Structural and Functional Analysis of Single Amino Acid Replacements in Proteins: Insights from Protein Evolution into the Disease Aetiology

A dissertation submitted for the degree of *Doctor of Philosophy*

Sung Sam Gong

UNIVERSITY OF CAMBRIDGE

Hughes Hall

Cambridge, England

Feburary 2011

Declaration

This dissertation is my own work and includes nothing which is the outcome of work done in collaboration except where specified in the text. I state that this dissertation is not substantially same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. This thesis does not exceed the word limit.

Sung Sam Gong
Cambridge, England
Feburary, 2011

To all the parents of scientists

Acknowledgements

I am grateful to all who are developing and maintaining biological databases, scientists submitting their invaluable data, and people who support open source programs and operating systems. I appreciate the Mogam Science Scholarship Foundation of Korea for the financial support during my second year PhD study, the Korean Scientists and Engineers Association in the U.K. (KSEAUK), Hughes Hall my college, and the Sanger fund from the Department of Biochemistry for their travel bursaries. I thank my colleagues Semin Lee, Adrian Schreyer, and Dr. Richard Bickerton for their collaboration on developing databases. I also thank other Biocomputing people Dr. Duangrudee Tanramluk, Dr. William Pitt, Bernardo Ochoa, Dr. Catherine Worth, Alicia Higueruelo, Jawon Song and Dr. Tammy Cheng. Also I do not forget generous help from Graham Eliff, Dr Cynthia Lampert Moore, Nikki Miller and Irene Knightley. I thank Prof. Sir Tom Blundell, my supervisor, for his kind and generous help and guidance during my PhD study and my MSc supervisor and friend Dr. Jong Bhak for his generous hardware supports and cheerful and inspiring chats. Lastly but with my best regards, I am very grateful for my parents, Jaenam Lee (이재남) and Gibok Gong (공기복), for their endless supports during my stay in Cambridge.

Abstract

High-throughput genomic sequencing has focused attention on understanding differences between species and between individuals. When this genetic variation affects protein sequences, the rate of amino acid substitution reflects both Darwinian selection for functionally advantageous mutations and selectively neutral evolution operating within the constraints of structure and function. During neutral evolution, whereby mutations accumulate by random drift, amino acid substitutions are constrained by factors such as the formation of intramolecular and intermolecular interactions and the accessibility to water or lipids surrounding the protein. In this thesis, I attempt to address structural and functional restraints that shape replacement of amino acids during protein evolution and apply the general rules in the study of amino acid variations associated with disease etiology.

I first focus on the use of amino acid substitution model and address how the description of amino acid replacement could be improved by discriminating local structural environments from the following four categories of functional restraints: i) protein-protein interactions, ii) protein-nucleic acid interactions, iii) protein-ligand interactions and iv) catalytic activity of enzymes. I characterize the impacts of various functional restraints on the conservation of amino acids in three-dimensional structures. To better understand how amino acids are substituted under their local environments — often defined by secondary structure, solvent accessibility and the existence of hydrogen-bonds from side-chains to main-chains or other side-chains — I quantify and rank the determinants of amino acid substitutions in the three-dimensional structures of proteins by the way they affect the rate of accepted substitutions. I show that solvent accessibility is the most important determinant, followed by the existence of hydrogen-bonds from the side-chain to main-chain functions and the nature of the element of secondary structure to which the amino acid contributes.

From the observation of amino acid replacements which are under restraints of the local structural and functional environments, I apply those principles in the study of human genetic variation from the following three categories: i) Mendelian disease-related

variants, ii) neutral polymorphisms and iii) cancer somatic mutations. I characterize structural and functional environments where the variants occur and compare how the environments are different amongst three groups. I show that various types of variants are under different degrees of structural and functional restraints, which affect their occurrence in human proteome. Then, I exemplify how the understanding of structural and functional restraints imposed on proteins could help identify genetic variations associated with a disease by demonstrating analysis of genetic variations responsible for type 1 diabetes. The genetic variations are from the John Todd group, Cambridge institute of medical research, and consist of 355 Single Nucleotide Polymorphisms (SNPs) within protein coding regions.

Finally, I describe a development web-based database system which houses structural and functional annotations of amino acid residues which have been used during this study. The system, which is named SAMUL, interconnects the Blundell group's in-house databases focused on molecular interactomes and external data sources such as PDB, UniProt and Ensembl. In addition, SAMUL accommodates amino acid variation and mutation data mentioned earlier and provides an interface in which people can navigate the mutations in the context of three-dimensional structure of proteins, if available, and interpret their severity in conjunction with the structural and functional environments where the variants occur at the wild type amino acid.

Table of Contents

Lists of Figures

Lists of Tables

Abbreviations

3D	three-dimensional
ESST	environment-specific substitution table
GWAS	genome-wide association study
LD	linkage disequilibrium
T1D	type 1 diabetes
SNP	single nucleotide polymorphism
nsSNP	non-synonymous single nucleotide polymorphism
UTR	untranslated region
PDB	protein data bank
PCA	principal component analysis
NH	amide
CO	carbonyl
SC	side chain
API	application programming interface
XML	extensible markup language
RDBMS	relational data base management system
DAS	distributed annotation system
HTTP	hypertext transfer protocol
SAMUL	systematic annotation of macro-molecules

Chapter 1
Introduction

1.1 Protein Evolution

1.1.1 Overview

An understanding of protein evolution requires not only knowledge of genomes, protein sequences, structures and functions but also an understanding of selective pressures at the level of the whole organism and the role of the protein in cells and whole organisms [1,2].

Insights into the relationship of protein structure, function and evolution began to emerge nearly fifty years ago as protein structures were determined for which there were multiple sequences. For example, insulin sequences from Fred Sanger in the 1950s [3,4] (See [5] for review) together with the three-dimensional structure from Dorothy Hodgkin a decade later [6,7] provided clues about the impacts of amino acid substitutions on tertiary structure and precursor activation, on quaternary interactions at dimer and hexamer interfaces, and on the putative receptor binding region [8]. Through the comparative analysis of insulins in different species, they observed that much of the sequence variation appeared to be selectively neutral; the accepted amino acids were able to fulfil the same structural and functional roles such as those occurring in the hystricomorph (e.g. rodents) insulins. However, these substitutions proved to be consistent with loss of ability to dimerise and stabilisation of the monomeric form. This presumably resulted from change of storage form, possibly related to zinc availability, and was therefore probably also selectively advantageous.

These observations lead Kimura and Ohta to develop the neutral theory of evolution, which states that the majority of evolutionary changes at the molecular level are caused by neutral drift, the acceptance of selectively neutral mutations [9,10]. They suggested that mutations that disrupt the existing structure and function of a molecule occur less

frequently in evolution than neutral mutations. This idea was elaborated by Zuckerkandl and colleagues in the functional density hypothesis, which proposes that the rate of evolution is determined by the proportion of all possible mutations which produce a protein that is functionally equivalent to the wild type [11,12]. More recently, Fraser *et al*. [13] demonstrated that proteins with many interaction partners evolve more slowly than those with few interaction partners [14,15], but this has been disputed [15].

1.1.2 Comparative analyses of homologous proteins

The first comparisons of primary and tertiary structures of homologous proteins forty to fifty years ago — globins, serine proteinases and lysozymes — focused on accessibility to water, usually called solvent accessibility, and showed that the solvent inaccessible cores of proteins tended to be closely packed, more hydrophobic and more conserved than the surface regions [16]. Analyses of the structures from many protein families show that this remains a useful generalization. These early analyses also focused on regular secondary structures, such as α-helices and β-sheets, which were immediately recognised to favour particular amino acids, so providing further constraints on evolutionary change [17,18,19].

Pauling and colleagues realized that the requirement for satisfaction of hydrogen-bonding potential of polypeptide mainchain peptide amide (NH) and carbonyl (CO) groups would not only give rise to regular secondary structures [20,21], but also make the mainchains of proteins more hydrophobic so that they could be buried in the core of a globular protein along with non-polar sidechains. It soon became evident that these features of mainchain hydrogen bonding restrict protein architectures to a limited set of super-secondary structures formed by combinations of secondary structures into globular units, such as β-sheets, jelly rolls, β-propeller, α-helical bundles, αβ-Rossman fold, αβ-barrel and many others. Mainchain hydrogen bonding also has important roles in the formation of complex arches and turns that link α-helices and β-strands [22,23,24].

Nevertheless many main-chain peptide CO and NH groups are left unsatisfied in their potential to form hydrogen bonds: an early analysis of hydrogen bonding revealed that ~40% of such groups do not form hydrogen bonds with mainchain atoms of other amino acids [25]. In general these occur at places where strands and helices terminate [25,26,27,28], bulge [29,30] or bend [31,32], but they are also common in polyproline or irregular, twisted strands [33,34] and in arches and turns [22,23,24,35,36]. The hydrogen-bonding potential of these motifs is satisfied by water molecules or by polar sidechains; when the sidechains are inaccessible they provide a strong restraint on neutral drift.

