Although it is a structure based predictor, it uses linear regression method to compute the scores from the three servers (cons-PPISP, Promate and PINUP) and uses it as an input. It further correlates individual scores for each residue through a series of optimization to produce a final prediction. The result of the meta-PPISP is tabulated and it also includes those from the constituent servers. Meta-PPISP was chosen because of the meta-method it utilizes in its prediction and it is freely available at http://pipe.scs.fsu.edu/meta-ppisp/

4.3.5 PPI-Pred (**P**rotein - **P**rotein **I**nterface **Pred**iction)

This simple web server utilizes a combination of support vector machine and surface patch analysis for predicting probable protein-protein interfacial residues (Bradford and Westhead, 2005). With its simple user interface, it gives four options for viewing the results of a prediction set: (a) a simple tabular listing of predicted highest scoring residue patches or (b) a color-coded sequence view showing predicted residues according to scoring patches or (c) a modified pdb file that shows the predicted residues in colors or (d) a surface view that shows the location of the three patches on the surface using Pred-viewer. The open-access web server is available at http://bmbpcu36.leeds.ac.uk/ppi_pred/index.html

4.3.6 PIER (Protein IntErface Recognition)

This predictor by far had the simplest user interface (Pdb code/file and Chain) and it was the fastest amongst all the protein-protein interface predictors. This predictor has a net prediction accuracy of 60% and also utilizes the patch generation algorithm. In addition to that, it inculcates a local statistical properties of the protein surface at the atomic-group level to the prediction method (Kufareva *et al* 2007).

The results (list of residues) are given in a tabular form with each predicted residue prioritized according to the likelihood of the residue to be in a protein-protein interface. The higher the probability, the more likely the residue would be located in an interface region of the protein. Buried residues are indicated in the results with a probability value of 0.0 while those residues likely to be interface regions are above 30. For this analysis, only those residues with a score above 30 were selected. The web server is available at the address: http://abagyan.ucsd.edu/PIER/pier.cgi?act=abstract

4.4 Computation of Protein-Protein Interface Residues

4.4.1 The Prediction Hit ("+")

All six web server predictors mentioned above were used to analyze interfacial residues that are present in human BTK domains- PH, SH2 and Kinase. Each amino acid residue for all domains that were analyzed by all 6 predictors, a **prediction hit** by each predictor for all predicted amino acid residue was indicated as "+".

4.4.2 Probability (Percentage) Score

A simple arithmetic scoring system was used to assign a score to each **prediction hit** of a residue in terms of a probability percentage. Hence, the sum of probability percentage for all 6 predictors is 100%, with each predictor having a **probability score** of 16.66% for any residue.

For example, if only two out of all 6 predictors predict a residue A to be an interfacial residue, then the probability score would be calculated as:

Probability score of residue A is 16.66% * 2 = 33.32%

In pursuit of simplicity and avoidance of ambiguity, the probability score for each residue were presented in the form of color codes since all possible scores are results of the total amount of predictors.

Table 4.1 showing the color code of probability percentages for a residue predicted to be in the protein-protein interface.

Color Code	Amount of Predictors	Probability percentages % (16.66 * amount of predictors)
🟩	1	16.66
🟦	2	33.32
🟨	3	49.98
🟥	4	66.64
🟦	5	83.30
🟧	6	99.96

4.4.3 Selection of Interfacial Residues

Considering a total of six prediction results from different protein-protein interface predictors, the consequent volume of data generated from the analysis aroused the need to filter only those residues that are most likely to be found in the protein-protein interface. Therefore, where two or more predictors highlight a residue to be in protein-protein interface, those predicted residues were highlighted.

4.5 Rendering of Protein Structure Images

Like the graphical interactive results obtained by using the PredUS (a protein-protein interface predictor web server) (Zhang *et al.*, 2011), similar efforts were made using the VMD (Visual Molecular Dynamics) software (Humphrey *et al.*, 1996). Only those interfacial residues predicted by two or more protein-protein interface predictors were rendered. Such graphical rendering of the protein structures portrayed the predicted residues in relation to: respective secondary structures (α-helices or β-strands); residue accessibility (surface or buried residues); region of the protein.

Graphical representations of the predicted residues in various conformations were designed in order to understand the region in which these interfaces are situated. In addition, possible implications of such predicted interfaces were analyzed by superimposing the predicted residues and the analyzed pathogenic SNPs dataset (associated with human BTK).

5 RESULTS

Domain-wise analysis of protein-protein interface residues in human BTK protein was carried out using six server-based predictors with the exception of the SH3 domain, as stated earlier in the previous section: materials and method. For simplicity and coherency, the results from the domain analysis were tabularized in each section.

5.1 Analysis of PH Domain

The pdb file (1BTK) for BTK PH domain contained 170 residues. Using the UNIPROT annotation data of BTK for interface prediction of the PH domain (residue 3-133), two (meta-PPISP and cons-PPISP) out of the six protein-protein interface predictors did not highlight any interfacial residue in the PH domain between residues 3-133. However, predictions were made by these two web predictors (meta-PPISP and cons-PPISP) for residues between 134-170: 142, 164, 163, 162, 161, 141, 140, 157, 165, 166, 167, 155 and 151.

Although the above predicted residue positions, between position 134-170, were in the pdb file (1BTK) of PH domain, they have been annotated by UNIPROT (The UNIPROT Consortium, 2012 as an active region in the human BTK protein known as 'Zinc finger'. Thus, predictions in the region were excluded from the analysis.

Also, another basis for this shared characteristic is because both web server predictors (Cons-PPISP and Meta-PPISP) share the same algorithm and are both from similar authors (Zhou group, 2007). Prediction results made by Meta-PPISP and cons-PPISP were all outside the region of the PH domain residues (3-133) as annotated by UNIPROT (Berman et *al.*, 2000) but still in the 1BTK pdb file (with a total number of 170 residues).

A total of 131 amino acid residues (between positions 3-133) from the PH domain were analyzed out of which 39 residues were predicted to be involved with a protein-protein interface. However, just one out of any of the six predictors predicted 37 of these 39 amino acid residues to be an interfacial residue (green colored in **Table 5.1**). Also, 2 of these 39 residues had a probability percentage of being an interface residue of 33.32% i.e. as described in **Table**

4.1, 2 out of any of the six predictors predicted them to be an interfacial residue (colored as light blue in **Table 5.1**). A graphical representation of all predicted residues can be seen in **Figure 5.4.** Correlating the predicted interface residues (39) and the known pathogenic SNPs in the PH domain protein, only one out of the recorded 28 polymorphic amino acid positions were predicted to be in the protein-protein interface: V64 had a pathogenic polymorphism amount of 4.

Table 5.1 The table shows a cross analysis of disease-causing SNPs and secondary structure in PH domain of BTK using selected protein-protein interface predictors. The "+" denotes *Prediction hit* of those residues that were predicted to be in the interface region by any one of the predictors.

SECONDARY STRUCTURE	UNIPROT (WILD) POSITION 3-133	RESIDUE	PATHOGENIC SNPs STATISTICS	INTERFACE PREDICTORS					
				Cons-PPISP	PredUS	SPPIDER	Meta-PPISP (cons-PPISP, PINUP, Promate)	PPI-Pred	PIER
	3	A				+			
	4	V				+			
	5	I				+			
B-strand	6	L							
	7	E				+			
	8	S							+
	9	I							+
	10	F	1						
	11	L	4						
	12	K	1						
	13	R							
	14	S	3						
	15	Q							
	16	Q							
	17	K	1						
B-strand	18	K							
	19	K	2						
	20	T							
	21	S							
	22	P							
	23	L							
	24	N							
B-strand	25	F	1						
	26	K							
	27	K	1						
	28	R	40						
	29	L							
	30	F							
	31	L	1						
	32	L	5						
	33	T	6						
	34	V							
	35	H							
	36	K							
	37	L	1						
	38	S	1						

	39	**Y**	7						
	40	**Y**	5						
	41	**E**							
	42	**Y**			+				
	43	**D**			+				
Turn	44	**F**			+				
	45	**E**			+				
	46	**R**			+				
	47	**G**			+				
B-strand	48	**R**			+				
	49	**R**							
	50	**G**							
	51	**S**							
	52	**K**			+				
	53	**K**							
	54	**G**							
	55	**S**			+				
	56	**I**	1						
	57	**D**							
Helix	58	**V**							
	59	**E**						+	
	60	**K**						+	
B-strand	61	**I**	6						
	62	**T**						+	
	63	**C**						+	+
	64	**V**	4					+	
	65	**E**						+	
	66	**T**						+	
	67	**V**							
	68	**V**							
	69	**P**							
	70	**E**							
	71	**K**							
	72	**N**				+			
	73	**P**							
	74	**P**							
Helix	75	**P**							
	76	**E**							
	77	**R**							
	78	**Q**							
	79	**I**							
	80	**P**							
	81	**R**							
	82	**R**							
	83	**G**							
	84	**E**							
	85	**E**							
	86	**S**							
	87	**S**							
	88	**E**							
	89	**M**			+				
	90	**E**			+				
	91	**Q**			+	+			
	92	**I**			+				
Helix	93	**S**							
	94	**I**			+				
	95	**I**			+				
	96	**E**			+				
	97	**R**							
	98	**F**	2						
	99	**P**							
B-strand	100	**Y**							
	101	**P**							

	102	F							
	103	Q	1						
	104	V							
	105	V						+	
	106	Y						+	
	107	D						+	
	108	E			+				
	109	G							
	110	P						+	
B-strand	111	L	1						
	112	Y	1						
	113	V	1						
	114	F							
	115	S	1						
	116	P							
	117	T	12						
	118	E							
Helix	119	E							
	120	L							
	121	R							
	122	K							
	123	R							
	124	W	3						
	125	I						+	
	126	H							
	127	Q	1						
	128	L							
	129	K						+	
	130	N				+			
	131	V							
	132	I						+	
	133	R						+	

Hyvönen *et al*. (1995) and Ferguson *et al*. (1995) described the structure of the PH domain of BTK to consist of two perpendicular anti-parallel β-sheets that make up the β-barrel, followed by a C-terminal amphipathic helix. In addition to the interface and pathogenic SNP analysis, further correlation between the secondary structures in the PH domain and the predicted interface residues as well as pathogenic SNPs residues was done.

Table 5.2 Table showing the distribution of pathogenic SNPs and predicted interfacial in correlation to their secondary structures in the PH domain of Human BTK protein

Color code	Secondary structures	Amount of SNPs positions in secondary structures (Total = 28)	Amount of residues predicted to be in interface region (Total = 39)
🟧	A-helices	2 (7.14%)	7 (17.95%)
🟪	B-strands	18 (64.29%)	13 (33.33%)
▨	Turns/loops	0 (0.00%)	4 (7.55%)

As seen in **Table 5.2**, 33.33% (13 residues) of the predicted interface residues are situated in β-strands while approximately 18% were found to be embedded in the helical secondary structure of the PH domain protein.

Using the VMD standalone computer software program (Humphrey *et al*., 1996), graphical illustrations of the location of the predicted interfacial residues were performed. Amongst the illustrations is a representation showing probable binding patches and pockets situated at the surface of the protein (Fig. 5.1a) and also in buried interfacial residues that is observable in the secondary structure representation (Fig. 5.1b).

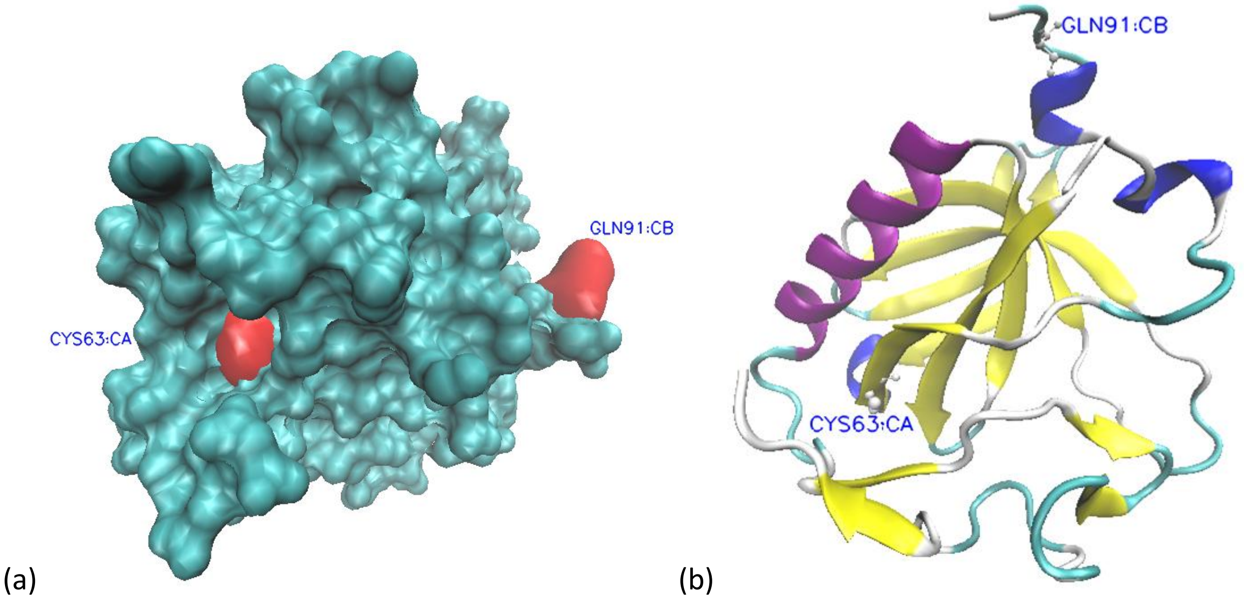
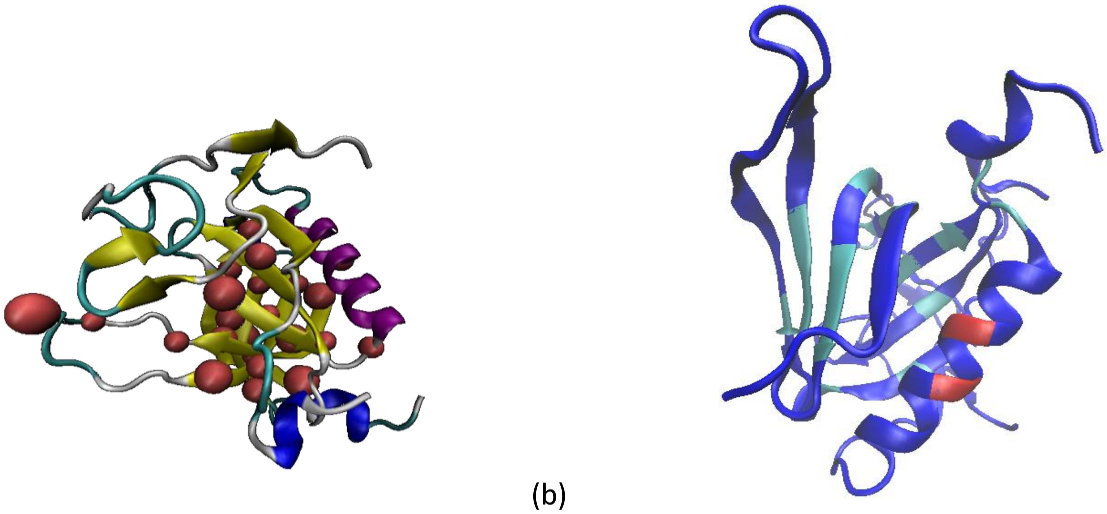