1.1.3 Knowledgebase for a comparative study

Insight into evolutionary relationships can be gained by grouping similar proteins and comparing sequences and structures of members of families and superfamilies — proteins that are homologous or descended from a common ancestor — to be found amongst the more than fifty thousand proteins for which architectures have been determined at high resolution. Several classification resources, as shown in Table 1-1, categorize proteins based on the degree of similarity but they differ in definition and method. Nevertheless, there is general agreement on the hierarchical order of overall topology or fold, superfamily, family and individual domain and many proteins adopt regular architectural arrangements of polypeptide chains — often called protein folds — which categorized into the same topology [2]. In addition, it is believed that members of superfamilies and families are likely to have arisen from a common ancestor by divergent evolution. SCOP [37] and CATH [38] are two well known databases of hierarchical protein structure classification. HOMSTRAD [39], PASS2 [40], Toccata [41] and FSSP [42] provide superimposed and aligned protein families with various annotations at the residue level. CE [43] also provides structure comparison and alignment. MMDB provides structure neighbour calculations such that each structure is linked to related three-dimensional domains [44]. Sequence based protein family databases include Pfam [45] and InterPro [46]. InterPro is a consortium of several member databases such as PROSITE [47], Pfam, Prints [48], ProDom [49], SMART [50] and TIGRFAMs [51]. Using curated or computed classification schemes of

proteins, homology detection can be achieved using sequence and/or structure similarity as implemented by Gene3D [52], Superfamily [53], PhyloFacts [54], CDD [55], PairsDB [56] and SMART. These databases and servers can be useful resources in the study of protein evolution. A comprehensive comparison of these databases and servers is available in Orengo *et al.*[57]

Table 1-1 A list of protein classification databases and similarity search servers

Type	Database	Method (tool)	URL	Summary
Structure classification	SCOP	Manual Curation	http://scop.mrc-lmb.cam.ac.uk/scop	Classification of structural protein families by human experts
	CATH	SSAP [58]	http://www.cathdb.info/	(Semi)automatic classification of protein structures into hierarchies
	HOMSTRAD	Comparer [59]	http://www-cryst.bioc.cam.ac.uk/~homstrad	Manual classification of protein homologues with JOY [60] format annotation
	PASS2	Comparer	http://caps.ncbs.res.in/campass/pass2.html	Database of protein structure alignments at the superfamily level
	TOCCATA	Baton (successor of Comparer)	http://www-cryst.bioc.cam.ac.uk/toccata/toccata.php	Database of protein structure alignments on the basis of SCOP family
	CE	CE	http://cl.sdsc.edu	Protein structure comparison and alignment using the combinatorial extension (CE) method
	MMDB	VAST [61]	http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml	Protein structures with visualization and comparison analysis
	FSSP	DaliLite [62]	http://ekhidna.biocenter.helsinki.fi/dali/start	Protein fold families structurally aligned on the basis of exhaustive 3D structure comparison
Sequence family	Pfam	HMM [63]	http://pfam.sanger.ac.uk	Collection of protein families with multiple sequence alignment using Hidden Markov Models (HMMs)
	InterPro	various methods	http://www.ebi.ac.uk/interpro	Protein families, domains, regions, repeats and sites in which identifiable features found in known proteins can be applied to new protein sequences

	ProDom	PSI-BLAST [64]	http://prodom.prabi.fr/prodom/current/html/home.php	Automatic classification of protein domain families on the basis of UniProt [65] knowledge database
	TIGRFAMS	HMM	http://www.jcvi.org/cms/research/projects/tigrfams/overview/	Protein families based on Hidden Markov Models (HMMs)
	PROSITE	manual curation	http://www.expasy.ch/prosite	Protein families and domains with human-curated annotations such as functional sites, sequence motifs and profiles
Homology detection	Gene3D	profile-HMM	http://gene3d.biochem.ucl.ac.uk/Gene3D/	Structural and functional annotation of protein sequences on the basis of CATH and profile-HMM library
	Superfamily	HMM [66]	http://supfam.cs.bris.ac.uk/SUPERFAMILY	Database of structural and functional protein annotations on the basis of SCOP (super)families using HMMs
	PhyloFacts	FlowerPower [67]	http://phylogenomics.berkeley.edu/phylofacts	Pre-calculated structural and phylogenomic analyses of protein families and domains
	CDD	CD-Search [68]	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	Multiple sequence alignments for ancient domains and full-length proteins
	PairsDB	BLAST, PSI-BLAST [64]	http://pairsdb.csc.fi	Protein sequences and BLAST and PSI-BLAST alignments between them
	SMART	HMM	http://smart.embl-heidelberg.de	Online tool for the exploration and comparative study of domain architectures in both proteins and genes

1.1.4 Restraints of amino acid conservation

From the comparative analyses of insulin structures, Blundell and colleagues suggested that amino acid substitutions were accepted during evolution in a way that satisfied restraints arising from structure and function [69]. Thus, the core of the protein tended to be relatively conserved [16] and residues in helices and strands were substituted in ways that maintained the overall stabilities of these secondary structures. Most interestingly, a glycine with a positive phi main-chain torsion angle[1] that allowed the chain to change direction sharply was conserved in all insulins. Substitutions of amino acids at positions involved in dimer formation retained their hydrophobic character in all species except the hystricomorpha. The conservation of B10 His in most mammalian, fish and bird insulins was evidence of restraints arising from the existence of a hexamer.

The insulin structure also provided a good evidence of restraints from functional interactions. Residues in a patch mainly on the surface of the monomer appeared to have greater restraints on their substitution than could be explained by retention of the structure of insulin throughout evolution; this observation provided the clues about restraints in evolution arising from function, in this case the binding of insulin with its receptor.

Thus, the analyses of insulins, along with parallel work on globins, lysozymes and serine proteinases, provided strong evidence for the conservation of tertiary structure during evolution, and emphasised the importance of considering restraints from protein interactions, in this case in terms of oligomers and receptor activation. They underlined the importance of local environment in the acceptance of amino acid substitutions during protein evolution. These are covered in more detail in section 1.2.2

However, there are many other restraints that are less well understood, but provide important pressures in evolution. They include those that arise from DNA packaging

[1] A positive dihedral angle around the nitrogen–α-carbon bonds in the protein main chain. For L-amino acids these bond angles are generally restricted to a negative value owing to steric hindrance from the side chains, but they can be positive when there is no side chain (Gly) or when polar side-chain interactions with the main-chain peptide units stabilize this. See Figure 1-1 for details.

and gene splicing and from the requirement of reliable and well coordinated gene expression [70,71,72], for example ubiquitously expressed proteins tend to evolve more slowly than tissue-specific genes. In addition, they arise from the process of protein folding [73,74], from the importance of retaining various conformational changes and flexibility that mediate functions in the cell, and from the need to avoid opportunistic interactions (interactions occurring by chance) and amyloid formation – aggregation of misfolded proteins into a highly ordered fibril-like structure [75,76]. Furthermore, in order to prevent accumulation of damaging proteins the protein degradation system must be finely controlled, especially for misfolded proteins resulting from mutations [77]. Recently, it has been found that epigenetic factors, such as DNA methylation and chromatic remodelling, have important roles in the regulation of gene expression [78], which eventually affects the evolution of proteins. Hence, an integrated approach is required comprehensively to understand protein evolution [79].

1.2 Amino acid substitution models

1.2.1 A brief history

Proteins existing in living organisms have been selected through the process of evolution. However, as mentioned previously, much of the amino acid variation between orthologues appears to be selectively neutral [9] as far as the whole organism is concerned and accepted amino acid substitutions result in equal fitness. It has been long understood that the rate and nature of accepted mutations or substitutions are different for the 20 amino acids in a protein.

Indeed, between the late 1960s and early 1970s, the different substitution rates and patterns for the 20 amino acids were first quantified by Margaret Dayhoff as the PAM (Percentile Accepted Mutation or Point Accepted Mutation) matrix based on 1572 observed mutations in 71 families of closely related proteins [80]. PAM measures evolutionary distance of divergence in a protein where the PAM1 matrix states that the rate of substitution if 1% of the amino acids has changed. Using this logic, Dayhoff derived matrices as high as PAM250. Richard Grantham introduced a measurement

describing differences between amino acids, which correlate with amino acid substitution frequencies by categorizing chemical dissimilarity between the encoded amino acids [81]. The methodology was further developed by Henikoff *et al.* [82] to reflect more divergent relationships of protein sequences. BLOSUM62 is now recognized as a standard measure of substitution rate for 20 amino acids in the sequence comparisons. Jones *et al.* [83] introduced a fast and automated approach based on a maximum parsimony counting method (known as JTT substitution model) and Gonnet *et al.* [84] introduced a different method to measure differences among amino acids using exhaustive pairwise alignments of the protein databases as they existed at that time. Whelan *et al.* [85] applied a maximum-likelihood method to estimate the rate for amino acid replacement (known as WAG). Recently, Le *et al.* [86] claimed that they further refined the WAG method by incorporating the variability of evolutionary rates across sites in the matrix estimation and using a much larger and diverse database. All these substitution models are based on the sequence alignments of closely related protein families without considering three-dimensional information of protein structures.

However, sequence alignments of homologues of known structure can be used to help quantify the restraints that arise from both protein structure and function in a family of proteins. The local environments of individual amino acid side-chains restrain the accumulation of amino acid substitutions as proteins undergo neutral evolution. As we have learnt from comparative analyses of protein structures in section 1.1.2, one of the strong restraints arises from the need to maintain three-dimensional structure in order to retain function.

Analyses of protein families or superfamilies led to the idea that propensities for amino acids [87] and their substitution patterns [88,89] might be systematically defined in terms of local structural environments. Solvent accessibility of the side-chain and occurrence in regular secondary structures were local environments used by most groups [87,90,91,92]. Two further classes of local environment were added to these by Overington *et al.* [89]: (i) amino acids with a positive phi main-chain torsion angle (learning from the B8 Gly of insulin) and (ii) amino acids with side chains that formed hydrogen bonds to main-chain or other side-chain functions (inspired by the conserved

serine and threonine residues of the crystallins and aspartic proteinases). Below, I describe a substitution matrix which describes exchangeability of amino acids as a function of local structural environments where the amino acids occur within three-dimensional structure of proteins.

1.2.2 ESST: Environment Specific Substitution Table

The Environment Specific Substitution Table (ESST) is a substitution table that considers structural restraints in the calculation of substitution patterns. Overington *et al.* [88,89] first calculated ESSTs from a set of homologous protein families whose three-dimensional structures were available. The rationale behind ESSTs is that the acceptance of substitution of an amino acid in an orthologous family is subject to its local tertiary environment. The local structural environments of amino acids include 1) main-chain conformation and secondary structure, 2) solvent accessibility and 3) hydrogen bonding between side-chain and main-chain (see Figure 1-1). 64 ESSTs can be derived from a combination of structural features; four from secondary structures (α-helix, β-strand, coil and residue with positive φ main-chain torsion angle; see Figure 1-2 for details), two from solvent accessibility (accessible and inaccessible), and eight (2^3) from hydrogen bonds to main-chain carbonyl or amide or to another side-chain (see Table 1-2). These combinations of structural features restrict possible substitutions of an amino acid and give rise to distinct patterns of substitution. Summing all 64 tables, leads to an environment-independent 20*20 matrix such as PAM [80] or BLOSUM [82]. Hence, ESST further divides the conventional substitution table into 64 matrices, which differ in the local tertiary environments of amino acids in protein three-dimensional structures. In Chapter 2, I describe how the calculation of ESSTs can be improved by using only amino acids that are not involved in catalytic activity, metal or ligand binding, nucleic acid or protein interactions and other molecular functions.

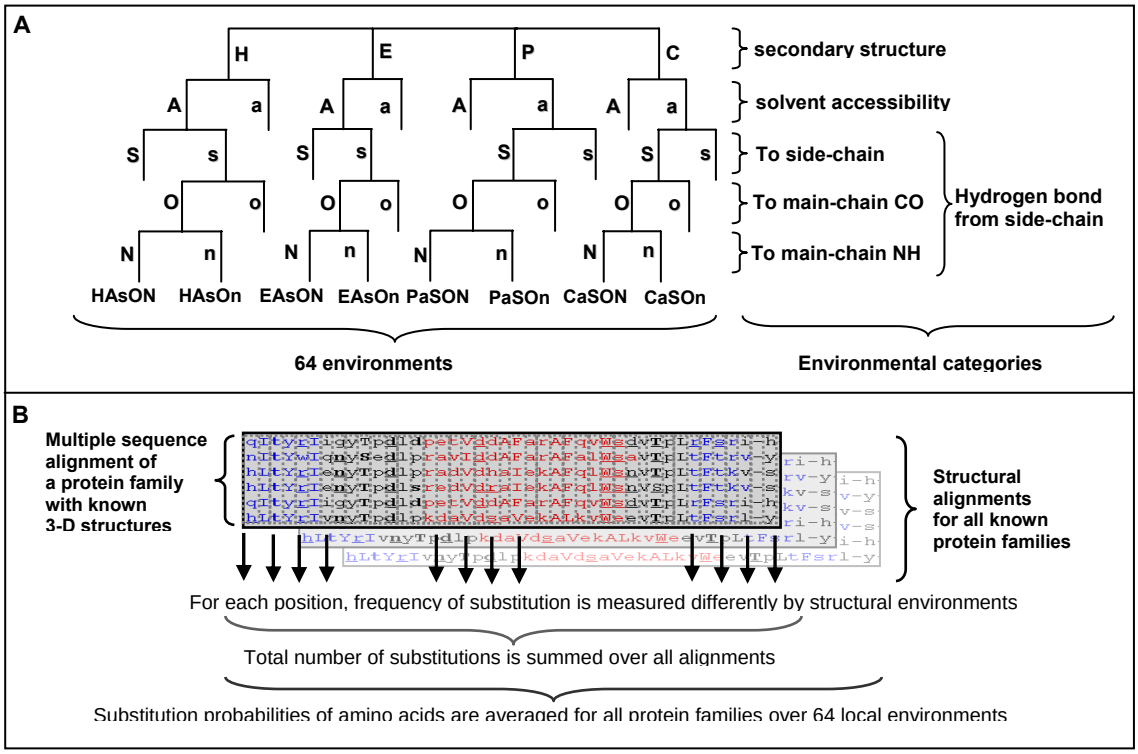


Figure 1-1 Environmental categories and a schematic diagram of ESST generation

A. Environment-specific substitution tables (ESSTs[2]) provide the basic evidence that amino acid substitutions are constrained in different ways in different local environments. Such tables exploit categories of amino acid local structural environments, such as main-chain conformation and secondary structure, solvent accessibility, and hydrogen bonding between side chains and either main-chain groups or other side chains. For example, in part a of the figure, amino acids can be classified into 1 of 64 environments: 4 from secondary structure (α-helix (H), β-strand (E), positive φ main-chain torsion angle (P) and coil (C)), 2 from solvent accessibility (accessible (A) and inaccessible (a)), and 8 from the existence (upper case) or absence (lower case) of hydrogen bonds from a side chain to another side chain (S and s), to a main-chain carbonyl group (O and o) and to a main-chain amide group (N and n). These combinations of structural features influence the substitution of amino acids and give rise to distinct patterns of amino acid substitutions.

B. ESSTs can be generated from homologous protein structure alignments in which each residue has been annotated with three-dimensional structural features (explained above) and assigned to one of the 64 environments in JOY [60] format: solvent inaccessible (upper case), solvent accessible (lower case), α-helix (red), β-strand (blue), hydrogen bond to main-chain amide group (bold) and hydrogen bond to main-chain carbonyl group (underlined). The frequency of amino acid substitutions is measured by each structural environment and averaged over all homologous protein families. Ulla [93] is a program that

[2] http://samul.org/ESST

generates ESSTs from a set of structure alignments, annotated in various structural and functional environments for amino acid residues.

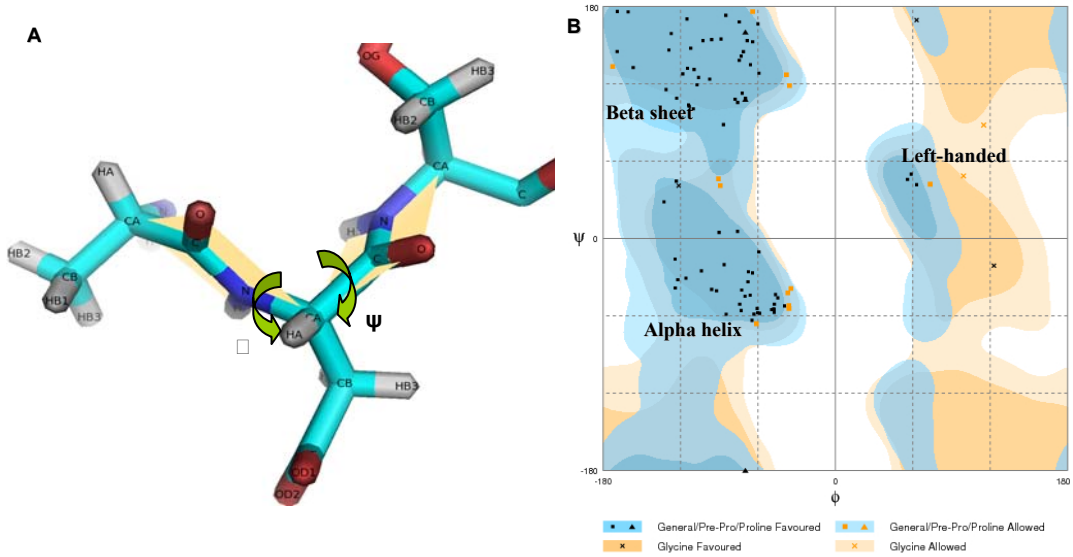


Figure 1-2 An example of backbone dihedral (or torsion) angles and Ramachandran plot

A. Two backbone dihedral angles, φ and ψ, are demonstrated using three amino acids, Val44-Asp45-Ser46, from a solution structure of the zinc finger CCCH domain containing protein (PDB: 2E5S). Asp45 is shown in the middle with Val44 and Ser46 on its left and right, respectively. Dihedral angle φ is an angle involving the backbone atoms C'-N-Cα-C', and ψ is a dihedral angle involving the backbone atoms N-C^{α}-C'-N. Two planes, spanning across peptide bonds, are highlighted in yellow with nitrogen and oxygen coloured in blue and red, respectively. Carbon and hydrogen are coloured in cyan and grey. This figure is drawn using PyMOL [94] with the BbPlane[3] script.

B. Once dihedral angles are calculated for every amino acid residue within a polypeptide chain, they can be plotted on X (φ) and Y (ψ) axis, which is known as the Ramachandran plot [95]. The plot visualises the possible conformations of φ and ψ angles from the same three-dimensional structure of a protein shown in **A**. Different elements of secondary structures are clustered distinctively and occupy unique regions as shown in the figure. However, certain regions of the plot are almost forbidden for an amino acid to occupy, because some torsion angles are physically and energetically unfavourable for atoms to adopt to prevent steric hindrance within the polypeptide chain. Hence, the Ramachandran plot could be used to assess the quality of a protein three-dimensional structure. This figure is drawn using the RAMPAGE web server [96].