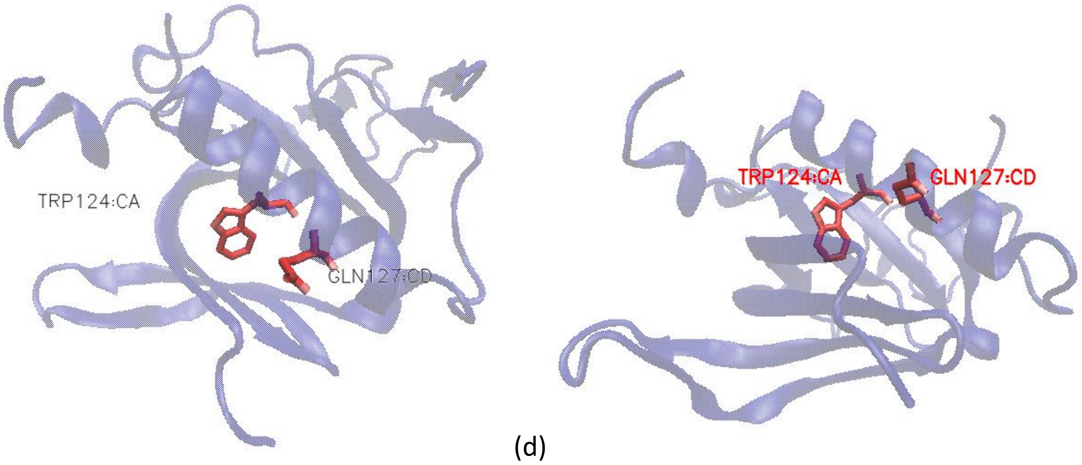
Figure 5.1 (a) Surface representation of the PH domain showing the location of those interface residues predicted by at least two predictors (in red= C63 and Q91) coded as color 'blue' in **Table 5.1.** (b) Also, cartoon structure of the PH domain showing the location of the predicted interface residues (C63 and Q91) by at least two predictors; the figure further depicts the location of the interface residues in the secondary structure of the PH domain.

Additionally, upon mapping the pathogenic SNPs data set of the PH domain protein to their corresponding secondary structure, it was observed that 18 out of 28 SNPs positions (~ 65%) occur in regions of β-sheets, some of which are buried residues within the structure of the protein. Just two out of the 28 pathogenic SNPs residue positions were found to be situated in α-helices. No known SNPs residue positions were observed in turns/loops.



(a) (b)

Figure 5.2 (a) Secondary structure representation showing the location of pathogenic SNPs residues (as red beads) in the PH domain. More of the SNPs residues (red balls) are located in the barrel structure of the beta-sheets than in α-helices. (b) Cartoon structure of the PH domain depicting those pathogenic SNPs residues highlighted according to their secondary structure: two residues occur in the helical structure (W124 and Q127); 18 residues are situated in the beta-strands (F10, L11, K12, K19, F25, K27, R28, L31, L32, T33, I56, I61, V64, Q103, L111, Y112, V113, S115). Most of the pathogenic SNPs are buried within the structure of the protein.



(c) (d)

Figure 5.2 (c) and (d) shows two angles of transparent secondary structures of PH domain highlighting the pathogenic SNPs residue (W124 and Q127) located in the helical structure.

Representations in **Figure 5.2** correlates with earlier studies by Neuvirth *et al.* (2004) that suggested β-strands to be more favored at protein-protein interfaces, and therefore residue variations at such regions will most likely lead to disease conditions as observed in XLA and other BTK-related dysfunctions. Also, the likelihood of pathogenic variations in β-sheets increases the risk of the resultant (domain) protein to be dysfunctional or unable to fold properly.

From the analysis of PPI prediction of the PH domain using the six predictors (PredUS, cons-PPISP, PPI-Pred, meta-PPSIP, SPPIDER and PIER), mapping of the predicted PPI residues to the structure of the protein was carried out based on the amount of predictors that predicted a particular residue to be in an interfacial region. A visualization of possible binding patches on the surface of the BTK PH domain protein is seen in figure 5.3.

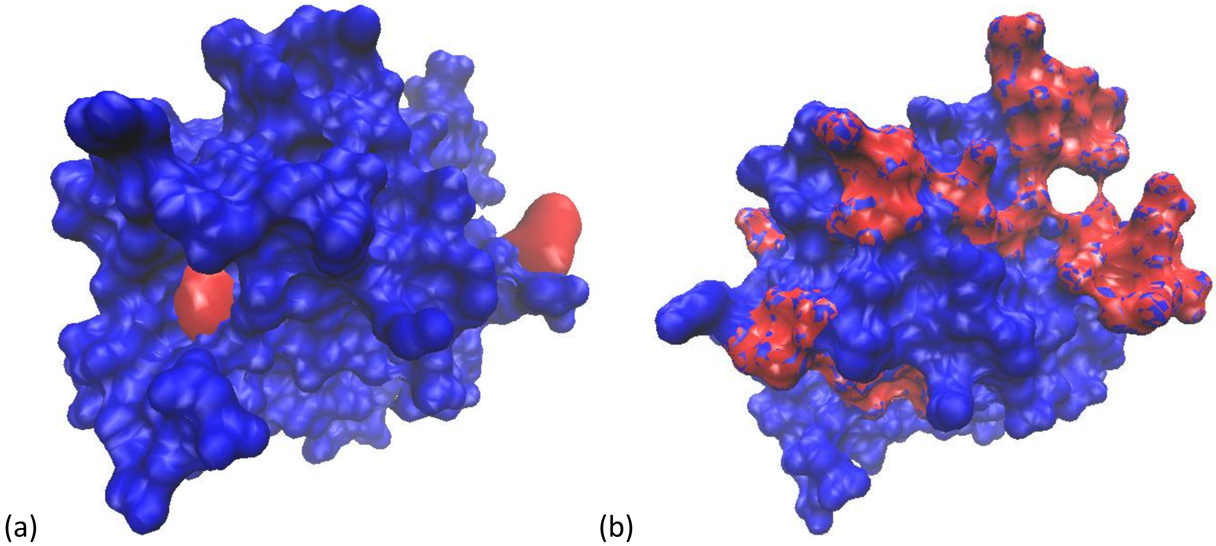


Figure 5.3 Surf representation showing possible binding patches mapped to the surface of the PH domain based on the predicted interface residues derived from the above analysis. The amount of prediction hit for each residue was ascertained and represented in the a color code as shown in **Table 5.1** (a) Two of the predictors highlighted just two residues to be an interface residue: C63 and Q91. (b) Here, predicted interface residues are depicted as binding patches on

the PH domain; only one out of any the six predictors predicted 39 residues to be in the interfacial region of the protein: A3, V4, I5, E7, S8, I9, Y42, D43, F44, E45, R46, G47, R48, K52, S55, E59, K60, T62, C63, V64, E65, T66, N72, M89, E90, Q91, I92, I94, I95, E96, V105, Y106, D107, E108, P110, I125, K129, N130, I132 and R133.

The amount of predictors highlighting a residue to be in the interfacial region, is directly proportional to the probability of that residue being in the interface region. Therefore, for PH domain of human BTK, residues C63 and Q91 (predicted by two of the six predictors as represented in **Figure 5.4a**) has the highest probability of being interface residues.

5.2 ANALYSIS OF SH3 DOMAIN

From the currently (August, 2012) available data set on XLA obtained from BTKbase (http://bioinf.uta.fi/BTKbase/) there were no entries for pathogenic SNPs existing in the SH3 domain of human BTK. Hopefully in the nearest future, on-going research efforts in pathogenic SNP occurring in human BTK would provide substantial scientific evidence pointing to the SH3 domain.

5.3 ANALYSIS OF SH2 DOMAIN

Examination of the pdb file 2GE9 for human BTK SH2 domain, which was obtained from the RCSB PDB (Berman et al.2000), showed stronger correlation between SNP residues and predicted interface residues. In accordance to the structure of SH2 domain (Huang *et al,* 2006) shown in **Figure 2.5b**, most of the predicted interfacial residues were situated in the β-strands due its abundance. **Table 5.4** shows the statistical distribution of the predicted interfacial residues in respect to the secondary structures of the protein.

Out of a total of the 97 residues in the pdb file of SH2 domain, 64 of them were predicted to be an interfacial residue by at least one or more of the six interface predictors used for the analysis. Residues W281, K284 and R288 all had a 67% probability percentage (i.e. they were predicted to be interface residues by four out of the six interface predictors) of being an interfacial residue, among which R288 has a known pathogenic SNP count of 40.

Table 5.3 The table shows a cross analysis of the prevalence of disease-causing SNPs in interface residues of human BTK SH2 domain. The "+" denotes those amino acid residues that were predicted to be in any interface region of the SH2 domain.

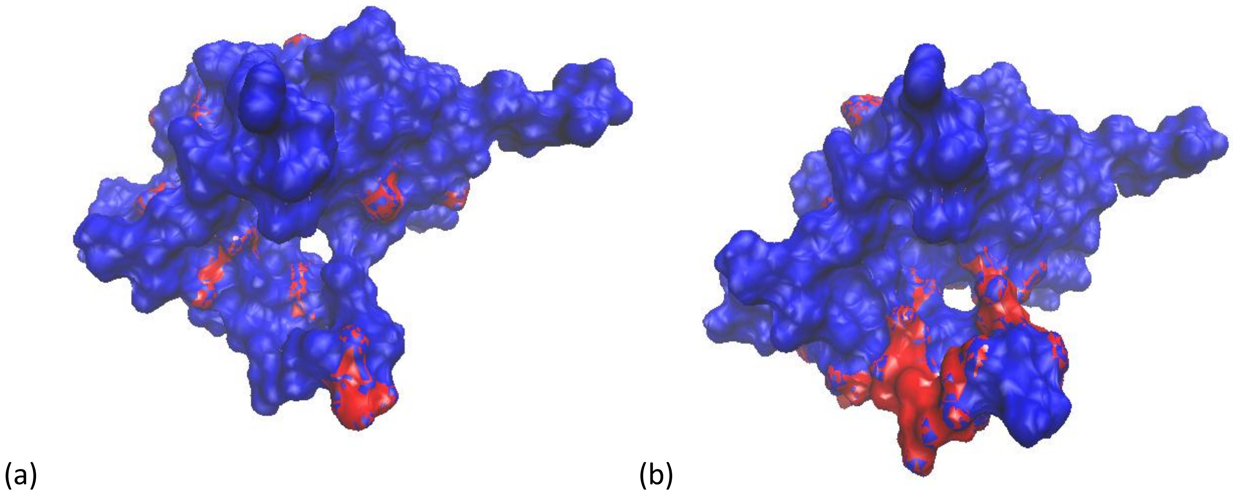
PDB SEQ POSITION	UNIPROT (WILD) POSITION	RESIDUE	SECONDARY STRUCTURE	PATHOGENIC SNP STATISTICS	INTERFACE PREDICTORS					
					consPPISP	PredUS	SPPIDER	Meta-PPISP (cons-PPISP, PINUP, Promate)	PPI-Pred	PIER
12	281	**W**	B-strand		+			+	+	+
13	282	**Y**						+	+	+
14	283	**S**			+			+	+	
15	284	**K**			+		+	+	+	
16	285	**H**			+			+		
17	286	**M**			+			+		
18	287	**T**			+					
19	288	**R**	Helix	40	+	+		+	+	
20	289	**S**				+				
21	290	**Q**								
22	291	**A**								
23	292	**E**				+			+	
24	293	**Q**								
25	294	**L**		1	+				+	
26	295	**L**		3					+	
27	296	**K**				+			+	
28	297	**Q**								
29	298	**E**								
30	299	**G**								
31	300	**K**							+	
32	301	**E**								
33	302	**G**		14	+				+	
34	303	**G**	B-strand		+			+	+	
35	304	**F**			+			+	+	
36	305	**I**			+				+	
37	306	**V**							+	
38	307	**R**		6	+		+	+		
39	308	**D**		1				+		
40	309	**S**						+		
41	310	**S**	B-strand							
42	311	**K**				+				
43	312	**A**								
44	313	**G**								
45	314	**K**								
46	315	**Y**	B-strand		+			+		
47	316	**T**		1	+			+	+	
48	317	**V**		1	+			+	+	
49	318	**S**		3	+			+	+	
50	319	**V**		1	+				+	+
51	320	**F**			+				+	
52	321	**A**							+	
53	322	**K**								
54	323	**S**								
55	324	**T**								
56	325	**G**								

57	326	**D**				+				
58	327	**P**								
59	328	**Q**								
60	329	**G**								
61	330	**V**	B-strand							
62	331	**I**				+				
63	332	**R**		1						
64	333	**H**		1						+
65	334	**Y**		2			+			+
66	335	**V**					+			+
67	336	**V**								
68	337	**C**	B-strand	1						
69	338	**S**								
70	339	**T**								
71	340	**P**	Turn							
72	341	**Q**								
73	342	**S**								
74	343	**Q**	B-strand							
75	344	**Y**					+			
76	345	**Y**				+				
77	346	**L**		2		+	+			+
78	347	**A**		1		+	+			+
79	348	**E**				+	+			
80	349	**K**				+	+			
81	350	**H**	B-strand			+	+			
82	351	**L**				+	+			
83	352	**F**					+			
84	353	**S**								
85	354	**T**								
86	355	**I**	Helix	3						
87	356	**P**								+
88	357	**E**				+				
89	358	**L**		3			+			
90	359	**I**					+			+
91	360	**N**					+			+
92	361	**Y**		4			+			+
93	362	**H**		2		+	+			+
94	363	**Q**				+	+			+
95	364	**H**		5			+			
96	365	**N**		1		+	+			
97	366	**S**		1		+	+			
98	367	**A**		2		+	+			
99	368	**G**				+	+			
100	369	**L**		2			+			
101	370	**I**		1			+			+
102	371	**S**		3			+			+
103	372	**R**		3		+	+			+
104	373	**L**	B-strand				+			+
105	374	**K**		3			+			+
106	375	**Y**								
107	376	**P**								+
108	377	**V**								

In comparison to the PH domain, structural differences in the amount of β-strands in SH2 domain appear to be higher than those in PH domain. Thus, this increases the likelihood of more interfacial residues being predicted in β-strands than α-helices for SH2 domain. The results from the analysis of the SH2 domain of the human BTK depicts strong statistical correlations between interfacial residues and occurrence of pathogenic SNP. In addition to β-strands being favored at interfacial regions (Neuvirth *et al.*, 2004), the amount of pathogenic SNPs found in β-strands- 41%, suggests that they constitute a functional site or "hot spots". These hot spots are known to be evolutionarily conserved and highly susceptible to variation (Chen and Zhou, 2005).