[3] http://pymolwiki.org/index.php/BbPlane

Table 1-2 Local structural environments

Local Structural Environments					Abbreviations
Secondary Structure	Solvent accessibility	Existence of hydrogen-bond from side-chain[4]			
		to other side-chain	to main-chain CO (carbonyl)	to main-chain NH (amide)	
Coiled coil	accessible	T	T	T	CASON
		T	T	F	CASOn
		T	F	T	CASoN
		T	F	F	CASon
		F	T	T	CAsON
		F	T	F	CAsOn
		F	F	T	CAsoN
		F	F	F	CAson
	inaccessible	T	T	T	CaSON
		T	T	F	CaSOn
		T	F	T	CaSoN
		T	F	F	CaSon
		F	T	T	CasON
		F	T	F	CasOn
		F	F	T	CasoN
		F	F	F	Cason
beta strand	accessible	T	T	T	EASON
		T	T	F	EASOn
		T	F	T	EASoN
		T	F	F	EASon
		F	T	T	EAsON
		F	T	F	EAsOn
		F	F	T	EAsoN
		F	F	F	EAson
	inaccessible	T	T	T	EaSON
		T	T	F	EaSOn
		T	F	T	EaSoN
		T	F	F	EaSon
		F	T	T	EasON
		F	T	F	EasOn
		F	F	T	EasoN
		F	F	F	Eason
alpha helix	accessible	T	T	T	HASON
		T	T	F	HASOn
		T	F	T	HASoN

[4] T: existence, F: non-existence

		T	F	F	HASon
		F	T	T	HAsON
		F	T	F	HAsOn
		F	F	T	HAsoN
		F	F	F	HAson
	inaccessible	T	T	T	HaSON
		T	T	F	HaSOn
		T	F	T	HaSoN
		T	F	F	HaSon
		F	T	T	HasON
		F	T	F	HasOn
		F	F	T	HasoN
		F	F	F	Hason
positive-phi mainchain torsion angle	accessible	T	T	T	PASON
		T	T	F	PASOn
		T	F	T	PASoN
		T	F	F	PASon
		F	T	T	PAsON
		F	T	F	PAsOn
		F	F	T	PAsoN
		F	F	F	PAson
	inaccessible	T	T	T	PaSON
		T	T	F	PaSOn
		T	F	T	PaSoN
		T	F	F	PaSon
		F	T	T	PasON
		F	T	F	PasOn
		F	F	T	PasoN
		F	F	F	Pason

Figure 1-3 demonstrates that amino acid substitution patterns are influenced by local structural environments. In particular, a solvent inaccessible environment (Figure 1-3B) restricts the possible substitution of amino acids most strongly, enhancing the diagonal of the substitution matrix, but secondary structure and the existence of side-chain hydrogen-bonds also lead to different substitution patterns (see Figure 1-3A and Figure 1-3C). These ESSTs also show that amino acids with sidechains that are hydrogen-bonded to mainchain NH (Figure 1-3D) and CO groups are more conserved than those with sidechains that are not hydrogen-bonded to mainchain NH or CO. This is

particularly evident when sidechains are inaccessible to solvent and when they form hydrogen bonds to mainchain amide NH groups, as shown by the bar chart in Figure 1-4. This implies that a crucial element in protein structure is the satisfaction of hydrogen-bond donor and acceptor properties of the mainchain NH and CO groups when the protein is folded. When these requirements are not satisfied by secondary structures, hydrogen bonds to sidechains might be conserved to meet this requirement [97]. In Chapter 3, I describe the relative importance of those local structural environments by an analysis of distances amongst the 64 tables – each characterised by a different set of restraints – followed by Principal Component Analysis (PCA).

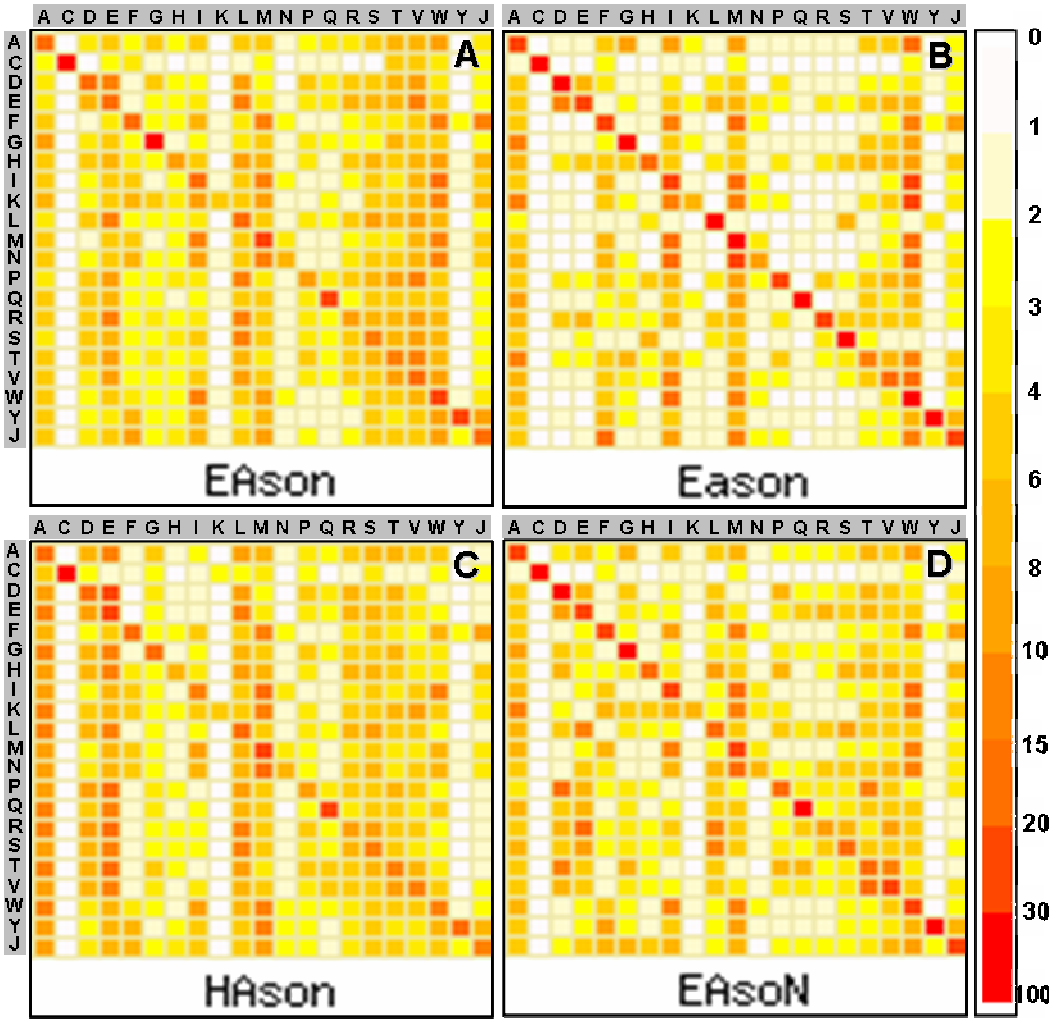


Figure 1-3 Four examples of ESSTs

A-**D** demonstrate that amino acid substitution patterns are influenced by local structural environment. **A**: solvent-accessible β-strand with no hydrogen-bonds from sidechains, **B**: solvent-inaccessible β-strand

with no hydrogen-bonds from sidechains, **C**: solvent-accessible α-helix with no hydrogen-bonds from sidechains, and **D**: solvent-accessible β-strand with only one hydrogen-bonds from sidechain to mainchain NH (see Table 1-2 for details). Matrices **B**-**D** differ from **A** by only one structural environment. The degree of amino acid conservation is represented as heatmap from 0% (non-conserved) to 100% (conserved). Note that the colour scale of percentage is not evenly distributed to emphasize the difference of amino acid substitution patterns amongst four matrices. The pictures were drawn with the Perl GD module[5].

Compared with traditional substitution tables (PAM, BLOSUM) derived from sequence information only, ESSTs were shown to give more precise and discriminating measures of substitution probabilities [98]. ESSTs have been shown to be useful in applications to secondary structure prediction [98] and sequence-structure homology recognition [92,99]. Recently, CRESCENDO, a computer software predicting functional residues from known three-dimensional structures of proteins, has been successful in prediction of functional residues by comparing the observed substitution patterns for amino acids which are under both functional and structural constrains with those that are predicted on the basis of structure alone [100].

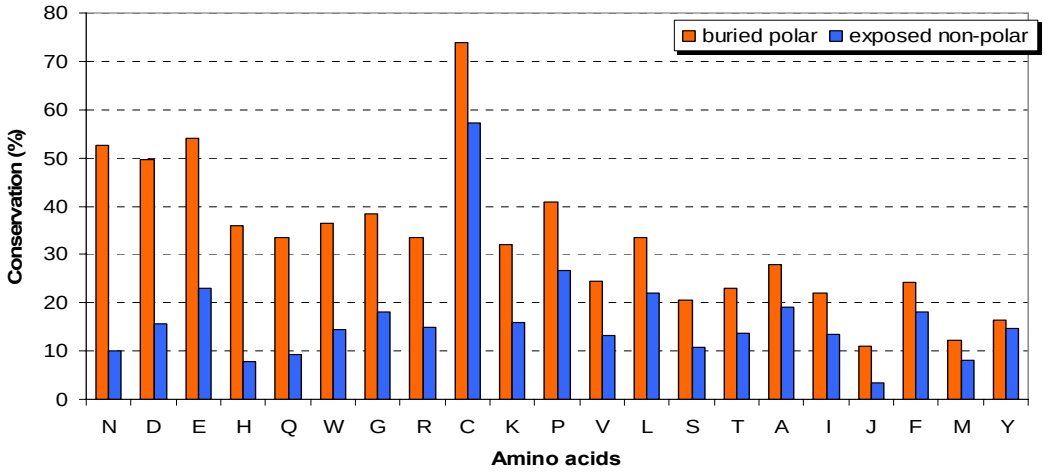


Figure 1-4 Differences in the probabilities of amino acid conservation between buried polar and exposed non-polar environments

Amino acids of 'buried polar' are from a substitution matrix 'HaSON' which represents solvent inaccessible sidechains that take part in hydrogen-bonds to mainchain functions or other sidechains,

[5] http://search.cpan.org/dist/GD/

whereas amino acids of 'exposed non-polar' are from 'HAson' which states solvent exposed sidechains that do not take part in hydrogen bonds (See Figure 1-1 and Table 1-2 for details). The probabilities of residue conservation, which are from the diagonal entries of corresponding substitution tables 'HaSON' and 'HAson', are plotted for 21 amino acids in descending order of the differences of probability scores. Note that amino acid 'C' represents half-cystine (disulphide bonded) and 'J' represents cysteine (non-disulphide bonded). Two matrices, HaSON and HAson, are chosen to illustrate how solvent accessibility (A/a) and hydrogen-bonds (SON/son) affect the degree of amino acid conservation.

1.3 Amino acid variations and diseases

1.3.1 Insights gained from Mendelian disease

Before the determination of the human genome sequence, analysis of genetic mutations focused on establishing the relationship between genotypes and their phenotypes, especially susceptibility to certain disease types [101,102]. However, there were no general methods identifying DNA sequences responsible for even simple Mendelian diseases until Botstein and colleagues developed a method which constructs a linkage map of the human genome, with restriction fragment length polymorphisms (RFLPs) as molecular markers in 1980 [103,104]. After this initial milestone, the human genetic linkage map and the methods and algorithms have been applied for connecting disease genes, traits or mutations with Mendelian diseases and successful in identification of 1,200 disease genes including classic examples of sickle-cell anaemia [105], hemochromatosis [106] , and lactose intolerance [107]. See [108,109] for reviews.

Detailed molecular analyses of protein structure and function have revealed that single amino acid substitutions or mutations are often responsible for certain disease types [110,111]. It has been claimed that ~60% of such Mendelian disease mutations arise from amino acid substitutions in their respective genes (see [108] for review). For most monogenic diseases, a single DNA variant resulting in an amino acid substitution is responsible for a certain disease type by affecting protein stability and thus function [112]. Hence, much effort has been expended to characterise the pattern of mutations in the context of sequences and structures of proteins in attempts to establish whether they are likely to be neutral or deleterious for the functions of the organism

[113,114,115,116]. Interestingly, most of the methods that aim to assess deleterious mutations are based on the principles observed from nature – to see whether the mutations conform to the neutral theory of protein evolution (see section 1.1.1), which selects against radical changes of amino acids. In this context, I seek to address structural and functional features of proteins that restrain genetic variation leading to single amino acid substitutions in Chapter 4. However, real challenges at present are from complex diseases that obscure the genetic basis responsible for molecular phenotypes.

1.3.2 Challenges from complex diseases

Linkage mapping, as mentioned earlier, has been successful for Mendelian diseases such as Huntington disease [117] and cystic fibrosis [118,119] where the causation of genotype and phenotype is straightforward and the diagnosis is unequivocal due to the monogenic nature of the diseases [108]. However, even before the first linkage map was completed, it was recognized that most human traits and diseases follow complex modes of inheritance. Hence it is not a trivial task underpinning genetic traits or variants responsible for complex diseases such as cancers and diabetes where the phenotypes are determined by coordination of multiple genes and interactions between genes and environmental factors that can affect gene expressions. In addition, unlike Mendelian diseases for which genetic variations in protein coding regions are responsible in most cases, it is reported that genetic variations in intergenic regions, introns, regulatory regions (e.g. transcription factor binding sites) and even synonymous mutations, which do not change amino acid types, can be responsible for complex diseases by affecting translational efficiency, mRNA stability, splicing control, post-translational modifications and chromosomal rearrangement [120,121,122].

The difficulties seem to start to be compensated by technical advancements in modern sequencing methods (see for [123] a review), which enable charting genetic variations between human individuals in a fast and high-accurate manner. A seminal project

initiated from the Wellcome Trust Case Control Study (WTCCS[6]) harnesses the power of such genotyping technologies to improve our understanding of the aetiological basis of causes of complex diseases such as type 1 diabetes, type 2 diabetes, coronary heart disease, hypertension, bipolar disorder, rheumatoid arthritis and Crohn's disease. For each disease type, genome sequence variations (single nucleotide polymorphisms or SNPs) are gathered by comparing the genetic make-up of the case group (disease) and the control group (normal). This allows identification of many SNPs and genes showing evidence of association with disease susceptibility [109,124,125,126]. In addition, the ENCODE project[7] (ENCyclopedia Of DNA Elements) aims to identify all functional elements in the human genome sequence and the 1000 Genome Project[8] aims to construct the most accurate human genetic variation map to support disease studies. Table 1-3 shows a selected list of database that compiles genetic variations available in the public domain. These international efforts look very promising, but there is still a long way to go to establish a complete understanding of disease mechanisms especially at the molecular level. In Chapter 5, I demonstrate how our understanding of molecular evolution learnt from amino acid replacements can help identify genetic variations related to disease by exemplifying an analysis of SNPs identified from genome-wide association study of Type 1 Diabetes.

[6] http://www.wtccc.org.uk/
[7] http://www.genome.gov/10005107
[8] http://www.1000genomes.org

Table 1-3 A compiled list of database for human genetic variations and diseases

Name	URL	Summary	Reference
HGMD	http://www.hgmd.cf.ac.uk/ac/index.php	A comprehensive core collection of data on published germline mutations in nuclear genes underlying human inherited disease.	[127]
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/	A free public archive for genetic variation within and across different species developed and hosted by the National Center for Biotechnology Information (NCBI) in collaboration with the National Human Genome Research Institute (NHGRI).	[128]
HGVbase (HGBASE)	http://www.hgvbaseg2p.org/	A catalogue of all known sequence variations (particularly single nucleotide polymorphisms (SNPs)) as a non-redundant set of records, which presents each variant in the context of its physical relationship to the nearest human gene.	[129,130]
ProTherm	http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/protherm.html	A collection of numerical data of thermodynamic parameters such as Gibbs free energy change, enthalpy change, heat capacity change, transition temperature etc. for wild type and mutant proteins, which are important for understanding the structure and stability of proteins.	[131]
ASEdb	www.asedb.org	A repository for energetics of sidechain interactions determined by alanine-scanning mutagenesis.	[132]
p53	http://www.bioinf.org.uk/p53/	Integrating mutation data and structural analysis of p53 tumor-suppressor protein.	[133]
G6PD	http://www.bioinf.org.uk/g6pd/	An integration of up-to-date mutational and structural data of human Glucose-6-phosphate dehydrogenase (G6PD) from various genetic and structural databases (Genbank, Protein Data Bank, etc.) and latest publications.	[134]
MutDB	http://mutdb.org/	Annotation of human variation data with protein structural information and other functionally relevant information.	[135]
SNPper	http://snpper.chip.org/	A web-based application designed to facilitate the retrieval and use of human SNPs for high-throughput research purposes.	[136]
ModSNP (SwissVar)	http://expasy.org/swissvar/	A portal to search variants in Swiss-Prot entries of the UniProt Knowledgebase (UniProtKB), and gives direct access to the Swiss-Prot Variant pages.	[137,138]
COSMIC	http://www.sanger.ac.uk/genetics/CGP/cosmic/	To store and display somatic mutation information and related details and contains information relating to human cancers.	[139,140]
TopoSNP	http://gila.bioengr.uic.edu/snp/toposnp/	An interactive visualization of disease and non-disease associated non-synonymous single nucleotide polymorphisms (nsSNPs) and displays geometric and relative entropy calculations.	[141]
LS-SNP	http://salilab.org/LS-SNP/	A genomic scale, computational pipeline that maps human SNPs in NCBI's dbSNP database [128] onto protein sequences in the SwissProt/TrEMBL databases.	[142,143]

SAAPdb	http://www.bioinf.org.uk/saap/	Integration of information on Single Amino Acid Polymorphisms (i.e. structurally expressed SNPs and mutations) with analysis of the likely structural effects of these amino acid mutations.	[144]
SNPeffect	http://snpeffect.vib.be/	Annotations for both non-coding and coding SNP, as well as annotations for the SwissProt set of human disease mutations.	[145,146,147]
SNP@Domain	http://snpnavigator.net	A web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences.	[148]
T1DBase	http://t1dbase.org	A public website and database that supports the type 1 diabetes (T1D) research community.	[149]
PolyDoms	http://polydoms.cchmc.org/polydoms/	A database to integrate the results of multiple algorithmic procedures and functional criteria applied to the entire Entrez dbSNP dataset. In addition to predicting structural and functional impacts of all nsSNPs, filtering functions enable group-based identification of potentially harmful nsSNPs among multiple genes associated with specific diseases, anatomies, mammalian phenotypes, gene ontologies, pathways or protein domains.	[150]
DMDM	http://bioinf.umbc.edu/dmdm/	A database in which each disease mutation can be displayed by its gene, protein or domain location. DMDM provides a unique domain-level view where all human coding mutations are mapped on the protein domain.	[151]
DVGa	http://www.ebi.ac.uk/dgva/page.php	A public catalogue of the large-scale insertions, deletions, duplications and rearrangements that are found in the genomes of individuals within a species.	[152]
1000 Genome	http://www.1000genomes.org	The project aims to find most genetic variants that have frequencies of at least 1% in the populations studied by sequencing many individuals lightly.	[153]
WTCCC	http://www.wtccc.org.uk/	To exploit progress in understanding of patterns of human genome sequence variation along with advances in high-throughput genotyping technologies, and to explore the utility, design and analyses of genome-wide association (GWA) studies	[154]

1.3.3 Computational methods to assess genetic mutations

Early analyses of protein structure showed that single amino acid substitutions or mutations are often disease associated [111]. Several studies have focused on the relationships between somatic mutations in the human genome (especially those in protein kinases [155,156] and other signalling pathway proteins [157]), and various human cancers. Recently, systematic resequencing of the cancer genome has revealed the frequency of genetic changes that are responsible for lung, breast and colorectal cancer [158,159]. Those genetic variations responsible for disease are now catalogued and accessible through web sites such as ModSNP [137], SwissVar [138], COSMIC [139] and HGMD [127] (see Table 1-3 for details).

For most monogenic diseases, a single DNA variant resulting in an amino acid substitution is responsible for the disease by affecting protein stability [111]. Therefore, methods that predict the effect of mutations on protein stability are useful for identifying possible disease associations [115,160]. Indeed, several computer programs successfully identify protein mutations that affect protein stability (see Table 1-4). These computer programs are generally classified into four categories: (1) physical potential approach; (2) statistical potential approach; (3) empirical potential approach; and (4) machine-learning approach.