Table 5.4 showing the distribution of pathogenic SNPs and predicted interfacial residues in secondary structures in the SH2 domain of human BTK protein

Color Code	**Secondary structure**	**Amount of SNPs positions (Total = 29)**	**Amount of residues predicted to be in interface region (Total = 64)**
🟧	A-helices	7 (24.14%)	14 (%)
🟪	B-strands	12 (41.38%)	30 (%)
⬜	Turns/loops	0 (0.00%)	0 (0.00%)



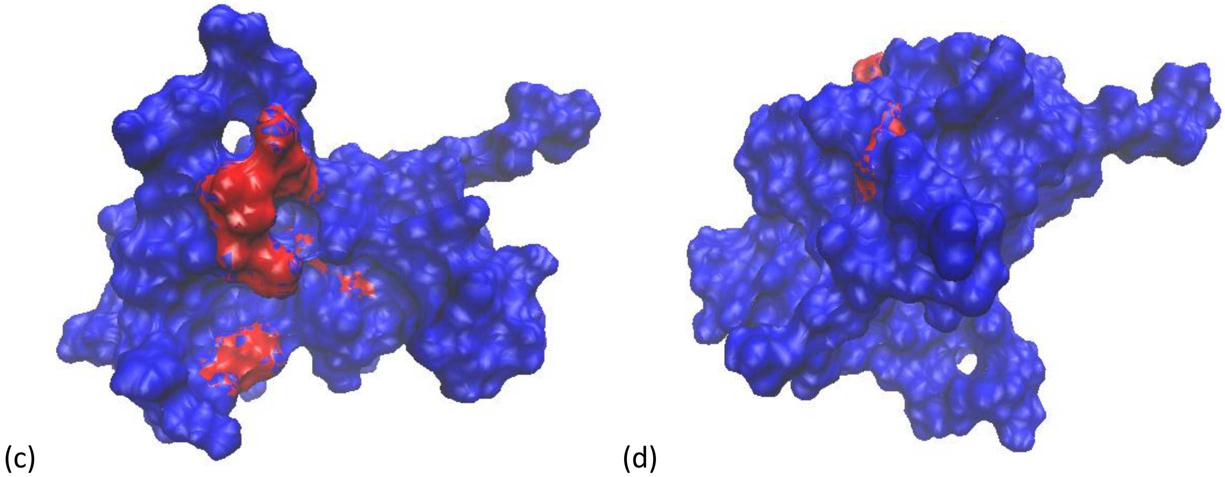


Figure 5.4 Structures of human BTK SH2 domain showing various binding patches formed by the sixty-four predicted interface residues based on the amount of web predictors. (a) Interfacial residues predicted by only one out of the six predictors: T287, S289, L295, K300, V306, D308, S309, K311, A321, D326, I331, H333, Y344, Y345, F352, P356, E357, L358, H365, L369 and P376. Of all these, six of them (**L295, D308, H333, L358, H364 and L369**) are known to contain pathogenic SNP that lead to XLA. (b) The highest amount of interface residues (26) that were predicted by only 2 out of the six predictors: H285, M286, E292, L294, K296, G302, I305, Y315, F320, Y334, V335, E348, K349, H350, L351, I359, N360, Y361, N365, S366, A367, G368, I370, S371, L373 and K374. Out of all these, **L294, G302, Y334, Y361, N365, S366, A367, I370, S371 and K374** were known to undergo pathogenic SNP that result into XLA.

(c) Interfacial regions formed by the residues predicted by any 3 out of the six interface predictors: Y282, S283, G303, F304, R307, T316, V317, S318, V319, L346, A347, H362, Q363 and R372. The interfacial region is compact and isolated to one part of the protein. More than half of the predicted interface residues are known to undergo pathogenic SNP that leads to XLA. They include: **R307, T316, V317, S318, V319, L346, A347, H362 and R372**. (d) Shows the location of residues W281, K284 and R288. Among the trio, R288 is the only residue known to undergo pathogenic SNP that leads to XLA; it has the highest solvent accessibility area as seen in **Figure 6.2** in the discussion.

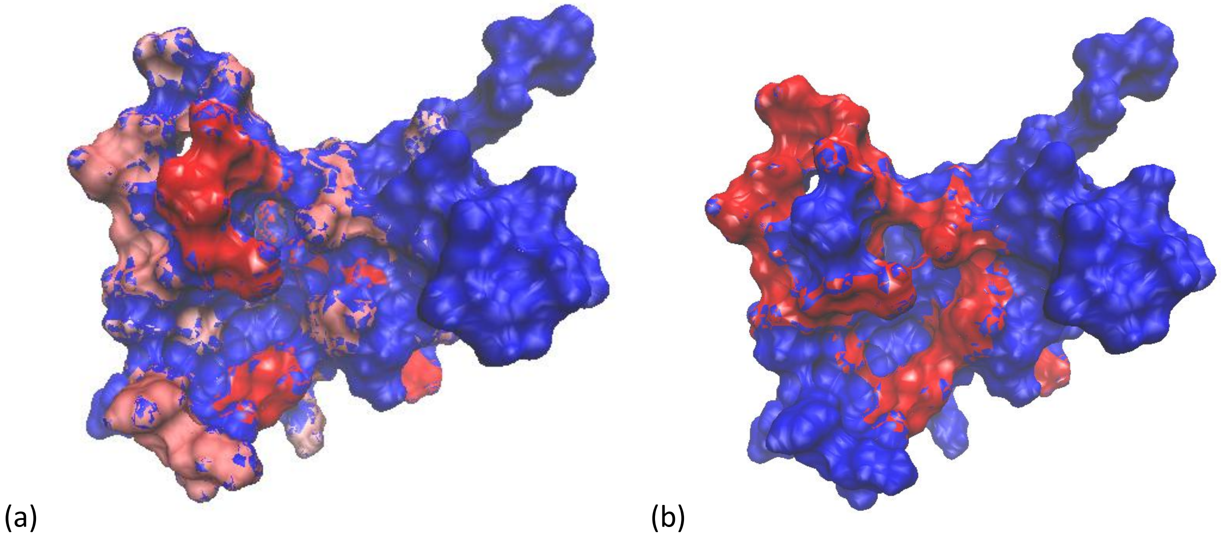


Figure 5.5 (a) Interfacial heat map showing the predicted interface regions/binding patches according to the probability percentages. The darker the shade, the higher the amount of protein interface predictors that highlighted the predicted residues. (b) Surface representation of all pathogenic SNPs predicted to occur in all interfacial residues of the SH2 domain. Overall, the human BTK SH2 domain appear to have multiple interacting regions mostly on the surface but also within the molecule, given the ratio of predicted interface residues to total number of residues in the protein.

5.4 ANALYSIS OF TK DOMAIN

For the analysis of the tyrosine kinase (TK) or kinase domain of BTK, the pdb file '3P08' with a resolution of 1.5Å was used an the input in all six protein interface predictors. The BTK kinase domain made of two chains with a residue length of about 254 aa. Thirty-two percent (83) of these 254aa residues were predicted to be in the interface region of the protein by any one out of the six protein interface predictors.

The highest probable (67%) interfacial residues, as seen in **Table 5.5**, were predicted by 4 out of the six predictors include: G409, T410, Q412, G414, V415 and Y617. 8 out of the 83 interface residues by predicted by 3 of the six predictors; another 19 of the 83aa interfacial residues were predicted by 2 of the six predictors while over 60% (50) of the 83 residues were predicted by just one out of the six interface predictors.

Cross analysis between the predicted interface residues and the residue positions of pathogenic variations in BTK kinase domain revealed that about 19% (16) of the predicted 83 interfacial residues undergo pathogenic SNP that is known to cause XLA. In comparison to PH and SH2 domains, about 1.6% and 38.8% of the predicted interfacial residues are known to undergo pathogenic variations that leads to XLA.

An amino acid of particular interest is arginine, which was one of the highest amino acids found in the BTK kinase interfacial regions (**Table 6.1**) as well as the amino acid residue having the highest amount of pathogenic variation in all domains analyzed. Detailed statistics about arginine and other amino acids have been highlighted in **Table 6.1** of the discussion section.

Table 5.5 A cross analysis of disease-causing SNPs in TK domain of BTK using selected protein interface predictors. The " +" denotes those residues that were predicted to be in the interface region.

PDB RESIDUE POSITION	UNIPROT (WILD) POSITION	RESIDUE	SECONDARY STRUCTURE	PATHOGENIC SNPS STATISTICS/POS	INTERFACE PREDICTORS					
					Consppisp	Predus	Sppider	Meta-ppisp (cons-ppisp, pinup, promate)	PPI-Pred n/a	PIER
10	402	**L**	B-strand	2						
11	403	**T**								
12	404	**F**				+	+			
13	405	**L**			+		+			
14	406	**K**			+		+			
15	407	**E**			+		+			
16	408	**L**		3	+		+			+
17	409	**G**			+		+	+		+
18	410	**T**			+		+	+		+
19	411	**G**	Turn				+	+		+
20	412	**Q**			+	+	+	+		+
21	413	**F**		1	+		+			+
22	414	**G**	B-strand	1	+		+	+		+
23	415	**V**			+		+	+		+
24	416	**V**			+		+			+
25	417	**K**			+		+			
26	418	**Y**		1			+			+
27	419	**G**		2						
28	420	**K**				+				
29	421	**W**				+				
30	422	**R**	Turn			+				
31	423	**G**								

32	424	**Q**								
33	425	**Y**	B-strand							
34	426	**D**								
35	427	**V**								
36	428	**A**					+			+
37	429	**I**		1	+		+			
38	430	**K**		6	+		+			
39	431	**M**			+		+			
40	432	**I**			+		+			
41	433	**K**			+		+			
42	434	**E**								
43	435	**G**					+			
44	436	**S**					+			
45	437	**M**								
46	438	**S**								
47	439	**E**	Helix							
48	440	**D**								
49	441	**E**								
50	442	**F**								
51	443	**I**								
52	444	**E**								
53	445	**E**		1						
54	446	**A**								
55	447	**K**								
56	448	**V**								
57	449	**M**								
58	450	**M**		1						
59	451	**N**				+				
60	452	**L**		2						
61	453	**S**				+				
62	454	**H**		2						
63	455	**E**				+				
64	456	**K**				+				
65	457	**L**								
66	458	**V**				+				
67	459	**Q**								
68	460	**L**	B-strand			+				+
69	461	**Y**								
70	462	**G**		4						
71	463	**V**								
72	464	**C**								
73	465	**T**								
74	466	**K**								
75	467	**Q**	B-strand							
76	468	**R**								
77	469	**P**								
78	470	**I**			+					
79	471	**F**								
80	472	**I**								
81	473	**I**								
82	474	**T**								
83	475	**E**								
84	476	**Y**		3						
85	477	**M**		1						+
86	478	**A**								
87	479	**N**						+		
88	480	**G**								+
89	481	**C**						+		+
90	482	**L**	Helix							
91	483	**L**				+		+		+
92	484	**N**				+		+		+
93	485	**Y**						+		
94	486	**L**		2						

95	487	**R**						+		+
96	488	**E**						+		
97	489	**M**	Helix					+		
98	490	**R**						+		
99	491	**H**								
100	492	**R**								
101	493	**F**								
102	494	**Q**								
103	495	**T**	Helix							
104	496	**Q**								
105	497	**Q**								
106	498	**L**								
107	499	**L**								
108	500	**E**				+				
109	501	**M**		1						
110	502	**C**		5						
111	503	**K**								
112	504	**D**		4						
113	505	**V**		1						
114	506	**C**		11						
115	507	**E**				+				
116	508	**A**		1						
117	509	**M**		12						
118	510	**E**				+				
119	511	**Y**		1						
120	512	**L**		5						
121	513	**E**								
122	514	**S**								
123	515	**K**				+				
124	516	**Q**								
125	517	**F**								
126	518	**L**		1						
127	519	**H**		2						
128	520	**R**		19			+			
129	521	**D**		4			+			
130	522	**L**		1						
131	523	**A**		2						
132	524	**A**	Helix	1						
133	525	**R**		18			+	+		
134	526	**N**		1						
135	527	**C**	B-strand	3						
136	528	**L**								+
137	529	**V**								
138	530	**N**								
139	531	**D**								
140	532	**Q**				+				
141	533	**G**								
142	534	**V**								
143	535	**V**	B-strand	3						
144	536	**K**								
145	537	**V**		1						
146	538	**S**		1						
147	539	**D**					+			
148	540	**F**		2						
149	541	**G**		2						
150	542	**L**	Helix	3						
151	543	**S**						+		
152	544	**R**		11			+			
153	545	**Y**								
154	546	**V**		1						
155	547	**L**								
156	548	**D**								

157	549	**D**	Helix	1						
158	550	**E**								
159	551	**Y**		1						
160	552	**T**								
161	553	**S**								
162	554	**S**								
163	555	**V**	Turn							
164	556	**G**								
165	557	**S**								
166	558	**K**								
167	559	**F**		1			+			
168	560	**P**						+		
169	561	**V**	Helix							
170	562	**R**		28				+		
171	563	**W**		1						
172	564	**S**								
173	565	**P**		2						
174	566	**P**	Helix	1						+
175	567	**E**		3						
176	568	**V**								
177	569	**L**		2						+
178	570	**M**								+
179	571	**Y**					+			
180	572	**S**								+
181	573	**K**								
182	574	**F**								
183	575	**S**		4						
184	576	**S**	Helix							
185	577	**K**		2						
186	578	**S**		2						
187	579	**D**		3						
188	580	**I**								
189	581	**W**		5						
190	582	**A**		8						
191	583	**F**		2						
192	584	**G**		7						
193	585	**V**		1						
194	586	**L**								
195	587	**M**		2						
196	588	**W**		2						
197	589	**E**		6						
198	590	**I**		1						
199	591	**Y**		2						
200	592	**S**	Turn	2						
201	593	**L**						+		
202	594	**G**		18						
203	595	**K**				+		+		
204	596	**M**								
205	597	**P**	Turn	2						
206	598	**Y**		4		+				
207	599	**E**				+	+	+		
208	600	**R**				+	+	+		
209	601	**F**						+		
210	602	**T**								
211	603	**N**	Helix				+			
212	604	**S**					+			
213	605	**E**								
214	606	**T**		2						
215	607	**A**		5			+			
216	608	**E**								
217	609	**H**				+				
218	610	**I**								
219	611	**A**								

220	612	**Q**		1						
221	613	**G**		7						
222	614	**L**				+				
223	615	**R**		3						
224	616	**L**		6						
225	617	**Y**				+	+	+		+
226	618	**R**					+	+		
227	619	**P**		6						
228	620	**H**					+	+		
229	621	**L**						+		
230	622	**A**		3						
231	623	**S**		3						
232	624	**E**	Helix							
233	625	**K**								
234	626	**V**		1						
235	627	**Y**					+			
236	628	**T**					+			
237	629	**I**								
238	630	**M**		6						
239	631	**Y**								
240	632	**S**								
241	633	**C**		1						
242	634	**W**		3						
243	635	**H**								
244	636	**E**								
245	637	**K**								
246	638	**A**	Helix							
247	639	**D**								
248	640	**E**								
249	641	**R**		15						
250	642	**P**		1						
251	643	**T**		1						
252	644	**F**	Helix	4						
253	645	**K**				+				
254	646	**I**								
255	647	**L**		4						
256	648	**L**		1						
257	649	**S**				+				
258	650	**N**								
259	651	**I**		2						
260	652	**L**		1						
261	653	**D**								
262	654	**V**								
263	655	**M**								

Table 5.6 Distribution of pathogenic SNPs and predicted interfacial residues in secondary structures of human BTK kinase domain

Color	**Secondary Structure**	**Amount of AA positions with SNP (Total = 91)**	**Amount of AA positions predicted as interfacial residues (Total = 83)**
🟧	A-helices	45 (49.45%)	19 (22.89%)
🟪	B-strands	10 (10.99%)	20 (24.10%)
⬛	Turns	5 (5.49%)	8 (9.64%)

From the analysis of kinase domain, more of the pathogenic SNPs were observed to occur in α-helices than β-strands; according to the results of the prediction in **Table 5.5**, β-strands were observed to be more dominant in interfacial regions.

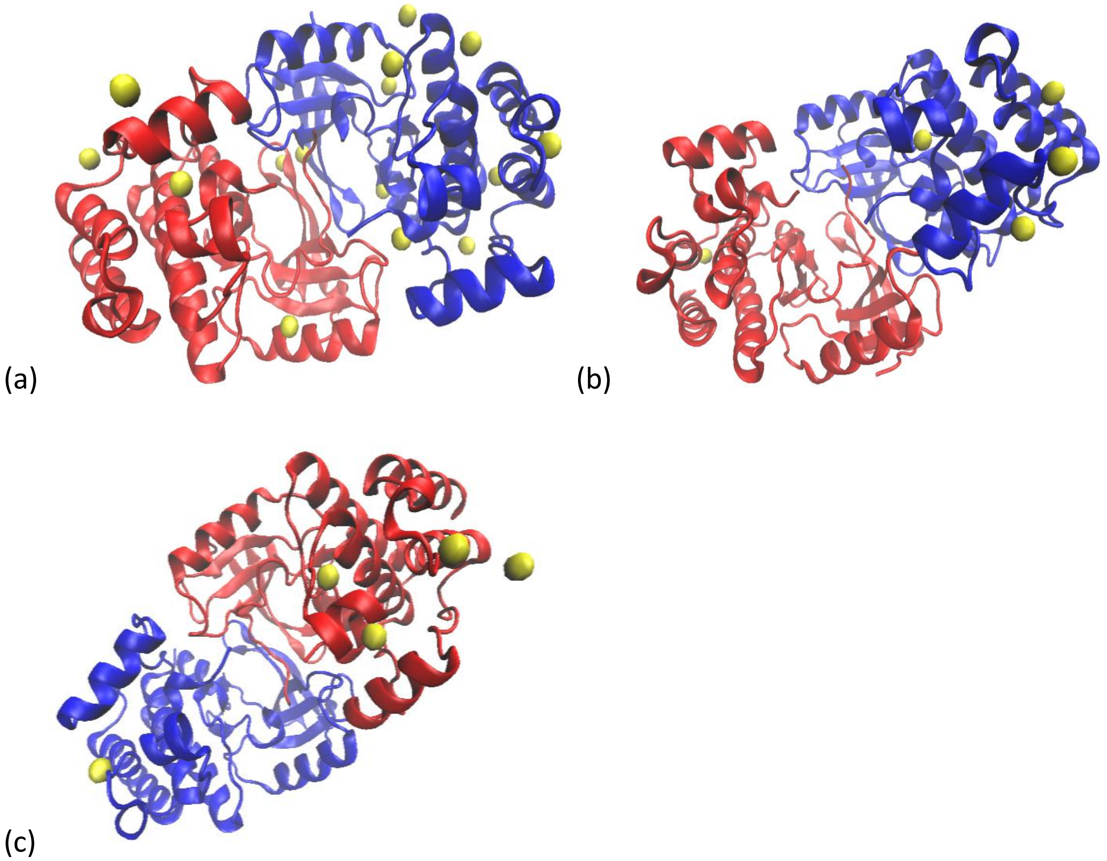


Figure 5.6 Structural representations of BTK kinase domain showing the predicted interfacial residues with probability percentage higher than 30% i.e. residues predicted by two or more protein interface predictors. (a) Cartoon representation of interfacial facial residues (F404, L405, K406, E407, K417, Y418, A428, I429, K430, M431, I432, K433, L460, C481, R487, R525, K595, R618 and H620) having a probability percentage of 33%. (b) Structural positions of all eight interfacial residues (L408, G411, F413, V416, L483, N484, E599 and R600) predicted by any 3 of the six interface predictors. Here, more interfacial residues (L408, G411, F413, V416, L483 and N484) are situated in chain A of the BTK kinase domain (c) Structural orientation of interface residues having a probability percentage of 67%. They include: G409, T410, Q412, G414, V415, Y617. Out of these, five of the interfacial residues (G409, T410, Q412, G414 and V415) are located in chain B of the kinase domain.

6 DISCUSSION AND CONCLUSION

6.1. Structural analysis and implication

6.1.1 PH Domain

In the analysis of PH domain, only one (V64) out of the 28 pathogenic SNPs residues was predicted to be situated in an interface region (see **Figure 6.1**). However, for the predicted interface residue, the amount of SNPs occurring in the β-strand structure is higher than those occurring in α-helices: 64.29% and 7.14% respectively (see **Table 5.3**). Therefore, since it has been suggested in earlier studies by Neuvirth *et al*. (2004) that interface regions have a preference for β-strands. This statistics regarding β-strands is in accordance to such studies.

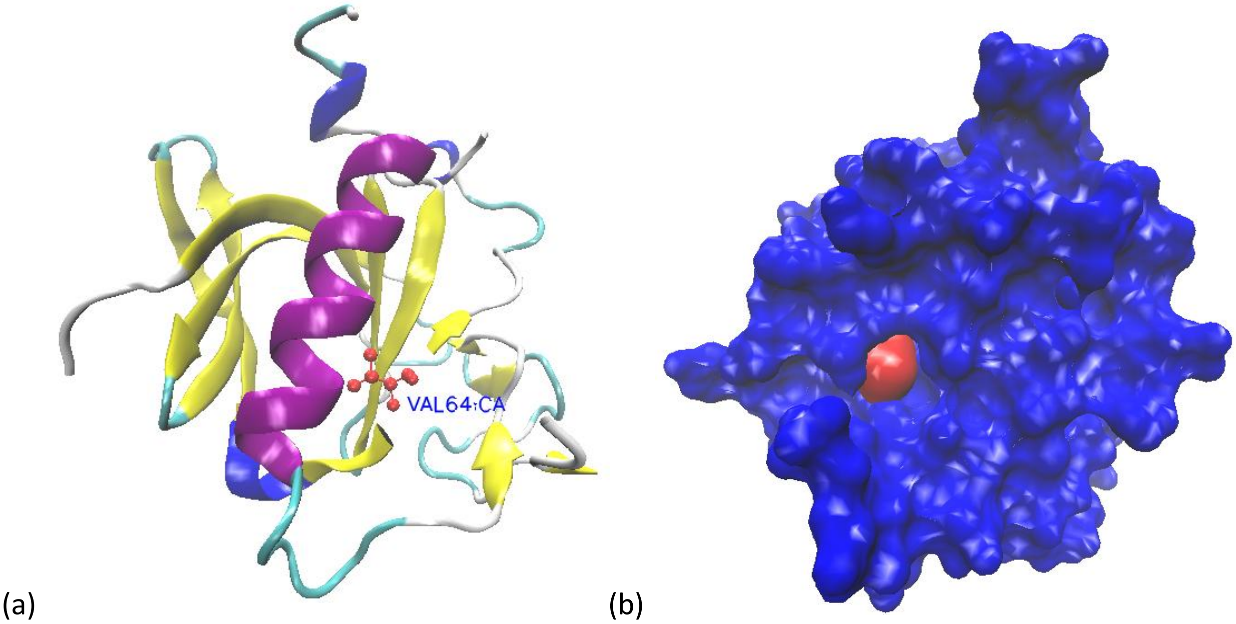


Figure 6.1 (a) Cartoon structure of BTK PH domain showing the location of the small hydrophobic valine residue located at position 67 (atom represented in red); it was discovered to be both a pathogenic SNPs residue and also an interface residue (V64). (b) Additionally, surface representation of the residue V64 (red patch) suggests that it has a relatively low accessibility.

Only the PIER protein interface predictor tool highlighted V64 as a likely interface residue. Kufareva *et al.* (2007) developed the protein-protein interface prediction tool- PIER based on local statistical properties of a protein surface derived at the level of atomic groups. This prediction web server functions by combining (i) statistically derived interatomic contact potentials, (ii) physical descriptors, such as observed solvent accessibility for separate atomic groups within amino acids, and (iii) sequence alignment based features, in particular, three different conservation scores (frequency-based, similarity matrix-based, and entropy-based) (Kufareva *et al.,* 2007).

The combination of the above methods was another reason PIER was chosen for this analysis. Porollo and Meller (2012) observed that the overall performance of PIER prediction tool is relatively dependent on the quality of pdb structure. Given 1BTK has a resolution of 1.60 Å from the PDB, we safely infer that the credibility of V64 being an interface residue is reliable.

The preference for secondary structures, such as β-strands, in interface regions may be due to the fact that β-strands are more flexible structures and are often able to form close contact across interfaces than α-helices (Neuvirth *et al.*, 2004). Variations occurring in such interfacial regions (known to be populated with β-strands) playing significant roles in biological processes is likely to account for certain observed phenotypic disease conditions as seen in the XLA.

As stated earlier were just one out of the 28 known pathogenic SNPs was predicted to be in interfacial region, this finding could suggest that most of the known pathogenic SNPs residues are embedded within the structure of the protein. For some of the predictors with a residue accessibility based algorithm, these predictors are unable to effectively analyze such residues and thereby producing false negative results. **Figure 5.2** shows the location of the pathogenic SNPs amongst which most are buried.

Shen and Vihinen (2004) suggested that the physiochemical properties of the amino acid residues found in the PH domain contain a substantial amount of hydrophobic residues. This implies that the high hydrophobicity of the residues buried in the protein are relatively involved in the folding of the functional quaternary BTK protein, thus, leading to the burying of most interfacial residues when the protein is unbound.

6.1.2 SH2 and TK Domain

Earlier studies have shown the conservation of two functionally important residues: R288 and R307, in the human BTK SH2 domain (Tzeng *et al.* 2000). From our analysis, **Figure 5.5d** shows the representation of those interface residues predicted by four out of the six predictors used. The residues W281, K284 and R288 form a cluster at the proximal end of the SH2 domain, thus, it suggests the likelihood of an interfacial region. Although no known SNP occur in W281 and K284, however high amount of SNP occurring at R288 further supports the likelihood of an interfacial region present at the proximal end of the human BTK SH2 domain.