PoPMuSiC [161,162] is a program to predict protein mutant stability changes by performing all possible point mutations in a given protein. The program uses different combinations of database-derived potentials according to the solvent accessibility of the mutated residues. DFIRE (distance-scaled, finite ideal-gas reference) [162] is a reference state for distance-dependent structure-derived potentials. DFIRE was used to construct a residue-specific all-atom potential of mean force from known structures, and the potential not only recognises more native proteins from decoy sets but also shows significant improvement in predicting stability changes on mutants compared to other physical-based, potential-based methods such as CHARMM [163] and GROMOS [164].

FOLDEF [165,166] quantitatively estimates the importance of interactions contributing to the stability of proteins. The program uses protein structure information at the atomic level, and takes into account various energy terms such as van der Waals interactions, solvation energy and electrostatic potential. The energy terms were balanced using empirical data obtained from protein engineering experiments. When compared to statistical potential-based method such as PoPMuSiC, FOLDEF produced better predictions for buried residues in which the effects of atomic interactions play dominant roles in stabilising protein structure. On the other hand, the statistical methods better describe thermodynamic properties of protein surface and show better performance for the impact of mutation of exposed residues [117].

MUpro [167] is a machine-learning approach based on support vector machines (SVMs) to predict the stability changes for single site mutations. MUpro first predicts whether a mutation will increase or decrease the stability of protein structure, then it predicts the stability change resulting from single site mutations. MUpro uses various sequence and structure information as input features, and the method was trained and tested against experimental mutation data from ProTherm database [131]. I-Mutant 2.0 [168] is another SVM-based tool for the prediction of protein stability changes upon single point mutations. I-Mutant 2.0 is a descendant of I-MUTANT [169] which is based on a neural network that can be also used to predict whether a mutation is stabilizing or destabilizing. I-Mutant 2.0 can predict the direction and the $\Delta\Delta G$ value of the protein stability changes upon single point mutation only from the protein sequence. AUTO-MUTE [170] is a combined approach to predict stability changes in protein mutants based on a four-body, knowledge-based and statistical contact potential, and machine-learning techniques.

Recently, the effect of mutations on the affinity of protein–protein interaction has been widely reviewed [171]. However, a systems approach is required to predict functional effect in the context of complex interaction networks. For this reason, there have been several efforts at interrogating genetic variations to understand their effects on protein structures and interaction network [143,145,172,173].

Table 1-4 Computer software and web applications to study the effects genetic mutations and disease associations

Name	URL	Summary	Reference
SDM	http://mordred.bioc.cam.ac.uk/~sdm/sdm.php	A statistical potential energy function developed to predict the effect that mutations on the stability of proteins.	[174]
PopMuSiC	http://babylone.ulb.ac.be/popmusic/	A statistical potential approach for the computer-aided design of mutant proteins with controlled stability properties. It evaluates the changes in stability of a given protein or peptide under single-site mutations, on the basis of the protein's structure.	[161,175]
SIFT	http://sift.jcvi.org/	An algorithm taking a query sequence and using multiple alignment information to predict tolerant and deleterious substitutions for every position of the query sequence.	[176]
DFIRE	http://sparks.informatics.iupui.edu/yueyang/DFIRE/dDFIRE-service	Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction.	[162]
FOLDEF	http://foldx.crg.es/	A computer algorithm to provide a fast and quantitative estimation of the importance of the interactions contributing to the stability of proteins and protein complexes using an empirical potential approach.	[165,166]
Polyphen	http://genetics.bwh.harvard.edu/pph/	A tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.	[173,177]
I-mutant	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant/I-Mutant.cgi	A neural network method that can be used to predict whether a mutation is stabilizing or destabilizing.	[169]
Panther	http://www.pantherdb.org/tools/	A library of protein families and subfamilies derived by the use of Hidden Markov Model (HMM) techniques indexed by a vocabulary of more than 500 biological functional terms (aka. subPSEC).	[178]
GROMOS	http://www.igc.ethz.ch/GROMOS/index	A force field for molecular dynamics simulation.	[164]
I-mutant 2.0	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi	SVN (Support Vector Machine) version of I-mutant.	[168]
PHD-SNP	http://gpcr.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP.htm	A decision tree with the SVM-based classifier coupled to the SVM-Profile trained on sequence profile information.	[179]
nsSNPAnalyzer	http://snpanalyzer.utmem.edu/	Web-based software which extracts structural and evolutionary information from a query nsSNP and uses a machine learning method called Random Forest to predict the nsSNP's phenotypic effect (the web is down at the time of this writing).	[180]
Pmut	http://mmb2.pcb.ub.es:8080/PMut/	Computer software aimed at the annotation and prediction of pathological mutations by retrieving a series of structural parameters such as volume parameters, secondary structure propensities, hydrophobicity descriptors and sequence potential, among others.	[181]

Mupro	http://mupro.proteomics.ics.uci.edu/	A machine-learning approach based on support vector machines (SVMs) to predict the stability changes for single site mutations.	[167]
CUPSAT	http://cupsat.tu-bs.de/	A tool to predict changes in protein stability upon point mutations.	[182]
FastSNP	http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp	An web-based application which prioritizes SNPs according to twelve phenotypic risks and putative functional effects, such as changes to the transcriptional level, pre-mRNA splicing, protein structure, etc.	[183]
SNPs3D	http://www.snps3d.org/	A website which assigns molecular functional effects of non-synonymous SNPs based on structure and sequence analysis.	[184]
ERIS	http://troll.med.unc.edu/eris/login.php	The Eris server calculates the change of the protein stability induced by mutations (ΔΔG) utilizing the recently developed Medusa modelling suite.	[185]
SAPRED	http://sapred.cbi.pku.edu.cn/	An automatic pipeline to predict the disease-association of SAPs using several novel attributes such as Structural Neighbor Profile and Nearby Functional Sites, in addition to incorporating other well-known attributes such as Residue Frequency and Conservation.	[186]
stSNP	http://ilyinlab.org/StSNP/	The structure SNP (StSNP) web server compares structural nsSNP distributions in many proteins or protein complexes. StSNP enables researchers to map nsSNPs onto protein structures by comparative modelling of structure with nsSNPs by MODELLER (http://salilab.org) and visualize their structural locations by using the multiple structure-sequence viewer Friend. Pathway information is provided from KEGG database.	[187]
SNAP	http://snap.humgen.au.dk/views/index.cgi	A sequence analysis web server providing a simple but detailed analysis of human genes and their variations.	[188]
AUTO-MUTE	http://proteins.gmu.edu/automute/	A combined approach to predict stability changes in protein mutants based on a four-body, knowledge-based and statistical contact potential, and machine-learning techniques.	[189]
Bongo	www.bongo.cl.cam.ac.uk/Bongo/	A Graph theoretic measure for estimation of structural and pathological impacts of non-synonymous SNP.	[190]
Omidios (SeqProfCod)	http://sgu.bioinfo.cipf.es/services/Omidios/	The Omidios web site takes a query SWISS-PROT id and searches for all annotated and predicted protein variants (nsSNP).	[191]
F-SNP	http://compbio.cs.queensu.ca/F-SNP/	It provides integrated information about the functional effects of SNPs obtained from 16 bioinformatics tools and databases.	[192]
CHARM	http://www.charmm.org/	A force field for molecular dynamics as well as the name for the molecular dynamics simulation and analysis package associated with them.	[163]

1.4 Thesis outline

In this thesis, I address structural and functional restraints that shape and affect the occurrence of amino acid substitution (or conservation) from the perspective of protein evolution and apply the general rules of amino acid replacement into disease-association study. This thesis describes how the knowledge learnt from protein evolution can help our understanding of genetic variations underlying disease aetiology.

In Chapter 2, I describe how the description of amino acid replacement could be improved by discriminating local structural environments from the following four categories of functional restraints: protein-protein interactions, protein-nucleic acid interactions, protein-ligand interactions and catalytic activity of enzymes. In Chapter 3, I seek to answer the following questions: 1) what determines the replacement of amino acids within a group of proteins presumably descended from a common ancestor, 2) could we measure the extent of their contributions and prioritize them? To address these questions, I focus on local structural environments (see Figure 1-1) of amino acids as major restraints on the possible substitutions of amino acids during protein evolution. In Chapter 4, I describe structural and functional restraints that shape the occurrence of single amino acid variations in human proteins. I try to identify differences in amino acid variations from the following three categories: i) Mendelian disease-related variants, ii) neutral polymorphisms and iii) cancer somatic mutations. In Chapter 5, as an extension of the previous chapter, I focus on a specific example of a complex disease – type 1 diabetes (T1D) – and present an analysis of genetic variations related with the disease. The genetic variations, which are presumably responsible for T1D, are from the group of Professor John Todd[9], Cambridge Institute of Medical Research, and consist of 355 SNPs. I exemplify how the understanding of structural and functional restraints imposed on proteins can help identify genetic variations associated with a disease. In Chapter 6, I introduce a web-based database system SAMUL, which houses structural and functional annotations of amino acid residues and their variants, which have been

[9] http://www-gene.cimr.cam.ac.uk/todd/index.html

the basis in this research. Lastly in Chapter 7, I discuss importance of maintaining the function of a protein and its role in restraining amino acid substitutions, especially where molecular recognition is crucial such as in enzyme active sites. Then, I summarize conclusions of my research and discuss limitations and future directions.

Chapter 2

Discarding Functional Residues from the Substitution Table Improves Predictions of Active Sites within Three-Dimensional Structures

Identification of residues responsible for a specific function of a protein can provide clues about the mechanism of action. Computational approaches to identifying functional residues have emerged as low cost alternatives to experimental methods by providing fast and large-scale analyses. Moreover, the demand for such approaches is increasing as more sequences become available from genome sequencing projects. In this chapter, I focus on the use of CRESCENDO to identify functional residues in proteins of known structure by comparing the amino acid substitutions observed in a family of proteins with those predicted on the basis of the protein structure. CRESCENDO uses Environment Specific Substitution Tables or ESSTs which define the way that accepted amino acid substitutions are influenced by the local structural environment. I describe how the calculation of ESSTs can be improved by using only amino acids that are not involved in catalytic activity, metal or ligand binding, nucleic acid or protein interactions and other molecular functions. My new substitution table can better describe the extent to which amino acid substitutions are under structural restraints. It should be of value in all applications of ESSTs, including their use in sequence-structure homology recognition, structure validation and structure prediction in addition to their use in identification of functional residues. These approaches should enhance the understanding of protein structure and function which is critically important in the post genomic era. Most of the material in this chapter has been published in PLoS Computational Biology[10] with the same title.