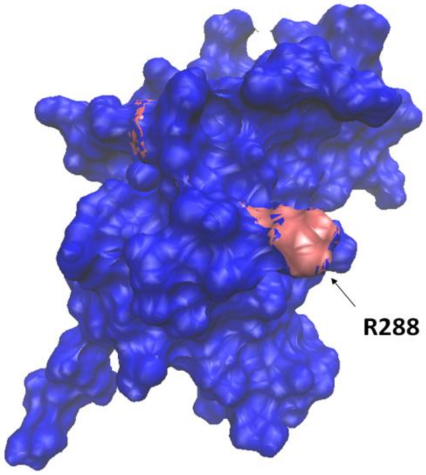


Figure 6.2 Surface representation of SH2 domain showing the accessibility of R288 (in red) to the surface of the protein. Although four out of the 6 interface predictors highlighted W281 and K284 as interface residues, they are however barely accessible to the surface of the protein. It is possible that residues W281 and K284 have intra-domain binding activity within the SH2 domain; they could as well be crucial residues of buried interfacial regions that are only functional when the conformation of the SH2 domain changes in accordance to specific binding partners.

In addition, Pekka *et al.*, (2000) highlighted six key residues that facilitates the phosphotyrosine-binding property of BTK SH2 domain in which pathogenic variations lead to XLA. These residues include: G302, R307, L358, Y361, H362 and I370. Similarly, results from interfacial residue prediction also correlated with all the above listed residues as seen in **Table 5.3**.

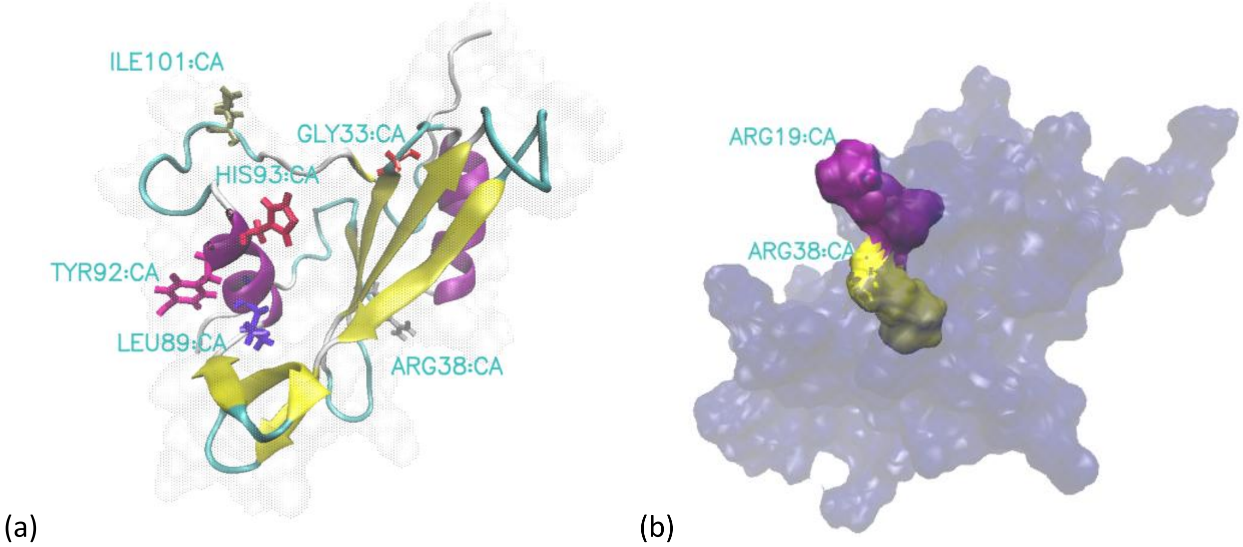


Figure 6.3 (a) Structural representation of predicted interfacial residues G302, R307, L358, Y361, H362 and I370 in BTK SH2 domain that are known to possess ligand-biding specificity to phosphotyrosine containing residues such as BLNK (Hashimoto *et al*. 1999). Given that pathogenic variations in functionally important residues ultimately lead to disease phenotypes (Pekka *et al.*,2000), inferably, the above listed residues are key components of interfacial regions/binding pockets in the SH2 domain. (b) Predicted (BTK SH2 domain) interfacial arginine residues R288 and R307 whose peptide-binding affinity in the phosphotyrosine binding pocket decreases by 200-fold in cases of XLA (Pekka *et al.*, 2000). As at the time of this analysis, R288 and R307 have a total pathogenic variation count of 40 and 6 respectively.

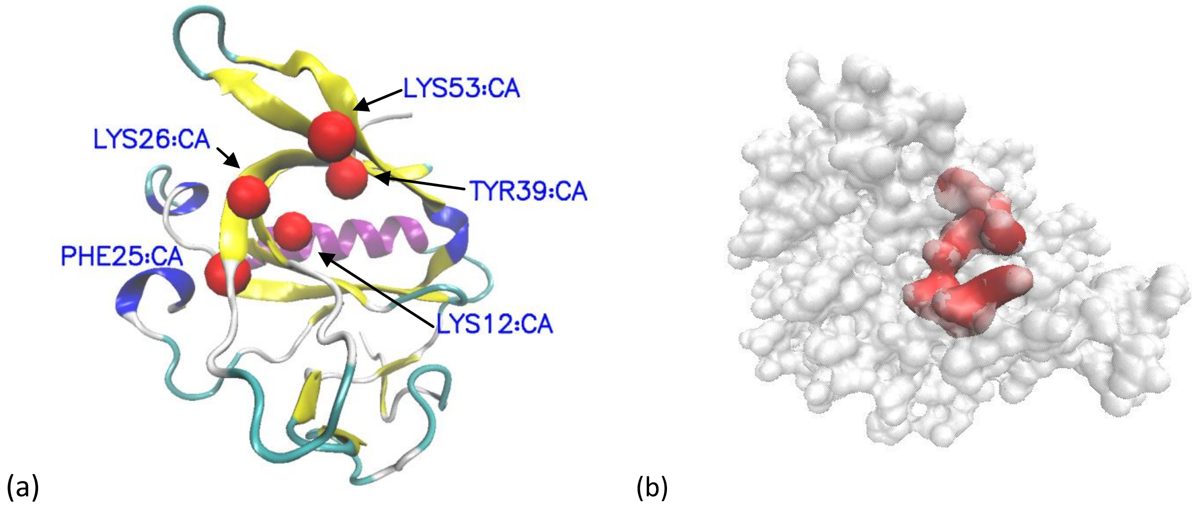
6.2. Effect on binding activity

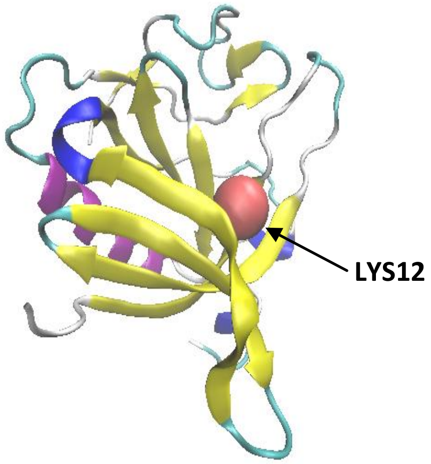
Primarily, PH domains are involved in protein–protein interactions or similar protein-lipid interactions and consequently these interactions or cellular localizations control the regulation of specific enzyme functions, as described by Musacchio *et al.* (1993); Lemmon *et* al. (1996) and Li *et al*. (1997). Amongst these interacting molecules, phospatidylinositol 3,4,5-trisphosphate (PIP3) is a known molecule whose interaction with the PH domain affects the overall activity of the BTK protein (Fakuda *et al.*, 1996).

Through computer modeling, observed residues that are essential for the binding of PIP3 to the PH domain are: K12, F25, R28 respectively (Ferguson *et al.,* 1995; Fakuda *et al*., 1996; Salim *et al*., 1996; Rameh *et al*. 1997). In addition, annotated binding sites from the Uniprot entry (Id-Q06187) also included R28. Other binding sites from Uniprot sequence annotation are: K26, Y39 and K53.

From our analysis, **Figure 6.1** shows the location of the only predicted interface residue- V64, that is also known to have pathogenic SNPs (V64D and V64F). Structural analysis by Hyvönen and Saraste (1997) showed that single nucleotide polymorphisms in residue V64 affect the stability and/or folding the PH domain because V64 is part of the hydrophobic core and is fully conserved in all proteins containing both the PH domain and BTK motif.

Analytic cross-referencing of the prediction results with existing literature earlier discussed revealed the insufficient ability of all six predictors to correctly predict the binding sites in the PH domain, except for the residue Lys12. Other residues F25, K26, R28, Y39 and K53 were not identified to be in a protein-protein interface region. On the other hand, the residues K12, F25, R28 and Y39 are implicated as pathogenic SNPs in the BTK protein (Li *et al.*, 1995; Vihinen *et al.*, 1995).





(c)

Figure 6.4 (a) Location of all residues (K12,F25,K26,R28,Y39 & K53) associated with PIP3 binding as suggested by previous literature (Ferguson *et al.,* 1995; Fakuda *et al*., 1996; Salim *et al*., 1996; Rameh *et al*. 1997). All residues are located in the β-barrel structure of the PH domain. (b) Surface representation of all six residues (from the previous literature) appear to form a "pocket"-like protein-protein interface with about half of the residues buried (c) Lysine12 (K12) was the only residue predicted by one of the six predictors.

Although the SNP residues had an uneven distribution, R28 was the only residue with a total of 40 pathogenic SNPs. R28 was not predicted to be an interface residue by any of the six predictors used during the analysis even though the R28 residue is situated in a β-strand as shown in the structure secondary structure organization of the PH domain from **Table 6.1.** More also, studies by Neuvirth *et al.* (2004) however suggests that a β-sheets are favored in interface regions. Thus, given a protein interface predictor with better accuracy of prediction, the extent of evolutionary conservation and mutagenic frequency of R28, it suggests that the residue-R28 constitutes a functional site in the PH domain of human BTK.

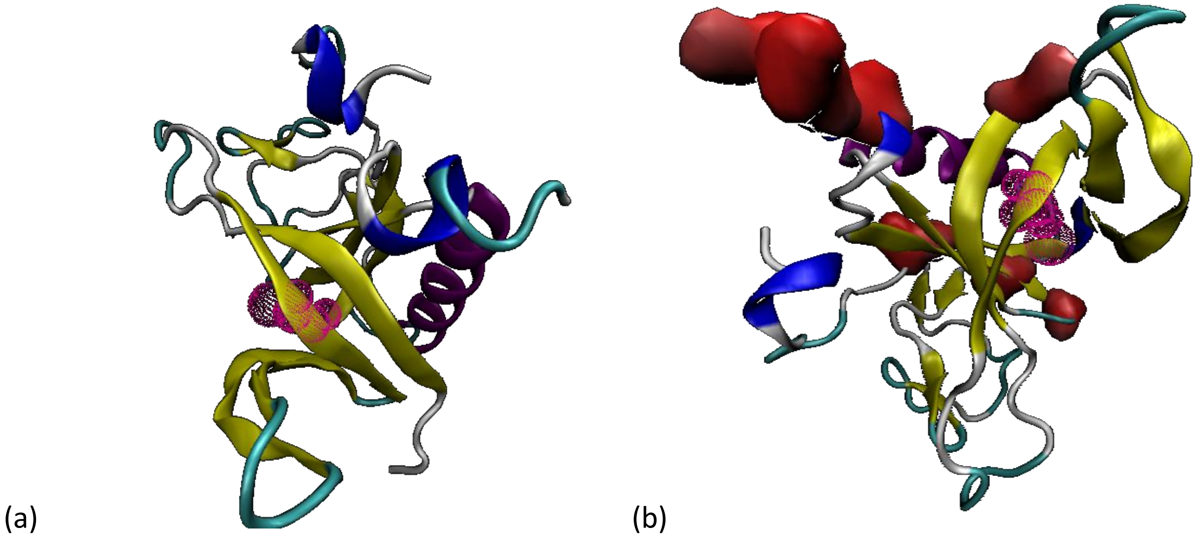


Figure 6.5 (a) The highly polymorphic R28 residue (dotted pink) is located on a β-strand buried within the structure of the PH domain protein. This might account for the inability of any of the predictors to be predict R28 as a probable protein-protein interface (b) The R28 residue is shown alongside the resultant predicted interface residues (red) from this analysis.

Also, in line with an earlier study by Vihinen *et al.* (1998), single nucleotide variations in R28 is associated to XLA in humans; given that β-sheets are more likely to be involved in interfacial regions. Likewise, one can infer that pathogenic variations in the β-strand structure where residue R28 is situated supports previous findings that associates single nucleotide variations in R28 as a genetic cause of XLA in humans (Vihinen *et al.*, 1998) and in experiments carried out in mice (Thomas *et al.*,1993; Rawlings *et al.*,1993).

6.3 Amino acid distribution across interfacial regions of BTK PH, SH2 and Kinase domains

Statistical analysis of all predicted interface residues, irrespective of the number of interface predictors, was carried out to investigate the presence of any correlation with those residues with pathogenic variations (SNP). Also, this was done to determine the possibility that certain amino acids are more favored than others in the interface regions of human BTK.

Table 6.1 Domain-wise distribution of interfacial and SNP residues in BTK domains: PH, SH2 and TK

Amino Acid Residue	**PH (170aa) Interface (SNP)**	**SH2 (97aa) Interface (SNP)**	**Tyrosine Kinase Domain (267aa) Interface (SNP)**	**Total Interface (SNP)**	Approximate Ratio Interface : SNP
A	1 (0)	3 (3)	2 (20)	6 (23)	1 : 4
*R	3 (42)	3 (53)	9 (110)	15 (205)	1 : 14
N	2 (0)	2 (1)	4 (1)	8 (2)	1 : 0.3
D	2 (0)	2 (1)	2 (13)	6 (14)	1 : 2.3
C	1 (0)	0 (1)	1 (20)	2 (21)	1 : 11
Q	1 (2)	1 (0)	2 (1)	4 (3)	1 : 0.8
E	7 (0)	3 (0)	7 (10)	17 (10)	1 : 0.6
*G	1 (0)	3 (8)	5 (45)	9 (53)	1 : 6
H	0 (0)	5 (8)	2 (4)	7 (12)	1 : 1.7
I	7 (7)	4 (4)	3 (4)	14 (15)	1 : 1.1
L	0 (16)	7 (12)	9 (33)	16 (61)	1 : 3.8
K	3 (5)	6 (3)	9 (8)	18 (16)	1 : 0.9
M	1 (2)	1 (0)	4 (21)	6 (23)	1 : 3.8
F	1 (4)	3 (0)	4 (10)	8 (14)	1 : 1.8
P	1 (0)	2 (0)	1 (13)	4 (13)	1 : 3.3
*W	0 (3)	1 (0)	1 (12)	2 (15)	1 : 8
Y	2 (13)	6 (6)	6 (12)	14 (31)	1 : 2.2
V	3 (5)	4 (2)	3 (8)	10 (15)	1 : 1.5
S	2 (5)	5 (7)	7 (14)	14 (26)	1 : 1.9
T	2 (9)	3 (1)	2 (3)	7 (13)	1 : 1.9
TOTAL	**40 (113)**	**64 (110)**	**83 (362)**		

From the statistical analysis of all predicted interface residues across the human PH, SH2 and TK domains, the net amino acid distribution was calculated independent of the residue positions. Results from the interface-to-SNP amino acid ratio in **Table 6.1** was considered domain-wise rather than generalizing the amino acid distribution across the entire molecule given that only PH, SH2 and TK domains were analyzed- they are not the only interfacial regions on the BTK molecule.

The BTK domains individually possess peculiar properties that capacitate them to interact with various signaling molecule, thus, enabling the molecule as a whole to carry out various biological processes. Qin and Chock (2001) demonstrated that pathogenic variations in any of these BTK domains: PH (R28C), SH2(R307A) or TK(R525Q), would reduce or totally inhibit the function of the whole BTK molecule.

Statistical analysis shows that lysine was the most abundant interfacial amino acid residue in all three domains (PH, SH2 and TK). **Figure 2.2** shows the schematic property of lysine: positively charged, polar, ability to form salt-bridges, amphipathic etc. Of interfacial importance from all these properties is the amphipathicity of lysin i.e. it allows hydrophobic interactions in the side chain close to the backbone while the terminal side chain remains polar. This unique feature enables part of the side chain to be buried within the protein and the remaining part on the surface of the protein- making it an optimal interfacial residue.

Glycine, tryptophan and arginine had an interface-to-SNP ratio of 1:6, 1:8 and 1:14 respectively. Even though arginine was the fourth most abundant residue across all interface regions of the three domains, however, it had the highest interface-to-SNP ratio. Cross referencing this result with the studies by Qin and Chock (2001) on implications of pathogenic variations on key arginine residues in BTK, results in **Table 6.1** suggests that arginine residues appear to fairly abundant in interface regions and they also constitute functional sites of the BTK PH, SH2 and TK domains.

Since evolutionary conserved residues: in BTK domains: PH (R28C), SH2(R307A) and TK(R525Q), are usually key residues of functional importance, therefore, the count of pathogenic variations of arginine residues across all three domains as seen in **Table 6.1** also strengthens the result that arginine is abundant at interfacial regions of the protein due to its function.

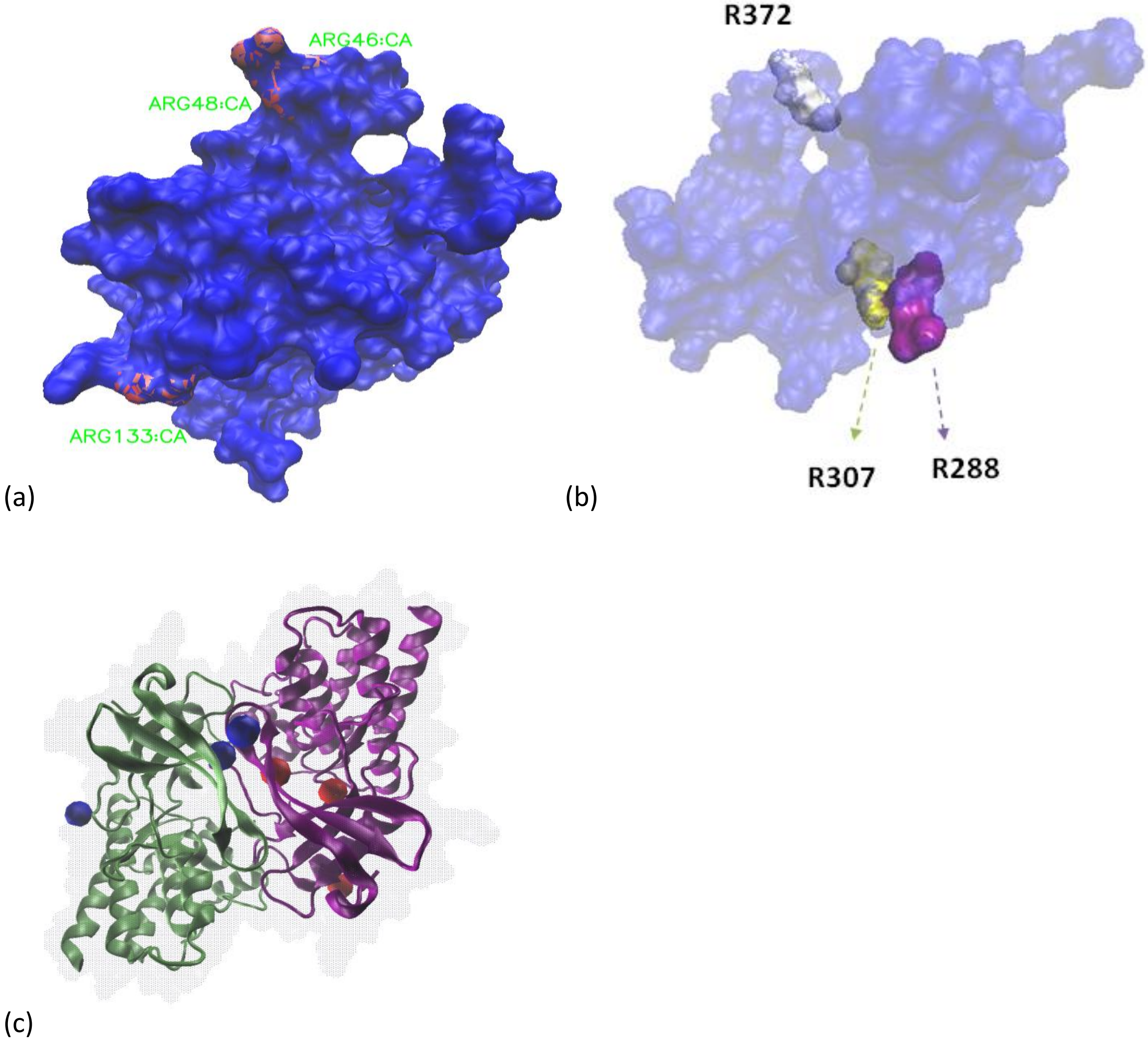


Figure 6.6 Structural representations of predicted interfacial arginine residues in BTK PH, SH2 and TK domains. (a) Surface representation of all three interfacial arginine residues R46, R48 and R133 in BTK PH domain. All residues are situated at the surface of the protein, residues R46 and R48 appears to be a constitute of binding site. (b) Interfacial arginine residues R288, R307 and R372 are shown in BTK SH2 domain. Here also, known phosphotyrosine interacting residues R288 and R307 appear as constituent residues of a bind site. (c) Dimeric structure of

BTK kinase domain showing the location of arginine residues. Most of the arginine residues are not easily accessible to the surface of the protein.

Adequate efforts and resources should be channeled into research aimed at illuminating the intricacies behind interactions occurring at interface regions of proteins. In the nearest future, the emergence of better interface predictors will greatly improve the accuracy of protein-protein interfacial region prediction. Furthermore, as more refined and precise methods of gene sequencing emerging, more evidence will emerge on the various types of pathogenic variations that occur in β-strand structure that are present in protein interfaces, both in homogenous or heterogeneous complexes- transient or obligatory.

Knowledge of interfacial regions on proteins including binding sites and docking regions are pivotal in the design of effective therapeutic agents in pharmaceutical industries as well as medicine. Adequate knowledge on specific interfacial residues that undergo disease-causing variations can elucidate which of the interfacial amino acid residues should be the target of an interface-focused gene therapy.

7 REFERENCES

Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics. 2004 Mar 1;20(4):477-86.

Anfinsen CB, Edsall JT, Richards FM (1972). Advances in Protein Chemistry. New York: Academic Press. pp. 99, 103.

Armon A, Graur D, Ben-Tal N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 16; 307(1):447-63.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000). The Protein Data Bank Nucleic Acids Research, 28: 235-242.

Betts MJ, Russell R.B. (2003). Amino acid properties and consequences of substitutions. In Bioinformatics for Geneticists, M.R. Barnes, I.C. Gray eds, Wiley.

Bibbins KB, Boeuf H, Varmus HE. Binding of the Src SH2 domain to phosphopeptides is determined by residues in both the SH2 domain and the phosphopeptides. Mol Cell Biol. 1993 Dec;13(12):7278-87.

Bouralexis S, Findlay DM, Atkins GJ, Labrinidis A, Hay S, Evdokiou A. Progressive resistance of BTK-143 osteosarcoma cells to Apo2L/TRAIL-inducedapoptosis is mediated by acquisition of DcR2/TRAIL-R4 expression: resensitisation with chemotherapy. Br J Cancer. 2003 Jul 7;89(1):206-14.

Bradshaw JM. The Src, Syk, and Tec family kinases: distinct types of molecular switches. Cell Signal. 2010. Aug;22(8):1175-84.

Bradford JR, Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics. 2005 Apr 15;21(8):1487-94.

Branden, Carl, and John Tooze. (1997). Introduction to Protein Structure. New York and London: Garland Publishing, Inc.

Brautigam CA, Wynn RM, Chuang JL, Machius M, Tomchick DR, Chuang DT. Structural insight into interactions between dihydrolipoamide dehydrogenase (E3) and E3 binding protein of human pyruvate dehydrogenase complex. Structure. 2006 Mar;14(3):611-21.

Brooker, Robert J., Widmaier, Eric P., Graham, Linda E., and Stiling, Peter D. (2008).Biology. New York: McGraw-Hill Pub.

Bross P, Corydon TJ, Andresen BS, Jorgensen MM, Bolund L, Gregersen N. (1999). Protein misfolding and degradation in genetic diseases. Hum Mutat 14:186–198.

Bruton, 0C. (1952). Agammaglobulinemia. Pediatrics 9, 722-728.

Campana, D., Farrant, J., Inamdar, N., Webster, A. D. B. & Janossy, G. (1990). Phenotypic features and proliferative activity of B cell progenitors in X-linked agammaglobulinemia. J. Immunol. 145, 1675-1680.

Chen H, Zhou HX. (2005). Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. Proteins. Oct 1;61(1):21-35.

Cheng G, Qian B, Samudrala R, Baker D. (2005). Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res;33(18):5861-7.

Cheng G, Ye ZS, Baltimore D. Binding of Bruton's tyrosine kinase to Fyn, Lyn, or Hck through a Src homology 3 domain-mediated interaction. Proc Natl Acad Sci US A. 1994 Aug 16;91(17):8152-5.

Conley ME, Brown P, Pickard AR, Buckley RH, Miller DS, Raskind WH. (1986). Expression of the gene defect in X-linked agammaglobulinemia. *N Engl J Med*.;315:564–567.

Danial NN, Gramm CF, Scorrano L, Zhang CY, Krauss S, Ranger AM, Datta SR, Greenberg ME, Licklider LJ, Lowell BB, Gygi SP, Korsmeyer SJ. BAD and glucokinase reside in a mitochondrial complex that integrates glycolysis and apoptosis. Nature. 2003 Aug 21;424(6951):952-6.

de Gorter DJ, Reijmers RM, Beuling EA, Naber HP, Kuil A, Kersten MJ, Pals ST, Spaargaren M. The small GTPase Ral mediates SDF-1-induced migration of B cells and multiple myeloma cells. Blood. 2008 Apr 1;111(7):3364-72.

DeLano WL. (2002). Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 12(1):14-20.

Elcock AH.(2001). Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol. 28;312(4):885-96.

Ferrer-Costa C, Orozco M, de la Cruz X. (2002). Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 315:771–786.

Ferguson KM, Lemmon MA, Schlessinger J, Sigler PB. (1995). Structure of the high affinity complex of inositol trisphosphate with a phospholipase C pleckstrin homology domain. Cell. Dec 15;83(6):1037-46.

Ferguson KM, Kavran JM, Sankaran VG, Fournier E, Isakoff SJ, Skolnik EY, Lemmon MA. (2000). Structural basis for discrimination of 3-phosphoinositides by pleckstrin homology domains. Mol Cell. 2000 Aug;6(2):373-84.

Fisher M, Zhang QC, Deng L, Dey F, Chen BY, Honig B, Petrey D. (2011). MarkUs: a server to navigate sequence–structure–function space. Nucleic Acids Res., 39:W357-W361.

Fuentes G, Oyarzabal J, Rojas AM. (2009). Databases of protein-protein interactions and their use in drug discovery. Curr Opin Drug Discov Devel. 2009 May;12(3):358-66.

Fukuda M, Kojima T, Kabayama H, Mikoshiba K. Mutation of the pleckstrin homology domain of Bruton's tyrosine kinase in immunodeficiency impaired inositol 1,3,4,5-tetrakisphosphate binding capacity. J Biol Chem. 1996 Nov 29;271(48):30303-6.

Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 1;43(2):89-102.

Golemis E. (2002) Protein-protein interactions : A molecular cloning manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. ix, 682 p.

Grosdidier S, Totrov M, Fernández-Recio J. Computer applications for prediction of protein-protein interactions and rational drug design. Adv Appl Bioinform Chem. 2009;2:101-23.

Guinamard R, Fougereau M, Seckinger P. The SH3 domain of Bruton's tyrosine kinase interacts with Vav, Sam68 and EWS. Scand J Immunol. 1997 Jun;45(6):587-95.