[10] *Gong S, Blundell TL (2008) Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. PLoS Comput Biol 4: e1000179.*

2.1 Introduction

Orthologous protein families are assumed to have diverged from a common ancestor, mainly by accepting mutations that are selectively neutral. The rate of evolution [9] is assumed to be constant over evolutionary time [194,195] and so evolutionary distances can be measured by analysing the substitutions of amino acids. The degree of conservation and the nature of substitutions of amino acids will be under many evolutionary restraints. One of those is dependent on the need to retain the protein tertiary structure and usually expressed as a tendency to maintain the local structural environments of individual amino acids [100].

An ESST[11] describes the substitution of amino acids in terms of a set of structural environments that restrict the allowable substitutions [88,89]. By defining the local structural environment of amino acid residues (secondary structure, solvent accessibility and formation of hydrogen bonds), distinct patterns of substitutions have been observed [89,196]. Environment-specific substitution tables store these substitution data quantitatively in the form of probabilities and therefore provide information about the existence of each amino acid in a particular environment and the probability of its being substituted by any other amino acid.

The ESST was improved and updated by Shi *et al.* [197] in 2001 by the use of the following features: 1) a clustering scheme to correct sampling bias, 2) a smoothing procedure to correct data sparsity, 3) using only high resolution structures in the alignments as a source of substitution matrices and 4) reduction of the bias caused by non-structural restraints. The last feature was designed to separate functional restraints from structural restraints when generating ESSTs. Because ESSTs take into account only structural environments, substitutions where the amino acids are conserved for functional reasons should not be counted in the calculation of matrices. Shi *et al.* took two kinds of functional residues into account in order to eliminate non-structural restraints that may cause a bias in the ESST. They were 1) residues involved in domain-

[11] http://samul.org/ESST

domain interactions and 2) those interacting with ligand. Such residues were masked in the alignment files and were not taken into account in the substitution counts. However, the masking appeared to have very little impact on the performance of FUGUE, a computer program for recognising distant homologues by sequence-structure comparison [197]. Chelliah *et al*. [198] further developed ESSTs by introducing functional restraints, particularly in enzymes, on amino acid substitutions as a new environment in addition to the 64 structural environments. They measured the Euclidean distance between every amino acid and the known functional residues and compared the degree of conservation in terms of the proximity with the functional residues. Their ESST, known as the function-dependent ESST, showed improvements in sequence-to-structure homology recognition.

In this chapter, I investigate the impacts of various functional restraints on the conservation of amino acids in three-dimensional structures. The functional residues are divided into four categories according to whether they are involved in 1) protein-protein interaction, 2) protein-nucleic acid interaction, 3) protein-ligand interaction and 4) catalytic reaction at enzyme active sites. Such residues will be under greater pressure to be conserved throughout the evolution process where they remain critically important to the activity of protein and thus the selective advantage of the organism. The degree of functional residue conservation is measured by masking the locations in the alignment file and then discarding them in the calculation of substitution probabilities. The substitution models are compared with the non-masking model, which counts those functional residues in the calculation of substitution probabilities. Also, relative contributions of four categories of functional residues are measured by making several masking tables in combinatorial fashion. The substitution models are tested by performing computational experiments using CRESCENDO [100], which is a program predicting functional residues from known three-dimensional structures of proteins and which should be more sensitive to the accuracy of the predicted substitution tables than FUGUE [197]. I show that the new ESST can find 16% more functional residues compared with the ESST of Shi *et al. [197]* for the same test-set. The new ESST is different from previous ones in that it covers a broader range of protein families, takes

into account more three-dimensional structures and considers a wider variety of functional residues that may bias amino acid substitution patterns.

2.2 Results and Discussion

2.2.1 Locating Functional Residues in Three-Dimensional Structures

Four categories of functional residues are considered in this study (Table 2-1). The first category of functional residues comprises catalytic residues of enzyme active sites, which are strongly conserved in orthologous families and often across superfamilies. CSA [199] and "ACT_SITE" records in UniProt [200] were used. The Catalytic Site Atlas (CSA) is a database of enzyme active sites and catalytic residues of enzymes whose 3D structures are available. It provides two types of entries: 1) original hand-annotated entries derived from the primary literature and 2) entries homologous to one of the original entries by sequence similarity. Only the hand curated entries were taken into account for reasons of reliability. The second category comprised amino acids involved in protein-protein interactions. Data concerning protein interactions were retrieved from InterPare [201] which is a database for interacting interfaces between protein domains. InterPare uses SCOP [37] for a domain definition and detects interacting domain pairs if there are at least five pairs of residues that fall within 5 Å distance between two adjacent domains. Residues interacting with nucleic acids comprise the third category. BIPA [202] and "DNA_BIND" records in UniProt were used for this category. BIPA is a database for protein-nucleic acid interactions, which defines the atomic interactions using a distance threshold of 5 Å for van der Waals contacts, and HBPLUS [203] default options for hydrogen bonds and water mediated hydrogen bonds. The final category comprises the ligand-binding residues. For this information, the following UniProt feature annotations were used: "BINDING", "METAL", "NP_BIND", and "CA_BIND" (see Table 2-1 for details).

The data from InterPare, CSA and BIPA are based on three-dimensional structures of proteins. Hence, those functional residues can be easily identified and mapped into PDB entries using chain and residue numbers as unique identifiers. However, as the

functional feature annotations from UniProt are based on sequence information, they must be mapped into their corresponding PDB entries. For this purpose, I developed a mapping protocol named “double-map” to align a sequence from UniProt with that of PDB at the residue level. This mapping protocol is critically important as the exact functional residues from the structural alignment should be identified and masked. The detailed algorithm of double-map is described in Materials and Methods.

Table 2-1 Four Categories of Functional Residues Considered in this Study

The versions of CSA [199] and UniProt [200] were 2.2.7 and 12.2, respectively. InterPare [201] was based on SCOP [37] version 1.71. The "Feature Identifier" is only for UniProt annotations.
(A: all masking, B: no protein-protein interaction, C: no active sites, D: active-site only)

Functional Category	**Database**	**Feature Identifier**	**Description**	**Masking Type**				**URL**
				A	**B**	**C**	**D**	
Protein-protein Interaction	**InterPare**	N/A	Database of domain-domain interaction interface	√		√		http://interpare.net
Catalytic activity	**CSA**	N/A	Database documenting enzyme active sites and catalytic residues in enzymes of 3D structure	√	√		√	http://www.ebi.ac.uk/thornton-srv/databases/CSA/
	UNIPROT	**ACT_SITE**	Amino acid(s) involved in the activity of an enzyme	√	√		√	http://www.uniprot.org
Protein-nucleic acid interaction	**BIPA**	N/A	Database of protein-nucleic acid interactions	√	√	√		N/A
	UNIPROT	**DNA_BIND**	Extent of a DNA-binding region	√	√	√		http://www.uniprot.org
Protein-ligand interaction	**UNIPROT**	**BINDING**	Binding site for any chemical group (co-enzyme, prosthetic group, etc.)	√	√	√		http://www.uniprot.org
		CA_BIND	Extent of a calcium-binding region	√	√	√		
		NP_BIND	Extent of a nucleotide phosphate-binding region	√	√	√		
		METAL	Binding site for a metal ion	√	√	√		

2.2.2 Structure Alignments and New Environment Specific Substitution Table

The new Environment Specific Substitution Table (ESST) was built based on the alignments of three-dimensional structures of proteins that belong to the same protein family. The PDB database was used as a source for the three-dimensional structures of proteins and SCOP as the definition of protein families and domains. SCOP version 1.71, which was used in this study, classifies 3004 families and 75930 domains from 27599 PDB entries. For each SCOP family, domains were clustered with sequence identity of 80% or more, after pre-processing the structure data (see Materials and Methods for details). Within a cluster defined in this way, a structure having the best resolution was selected as a representative for the structure alignments. This process yielded 1187 SCOP families having 5833 domains from 4309 PDB entries. These final alignments, which are shown as "ALL" in the matrix type of Table 2-2, were used as a source for the calculation of substitution tables.

Table 2-2 shows 17 ESSTs and compares the numbers of structures and the functional residues masked from the alignments. The four matrix types, OLD, ENZ, NOENZ and ALL, differ in the alignment source. "OLD" is based on the 177 HOMSTRAD families, from which the ESST of Shi *et al.* [197] was derived. "ENZ" is for the 221 enzyme-specific SCOP families whose members contain at least one "ACT_SITE" residue or CSA hand-curated entry. "NOENZ", the opposite of "ENZ", does not contain any "ACT_SITE" annotations or CSA entries at all. These two matrix types are prepared in order to assess the effect of alignment sources on the substitution patterns of amino acids. "ALL" is based on 1187 SCOP families described above. SCOP families that belong to ENZ and NOENZ are subsets of the ALL type and do not overlap as they include different SCOP families. Each matrix type is further divided into several subtypes (A, B, C and D) that differ in the masking sources of functional residues (see Table 2-1). This is to investigate the effect of a specific category of functional residues by comparing the differences in the substitution patterns. For example, the effect of masking enzyme active sites can be measured by calculating the difference between two

matrices D and X, because X does not mask any functional residues whereas D masks only active site residues. I made random-masking models (R), in order to assess the value of masking models in benchmarking the new ESSTs. The new ESSTs mask more functional residues than the ESST (J) of Shi *et al.*, because the models take into account a broad range of structural families and functional residues. ESSTs and structure alignments in Table 2-2 are also available from http://samul.org/ESST.