Güldener U, Münsterkötter M, Kastenmüller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, García-Martínez J, Pérez-Ortín JE, Michael H, Kaps A, Talla E, Dujon B, André B, Souciet JL, De Montigny J, Bon E, Gaillardin C, Mewes HW. CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D364-8.

Gushchina LV, Gabdulkhakov AG, Nikonov SV, Filimonov VV. High-resolution crystal structure of spectrin SH3 domain fused with a proline-rich peptide. J Biomol Struct Dyn. 2011 Dec;29(3):485-95.

Gustafsson MO, Hussain A, Mohammad DK, Mohamed AJ, Nguyen V, Metalnikov P, Colwill K, Pawson T, Smith CI, Nore BF. Regulation of nucleocytoplasmic shuttling of Bruton's tyrosine kinase (Btk) through a novel SH3-dependent interaction with ankyrin repeat domain 54 (ANKRD54). Mol Cell Biol. 2012 Jul;32(13):2440-53.

http://www.yellowtang.org/images/levels_of_protein_s_c_la_784.jpg : Moodle page on Yellow Tang. College of Siskiyous (Chemistry unit) Biology department. 2012. Accessed 22nd September, 2012.

Hansson H, Mattsson PT, Allard P, Haapaniemi P, Vihinen M, Smith CI, Hard T. Solution structure of the SH3 domain from Bruton's tyrosine kinase. Biochemistry. 1998 Mar 3;37(9):2912-24.

Hashimoto S, Iwamatsu A, Ishiai M, Okawa K, Yamadori T, Matsushita M, Baba Y, Kishimoto T, Kurosaki T, Tsukada S. Identification of the SH2 domain binding protein of Bruton's tyrosine kinase as BLNK--functional significance of Btk-SH2 domain in B-cell antigen receptor-coupled calcium signaling. Blood. 1999 Oct 1;94(7):2357-64.

Haslam RJ, Koide HB, Hemmings BA. Pleckstrin domain homology. Nature. 1993 May 27;363(6427):309-10.

Hosur R, Xu J, Bienkowska J, Berger B. iWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions. J Mol Biol. 2011 Feb 4;405(5):1295-310.

Hu Z, Ma B, Wolfson H, Nussinov R. (2000). Conservation of polar residues as hot spots at protein interfaces. Proteins 1;39(4):331-42.

Huang KC, Cheng HT, Pai MT, Tzeng SR, Cheng JW. Solution structure and phosphopeptide binding of the SH2 domain from the human Bruton's tyrosine kinase. J Biomol NMR. 2006 Sep;36(1):73-8.

Hubbard TJP, Aken BL, Beal1 K, Ballester1 B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Overduin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez KM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A and Birney E. (2007). Ensembl. *Nucleic Acids Res*. 2007 Vol. 35, Database issue:D610-D617.

Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38. http://www.ks.uiuc.edu/Research/vmd/

Hyvönen M, Macias MJ, Nilges M, Oschkinat H, Saraste M, Wilmanns M. (1995). Structure of the binding site for inositol phosphates in a PH domain. EMBO J.2;14(19):4676-85.

Hyvönen M and Saraste M. (1997). Structure of the PH domain and Btk motif from Bruton's tyrosine kinase: molecular explanations for X-linked agammaglobulinaemia. The EMBO Journal; (16): 3396 - 3404.

Il'ina VL, Korogodin VI, Fajszi C. [Relation between the frequency of various types of reverse mutations in yeasts auxotrophic for adenine and the adenine content of the medium]. Genetika. 1987 Apr;23(4):637-42.

Jackson S, Sugiman-Marangos S, Cheung K, Junop M. Crystallization and preliminary diffraction analysis of truncated human pleckstrin. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2011 Mar 1;67(Pt 3):412-6.

Jin R, Kaneko H, Suzuki H, Arai T, Teramoto T, Fukao T, Kondo N. Age-related changes in BAFF and APRIL profiles and upregulation of BAFF and APRIL expression in patients with primary antibody deficiency. Int J Mol Med. 2008 Feb;21(2):233-8.

John Toon. (2010) Protein-Protein Interfaces: Researchers Use Artificial Proteins to Understand Interactions Key to Cellular Processes. Georgia Tech: Research News. Research News & Publication Office, Georgia Institute of Technology; Atlanta, Georgia (30308, USA).

Jones S, Thornton JM. (1996). Principles of protein-protein interactions. Proc Natl Acad Sci U S A. 9;93(1):13-20.

Jones, S and Thornton, J. M. (1997). Prediction of protein-protein interaction sites using patch analysis. J. Mol. Biol. 272: 133-143.

Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V. Predicting protein-protein interface residues using local surface structural similarity. BMC Bioinformatics. 2012 Mar 18;13:41.

Joseph RE, Severin A, Min L, Fulton DB, Andreotti AH. SH2-dependent autophosphorylation within the Tec family kinase Itk. J Mol Biol. 2009 Aug 7;391(1):164-77.

Kang SW, Wahl MI, Chu J, Kitaura J, Kawakami Y, Kato RM, Tabuchi R, Tarakhovsky A, Kawakami T, Turck CW, Witte ON, Rawlings DJ. (2001). PKC beta modulates antigen receptor signaling via regulation of Btk membrane localization. EMBO J. 15;20(20):5692-702.

Keskin O, Nussinov R. (2005). Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. Protein Eng Des Sel ;18(1):11-24.

Khan S. and Vihinen M. (2010). Performance of protein stability predictors. Hum Mutat; 31(6): 675-84.

Kleanthous, C. (2000). Protein-Protein Recognition: Frontiers in Molecular Biology, Oxford University Press, Oxford, UK.

Kneidinger M, Schmidt U, Rix U, Gleixner KV, Vales A, Baumgartner C, Lupinek C, Weghofer M, Bennett KL, Herrmann H, Schebesta A, Thomas WR, Vrtala S, Valenta R, Lee FY, Ellmeier W, Superti-Furga G, Valent P. The effects of dasatinib on IgE receptor-dependent activation and histamine release in human basophils. Blood. 2008 Mar 15;111(6):3097-107.

Korogodin VI, Korogodina VL, Fajszi C, Chepurnoy AI, Mikhova-Tsenova N, Simonyan NV. On the dependence of spontaneous mutation rates on the functional state of genes. Yeast. 1991 Feb;7(2):105-17.

Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R. PIER: protein interface recognition for structural proteomics. Proteins. 2007 May 1;67(2):400-17.

Landgraf R., Xenarios I., Eisenberg D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol. Apr 13;307(5):1487-502.

Lederman H. M., Winkelstein J. A. (1985). X-linked agammaglobulinemia: an analysis of 96 patients. Medicine (Baltimore);64:145–156.

Lemmon MA, Ferguson KM, Schlessinger J. PH domains: diverse sequences with a common fold recruit signaling molecules to the cell surface. Cell. 1996 May 31;85(5):621-4.

Lemmon MA. Pleckstrin homology (PH) domains and phosphoinositides. Biochem Soc Symp. 2007;(74):81-93.

Li B, Kihara D. Protein docking prediction using predicted protein-protein interface. BMC Bioinformatics. 2012 Jan 10;13:7.

Li T, Tsukada S, Satterthwaite A, Havlik MH, Park H, Takatsu K, Witte ON. Activation of Bruton's tyrosine kinase (BTK) by a point mutation in its pleckstrin homology (PH) domain. Immunity. 1995 May;2(5):451-60.

Li Z, Wahl MI, Eguinoa A, Stephens LR, Hawkins PT, Witte ON. Phosphatidylinositol 3-kinase-gamma activates Bruton's tyrosine kinase in concert with Src family kinases. Proc Natl Acad Sci U S A. 1997 Dec 9;94(25):13820-5.

Liang S, Zhang C, Liu S, Zhou Y. (2006). Protein binding site prediction using an empirical scoring function. Nucleic Acids Res. Aug 7;34(13):3698-707.

Lichtarge O, Sowa ME. (2002). Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol. Feb;12(1):21-7.

Lijnzaad P, Argos P. (1997). Hydrophobic patches on protein subunit interfaces: characteristics and prediction. Proteins. Jul;28(3):333-43.

Lindvall JM, Blomberg KE, Väliaho J, Vargas L, Heinonen JE, Berglöf A, Mohamed AJ, Nore BF, Vihinen M, Smith CI . (2005). Bruton's tyrosine kinase: cell biology, sequence conservation, mutation spectrum, siRNA modifications, and expression profiling. Immunol Rev. Feb;203:200-15.

Liu J, Fitzgerald ME, Berndt MC, Jackson CW, Gartner TK. Bruton tyrosine kinase is essential for botrocetin/VWF-induced signaling and GPIb-dependent thrombus formation in vivo. Blood. 2006 Oct 15;108(8):2596-603.

Lo Conte L, Chothia C, Janin J. (1999). The atomic structure of protein-protein recognition sites. J Mol Biol. Feb 5;285(5):2177-98.

Lougaris V, Ferrari S, Cattalini M, Soresina A, Plebani A. Autosomal recessive agammaglobulinemia: novel insights from mutations in Ig-beta. Curr Allergy Asthma Rep. 2008 Sep;8(5):404-8.

Luu, Dao; Rusu, Alin; Walter, Vincent; Linard, Benjamin; Poidevin, Laetitia; Ripp, Raymond; Moulinier, Luc; Muller, Jean; Raffelsberger, Wolfgang; Wicker, Nicolas; Lecompte, Odile; Thompson, Julie; Poch, Olivier; Nguyen, Hoan (2012). KD4v: Comprehensible Knowledge Discovery System For Missense Variant. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W71-5.

Ma B, Shatsky M, Wolfson HJ, Nussinov R. (2002). Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. Protein Sci. Feb;11(2):184-97.

Martincorena I, Seshasayee AS, Luscombe NM. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature. 2012 May3;485(7396):95-8.

Mattsson PT, Lappalainen I, Bäckesjö CM, Brockmann E, Laurén S, Vihinen M, Smith CI. Six X-linked agammaglobulinemia-causing missense mutations in the Src homology 2 domain of Bruton's tyrosine kinase: phosphotyrosine-binding and circular dichroism analysis. J Immunol. 2000 Apr 15;164(8):4170-7.

Mayer BJ, Ren R, Clark KL, Baltimore D. A putative modular domain present in diverse signaling proteins. Cell. 1993 May 21;73(4):629-30.

Mayer BJ, Baltimore D. Signalling through SH2 and SH3 domains. Trends Cell Biol. 1993 Jan;3(1):8-13.

Mayer BJ. SH3 domains: complexity in moderation. J Cell Sci. 2001 Apr;114(Pt7):1253-63.

McKinney RE, Katz SL, Wilfert CM. (1987). Chronic enteroviral meningoencephalitis in agammaglobulinemic patients. Rev Infect Dis ;9:334–356.

Meffre E, LeDeist F, de Saint-Basile G, Deville A, Fougereau M, Fischer A, Schiff C. A non-XLA primary deficiency causes the earliest known defect of B cell differentiation in humans: a comparison with an XLA case. Immunol Lett. 1997 Jun 1;57(1-3):93-9.

Mestas J, Hughes CC. Of mice and not men: differences between mouse and human immunology. J Immunol. 2004 Mar 1;172(5):2731-8.

Miller, S. (1989). The structure of interfaces between subunits of dimeric and tetrameric proteins, Protein Sci. 11: 184-197.

Mohamed AJ, Vargas L, Nore BF, Backesjo CM, Christensson B, Smith CI. (2000). Nucleocytoplasmic shuttling of Bruton's tyrosine kinase. J Biol Chem. 22; 275(51): 40614-9.

Mohamed AJ, Yu L, Bäckesjö CM, Vargas L, Faryal R, Aints A, Christensson B, Berglöf A, Vihinen M, Nore BF, Smith CI. (2009). Bruton's tyrosine kinase (Btk): function, regulation, and transformation with special emphasis on the PH domain. Immunol Rev;228(1):58-73.

Monera OD, Sereda TJ, Zhou NE, Kay CM, Hodges RS. (1995). Relationship of sidechain hydrophobicity and alpha-helical propensity on the stability of the single-stranded amphipathic alpha-helix. J Pept Sci. 1(5):319-29.

Moschese V, Orlandi P, Di Matteo G, Chini L, Carsetti R, Di Cesare S, Rossi P. Insight into B cell development and differentiation. Acta Paediatr Suppl. 2004 May;93(445):48-51.

Musacchio A, Gibson T, Lehto VP, Saraste M. SH3--an abundant protein domain in search of a function. FEBS Lett. 1992 Jul 27;307(1):55-61

Musacchio A, Gibson T, Rice P, Thompson J, Saraste M. The PH domain: a common piece in the structural patchwork of signalling proteins. Trends Biochem Sci. 1993 Sep;18(9):343-8.

Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol. 2004 Apr 16;338(1):181-99.

Nguyen JT, Turck CW, Cohen FE, Zuckermann RN, Lim WA. Exploiting the basis of proline recognition by SH3 and WW domains: design of N-substituted inhibitors. Science. 1998 Dec 11;282(5396):2088-92.

Nore BF, Vargas L, Mohamed AJ, Brandén LJ, Bäckesjö CM, Islam TC, Mattsson PT, Hultenby K, Christensson B, Smith CI. (2000). Redistribution of Bruton's tyrosine kinase by activation of phosphatidylinositol 3-kinase and Rho-family GTPases. Eur J Immunol. 30(1):145-54.

Ofran Y, Rost B. (2003). Analysing six types of protein-protein interfaces. J Mol Biol. 10;325(2):377-87.

Okoh MP, Vihinen M. (2002). Interaction between Btk TH and SH3 domain. Biopolymers. 15; 63(5):325-34.

Ottilie S, Diaz JL, Horne W, Chang J, Wang Y, Wilson G, Chang S, Weeks S, Fritz LC, Oltersdorf T. Dimerization properties of human BAD. Identification of a BH-3 domain and analysis of its binding to mutant BCL-2 and BCL-XL proteins. J Biol Chem. 1997 Dec 5;272(49):30866-72.

Patel HV, Tzeng SR, Liao CY, Chen SH, Cheng JW. SH3 domain of Bruton's tyrosine kinase can bind to proline-rich peptides of TH domain of the kinase and p120cbl. Proteins. 1997 Dec;29(4):545-52.

Pawson T, Raina M, Nash P. Interaction domains: from simple binding events to complex cellular behavior. FEBS Lett. 2002 Feb 20;513(1):2-10.

Phizicky E. M. and Fields S. (1995) Protein-protein interactions: Methods for detection and analysis. Microbiol Rev. 59, 94-123.

Porollo A, Meller J. (2007). Prediction-based fingerprints of protein-protein interactions. Proteins. 15;66(3):630-45.

Porollo A, Meller J. (2012). Computational Methods for Prediction of Protein-Protein Interaction Sites. In: Protein-Protein Interactions - Computational and Experimental Tools; W. Cai and H. Hong, Eds. InTech 2012; 472: pp. 3-26.

Qin S, Chock PB. Bruton's tyrosine kinase is essential for hydrogen peroxide-induced calcium signaling. Biochemistry. 2001 Jul 10;40(27):8085-91.

Qin S, Zhou HX. (2007) A holistic approach to protein docking. Proteins. Dec 1;69(4):743-9.

Qin S, Zhou HX. (2007a). meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics. 15;23(24):3386-7.

Qiu Y, Robinson D, Pretlow TG, Kung HJ. (1998). Etk/Bmx, a tyrosine kinase with a pleckstrin-homology domain, is an effector of phosphatidylinositol 3'-kinase and is involved in interleukin 6-induced neuroendocrine differentiation of prostate cancer cells. Proc Natl Acad Sci U S A. 31;95(7):3644-9.

Rafael A Jordan, Feihong Wu, Drena Dobbs and Vasant Honavar. ProtinDb: A data base of protein-protein interface residues. In preparation.

Rameh LE, Arvidsson Ak, Carraway KL 3rd, Couvillon AD, Rathbun G, Crompton A, VanRenterghem B, Czech MP, Ravichandran KS, Burakoff SJ, Wang DS, Chen CS, Cantley LC. A comparative analysis of the phosphoinositide binding specificity of pleckstrin homology domains. J Biol Chem. 1997 Aug 29;272(35):22059-66.

Rawlings D. J., Saffran D. C., Tsukada S., Largaespada D.A., Grimaldi J. C., Cohen L., Mohr R. N., Bazan J. F., Howard M., Copeland N. G., Jenkins N. A. & Witte O. N. (1993). Mutation of unique region of Bruton's tyrosine kinase in immunodeficient XID mice. *Science* 261, 358-61.

Sacristán C, Tussié-Luna MI, Logan SM, Roy AL. Mechanism of Bruton's tyrosine kinase-mediated recruitment and regulation of TFII-I. J Biol Chem. 2004 Feb 20;279(8):7147-58.

Salim K, Bottomley M J, Querfurth E, Zvelebil M J, Gout I, Scaife R, Margolis R L, Gigg R, Smith E I E, Driscoo P C, Waterfield M D,Panayotou G (1996) EMBO J 15:6241–6250.

Saraste M, Hyvönen M. Pleckstrin homology domains: a fact file. Curr Opin Struct Biol. 1995 Jun;5(3):403-8.

Sedivá A, Smith CI, Asplund AC, Hadac J, Janda A, Zeman J, Hansíková H, Dvoráková L, Mrázová L, Velbri S, Koehler C, Roesch K, Sullivan KE, Futatani T, Ochs HD. Contiguous X-chromosome deletion syndrome encompassing the BTK, TIMM8A, TAF7L, and DRP2 genes. J Clin Immunol. 2007 Nov;27(6):640-6.

Shen B, Vihinen M. (2004). Conservation and covariation of PH domain sequences: Physicochemical profile and information theoretical analysis of XLA-causing mutations in Btk PH domain. Protein Engineering, Design & Selection 17(3) pp. 267±276.

Sideras P, Müller S, Shiels H, Jin H, Khan WN, Nilsson L, Parkinson E, Thomas JD, Brandén L, Larsson I. (1994). Genomic organization of mouse and human Bruton's agammaglobulinemia tyrosine kinase (Btk) loci. J Immunol. 15;153(12):5607-17.

Smith, C. I. E., Baskin, B., Humire-Greiff, P. X., Zhou, J.-n., Olsson, P. g., Manier, H. S., Kjellen, P., Lambris, J. D., Christensson, B., Hammarstrom, L., Bentley, D., Vetrie, D., Islam, K. B., Vorechovsky, I. & Sideras, P. (1994) J. Immunol. 152, 557-565.

Smith, C. I. E., K. B. Islam, I. Vořechovský, O. Olerup, E. Wallin, H. Rabbani, B. Baskin, L. Hammarström. (1994). X-linked agammaglobulinemia and other immunoglobulin deficiencies. Immunol. Rev. 138: 159.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat;21(6):577-81.

Takata M, Kurosaki T. A role for Bruton's tyrosine kinase in B cell antigen receptor-mediated activation of phospholipase C-gamma 2. J Exp Med. 1996 Jul 1;184(1):31-40.

Takatsu K. [Role of interleukin-5 in immune regulation and inflammation]. Nihon Rinsho. 2004 Oct;62(10):1941-51.

Taylor WR. (1986). The classification of amino acid conservation. J Theor Biol. 21;119(2):205-18.

The UniProt Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 40: D71-D75. http://www.uniprot.org/

Thomas JD, Sideras P, Smith CIE, Vorechovský I, Chapman V, Paul WE. (1993). Colocalization of X-linked agammaglobulinemia and X-linked immunodeficiency genes. Science 261, 355-358.

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(5):703-14.

Tjong H, Qin S, Zhou HX. (2007). PI2PE: protein interface/interior prediction engine. Nucleic Acids Res. 35 (Web Server issue):W357-62.

Tjong H, Zhou HX. (2007).DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic Acids Res;35(5):1465-77.

Tzeng SR, Pai MT, Lung FD, Wu CW, Roller PP, Lei B, Wei CJ, Tu SC, Chen SH, Soong WJ, Cheng JW. Stability and peptide binding specificity of Btk SH2 domain: molecular basis for X-linked agammaglobulinemia. Protein Sci. 2000 Dec;9(12):2377-85.

Väliaho J, Smith CI, Vihinen M. (2006). BTKbase: the mutation database for X-Linked agammaglobulinemia. Hum Mutat. 12;27(12):1209-1217.

Vargas L, Nore BF, Berglof A, Heinonen JE, Mattsson PT, Smith CI, Mohamed AJ. (2002). Functional interaction of caveolin-1 with Bruton's tyrosine kinase and Bmx. J Biol Chem. 15;277(11):9351-7.

Várnai P, Rother KI, Balla T. (1999). Phosphatidylinositol 3-kinase-dependent membrane association of the Bruton's tyrosine kinase pleckstrin homology domain visualized in single living cells. J Biol Chem. 16;274(16):10983-9.

Vetrie, D., Vořechovský, I., Sideras, P., Holland, J., Davies, A., Flinter, F., Hammarström, L., Kinnon, C., Levinsky, R., Bobrow, M., Smith, C. I. E. & Bentley, D. R. (1993). The gene involved in X-linked agammaglobulinaemia is a member of the src family of protein-tyrosine kinases. Nature (London) 361, 226-233.

Vihinen M, Nilsson L, Smith C.I.E. (1994). Tec homology (TH) adjacent to the PH domain. FEBS Lett. 350: 263.

Vihinen M, Zvelebil MJ, Zhu Q, Brooimans RA, Ochs HD, Zegers BJ, Nilsson L, Waterfield MD, Smith CI. (1995). Structural basis for pleckstrin homology domain mutations in X-linked agammaglobulinemia. Biochemistry 7;34(5):1475-81.

Vihinen M, Mattsson PT, Smith CI. (1997). BTK, the tyrosine kinase affected in X-linked agammaglobulinemia. Front Biosci. Jan 1;2:d27-42.

Vihinen M, Brandau O, Brandén LJ, Kwan SP, Lappalainen I, Lester T, Noordzij JG, Ochs HD, Ollila J, Pienaar SM, Riikonen P, Saha BK, Smith CI (1998). BTKbase, mutation database for X-linked agammaglobulinemia (XLA). Nucleic Acids Res. 1;26(1):242-7.

Vihinen M, Kwan SP, Lester T, Ochs HD, Resnick I, Väliaho J, Conley ME, Smith CI. (1999). Mutations of the human BTK gene coding for bruton tyrosine kinase in X-linked agammaglobulinemia. Hum Mutat 13(4):280-5.

Vihinen M, Mattsson PT, Smith CI. (2000). Bruton tyrosine kinase (BTK) in X-linked agammaglobulinemia (XLA*). Front Biosci. 5:* D917-28.

Wang XC. (2004). [Clinical features of X-linked agammaglobulinemia: analysis of 8 cases]. Zhonghua Er Ke Za Zhi. 2004 Aug;42(8):564-7.

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

Yamadori T, Baba Y, Matsushita M, Hashimoto S, Kurosaki M, Kurosaki T, Kishimoto T, Tsukada S. Bruton's tyrosine kinase activity is negatively regulated by Sab, the Btk-SH3 domain-binding protein. Proc Natl Acad Sci U S A. 1999 May 25;96(11):6341-6.

Yi Q, Suzuki-Inoue K, Asazuma N, Inoue O, Watson SP, Ozaki Y. Docking protein Gab2 positively regulates glycoprotein VI-mediated platelet activation. Biochem Biophys Res Commun. 2005 Nov 18;337(2):446-51.

Yu L, Mohamed AJ, Vargas L, Berglöf A, Finn G, Lu KP, Smith CI. Regulation of Bruton tyrosine kinase by the peptidylprolyl isomerase Pin1. J Biol Chem. 2006 Jun 30;281(26):18201-7.

Yue P, Li Z, Moult J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353: 459–473.

Zhang QC, Petrey D, Norel R, and Honig BH. (2010). Protein interface conservation across structure space. Proc Natl Acad Sci USA, 107, 10896-10901.

Zhang QC, Deng L, Fisher M, Guan J, Honig B, Petrey D. (2011) .PredUs: a web server for predicting protein interfaces using structural neighbors. Nucleic Acids Res;39(Web Server issue):W283-7.

Zhou H-X, Shan Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins ;44:336–343.

Zhou H.X., Qin S. (2007). Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics. 1;23(17):2203-9.

8 APPENDICES

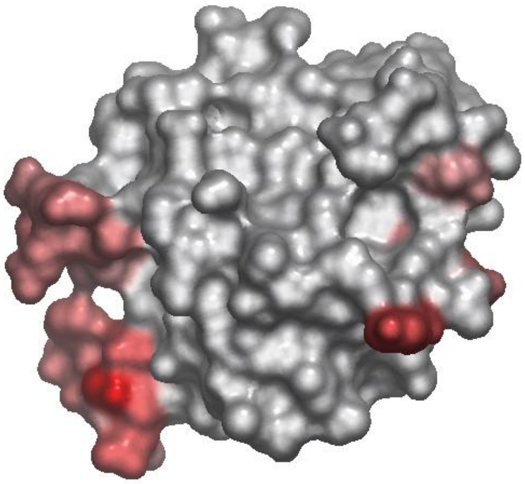


Figure 8.1 Heat map representation of human BTK PH domain showing interfacial regions. Image generated from PredUs protein interface predictor web server designed by Zhang *et al.* (2010); Zhang *et al.*(2011) and Fisher *et al.*(2011).

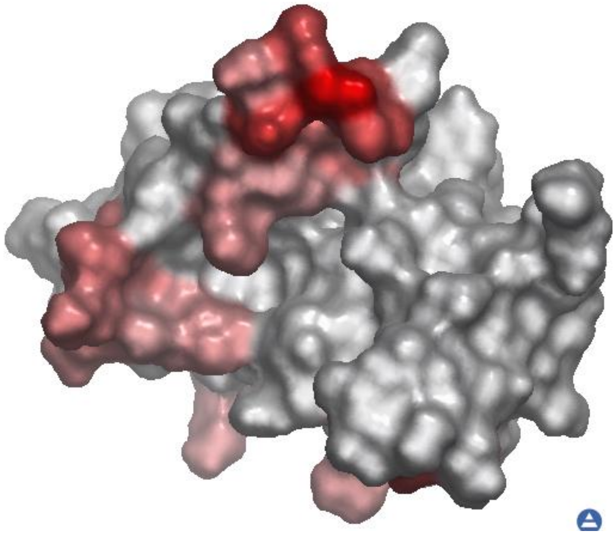


Figure 8.2 Heat map showing interfacial regions of the human BTK SH2 domain. Image generated from PredUs protein interface predictor web server designed by Zhang *et al.* (2010); Zhang *et al.*(2011) and Fisher *et al.*(2011).

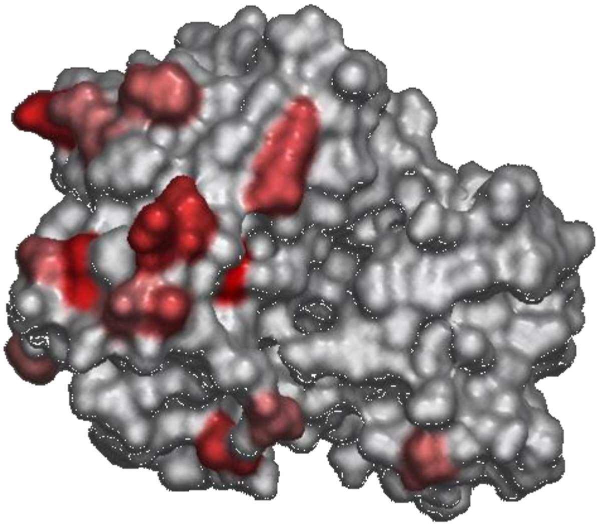


Figure 8.3 Heat map showing interfacial regions of the human BTK kinase domain. Image generated from PredUs protein interface predictor web server designed by Zhang *et al.* (2010); Zhang *et al.*(2011) and Fisher *et al.*(2011).