Table 2-2 17 ESSTs and the Number of Functional Residue Masked from the Alignments.

New ESSTs were based on the structure alignments of SCOP families [37]. ENZ is 221 enzyme-specific SCOP families which contain at least one ACT_SITE annotation of UniProt [200] or hand-curated CSA entry [199]. NOENZ is the opposite of ENZ. NOENZ does not even contain the predicted entries of CSA. ALL is the final alignment source obtained from the filtering process (see Materials and Methods). Note that ENZ is not an absolute complement of NOENZ; ENZ does not include any predicted active site from the CSA. Hence, ENZ and NOENZ do not add up to ALL. The masking sources of A, B, C and D are in Table 2-1. X is for non-masking and R is for random-masking. R is set as a control to see the significance of removing functional residues from the substitution models. The ESST of Shi *et al.* (OLD-J) [197] is based on 177 HOMSTRAD families, which consist of 706 structures. And which masks 2,048 resides involved in (1) interactions with heteroatoms and (2) domain-domain interactions. OLD-X and OLD-R is non-masking and random-masking model of OLD-J.

Alignment Source	Number			Matrix Type	Masking Type	Masking residues[b]	%Mask[c]
	family	structure	residue[a]				
HOMSTRAD	177	706	146,437	**OLD**	X	0	0.00
					J	2,048	1.40
					B	4,601	3.14
					R	4,601	3.14
SCOP	221	902	235,588	**ENZ**	X	0	0.00
					A	37,808	16.05
					B	6,195	2.63
					C	36,265	15.39
					D	1,615	0.69
					R	37,808	16.05
	566	2,556	384,618	**NOENZ**	X	0	0.00
	1,187	5,833	1,096,027	**ALL**	X	0	0.00
					A	198,411	18.10
					B	21,830	1.99
					C	191,377	17.46
					D	1,840	0.17
					R	198,411	18.10

[a] number of all residues

[b] number of masking residues

[c] %Mask = number of masking residues / number of all residues * 100

2.2.3 Differences between Substitution Tables: the Effects of Alignment Source and Masking

The new ESSTs differ from those of Shi et al. [197] in the source of structure alignments and the categories (and the number) of functional residues removed from the alignments. The differences between 17 substitution tables were measured and investigated in terms of 1) the conservation probability of amino acids (P_{CONS}) and 2) the distance (DIST) between ESSTs (see Materials and Methods). I first looked at the different sources of structure alignments to assess their effects on the amino acid conservation in the substitution table. For this purpose, the non-masking models (X) from four alignment sources (OLD, ENZ, NOENZ and ALL) were compared. Figure 2-1A plots the P_{CONS} of 21 amino acids (P_{CONS} in Table 2-3). The conservation probability in the figure is averaged over the diagonal entries (i.e. those amino acids that are not substituted) from 64 ESSTs for each model. The overall degree of conservation is 28.93, 29.10, 32.08 and 36.73% for NOENZ, ALL, ENZ and OLD respectively (see Table 2-3 for details). All the amino acids in OLD-type are more conserved than those of ALL-type, and the number of structures and families in the alignment may affect the P_{CONS}. In addition, the definition of protein families and domains of HOMSTRAD is more stringent than those of SCOP. This will make the sequences less divergent and the alignments more conserved. Table 2-4 shows the P-values measured by chi-square test to see how significantly the amino acid conservation probabilities (shown in Table 2-3), are different each other. Most of matrices within ENZ-NOENZ and OLD-ALL pairs are significantly different each other, whereas matrices within the same matrix-type are not least different. Similarly, the distance of substitution tables (see Table 2-5) shows that NOENZ and ENZ are the most distant (507) among four tables and NOENZ and ALL are the closest. This is clear as NOENZ and ENZ do not share any families but all the families in NOENZ belong to ALL. The farthest substitution tables (highlighted in bold in Table 2-5) agree well with the least P-value scores (see Table 2-4) within a pair of matrix-type. Figure 2-1A shows that amino acids R, K, H and S of ENZ-type are more conserved than those from NOENZ by 17, 14.2, 8.5 and 7%, respectively. However, C and W from ENZ are less conserved than those of NOENZ by 24% and 9%.

Table 2-3 Probability of Residue Conservation

For each masking type, the diagonal entries (not substituted entries) are averaged over 64 ESSTs. Note that there are 21 amino acids (J for cysteine and C for half-cysteine) in this table. See Table 2-2 for the definitions of 'Matrix types' and 'Masking types'.

Matrix types		OLD				ENZ						NOENZ	ALL					
Masking types		X	J	B	R	X	A	B	C	D	R	X	X	A	B	C	D	R
Amino acids	A	29.84	30.08	29.84	29.72	25.84	26.92	26.02	26.89	25.88	25.98	23.66	23.36	23.90	23.25	23.89	23.38	23.33
	C	76.94	77.00	63.29	77.28	60.51	60.77	59.41	61.32	59.77	60.11	84.57	75.90	75.95	75.98	76.04	75.81	74.33
	D	44.89	44.52	41.34	45.07	41.38	38.16	37.91	40.57	38.91	42.31	35.84	38.39	36.01	35.32	37.03	37.39	38.65
	E	32.36	32.17	29.16	32.15	33.27	31.95	31.27	33.72	31.22	33.36	27.83	29.66	28.72	27.92	29.66	28.51	30.10
	F	33.99	32.71	34.14	33.88	26.07	25.60	25.26	25.68	26.09	26.19	25.00	23.53	23.96	23.32	24.07	23.54	23.97
	G	53.50	53.00	52.32	53.81	53.24	50.74	49.63	50.77	53.25	53.73	45.99	47.32	45.83	45.19	45.84	47.32	47.30
	H	38.72	32.06	28.39	38.90	33.31	32.36	31.20	33.88	31.20	34.29	24.83	24.78	23.30	22.66	24.14	23.85	25.60
	I	26.61	26.94	26.54	26.32	23.25	23.38	22.94	23.36	23.28	23.85	21.10	20.94	21.05	20.67	21.06	20.94	20.85
	J	31.33	15.68	17.45	32.03	15.20	11.86	11.30	15.10	11.84	14.95	16.79	14.72	9.31	9.37	11.32	13.36	14.30
	K	34.03	34.05	28.59	33.78	38.20	33.40	32.96	33.22	38.19	37.54	24.00	33.01	27.33	27.08	27.68	32.68	32.34
	L	36.04	35.89	36.10	36.08	31.36	31.99	31.10	31.95	31.39	31.81	30.41	29.25	29.86	29.16	29.87	29.27	29.54
	M	18.13	17.86	17.27	17.96	15.70	16.17	15.85	16.15	15.73	15.28	10.61	11.50	11.19	11.32	11.22	11.51	11.73
	N	30.16	29.96	28.19	30.33	30.97	31.56	30.65	31.74	30.79	29.92	22.36	25.93	25.69	25.13	25.77	25.88	26.02
	P	45.46	45.55	45.45	45.48	43.80	44.29	43.88	44.30	43.82	44.35	39.01	38.43	38.62	38.48	38.58	38.43	38.61
	Q	24.89	24.85	24.99	25.04	20.29	20.07	20.42	20.11	20.28	20.40	16.90	16.63	16.20	16.51	16.25	16.63	16.62
	R	35.70	34.70	33.74	35.54	40.00	40.42	40.05	40.51	39.87	40.46	22.97	30.45	29.82	30.17	29.96	30.36	31.42
	S	29.10	29.00	28.07	29.21	25.36	23.62	23.54	24.35	24.65	25.17	18.34	21.74	21.00	20.67	21.29	21.49	21.98
	T	30.08	29.95	28.94	30.00	24.88	23.21	22.78	23.31	24.87	24.60	21.79	22.62	22.38	21.84	22.40	22.61	22.87
	V	30.05	29.97	30.05	30.11	27.56	26.29	26.00	26.27	27.58	27.52	24.69	24.12	24.24	23.75	24.23	24.13	24.16
	W	49.62	49.68	50.04	49.43	34.52	36.31	34.39	36.27	34.56	36.34	43.62	32.81	34.15	32.83	34.16	32.84	33.37
	Y	39.95	39.63	39.55	40.15	29.07	29.07	28.50	30.07	28.23	29.52	27.11	25.92	25.18	25.68	25.64	25.59	25.71
Average		36.73	35.49	33.97	36.77	32.08	31.34	30.72	31.88	31.49	32.27	28.93	29.10	28.27	27.92	28.58	28.83	29.18

Table 2-4 A P-value matrix of chi-square test based on the residue conservation scores

The chi-squared test was used to measure how significantly the amino acid conservation scores, shown in Table 2-3, are different from each other (see Materials and Methods). The P-value ranges from 0, which says most significant (most different), to 1, least significant (no difference at all). P-values less than 0.05 (significantly different) are highlighted in red. Pairs of matrix-type are shaded alternately.

Matrix type [1]	Masking type [2]	OLD				ENZ						NOENZ	ALL					
		X	J	B	R	X	A	B	C	D	R	X	X	A	B	C	D	R
OLD	X	1.00000	0.98211	0.86918	1.00000	0.11628	0.03931	0.01292	0.14342	0.02839	0.15550	0.00023	0.00081	0.00002	0.00001	0.00012	0.00030	0.00127
	J		1.00000	0.99982	0.54597	0.56495	0.47848	0.28773	0.62653	0.42154	0.64783	0.00885	0.03881	0.00795	0.00322	0.01894	0.02576	0.05610
	B			1.00000	0.38656	0.47653	0.53937	0.40450	0.63999	0.40290	0.56446	0.01092	0.03465	0.01231	0.00585	0.02593	0.02586	0.05663
	R				1.00000	0.09888	0.03210	0.01034	0.12342	0.02286	0.13323	0.00020	0.00067	0.00002	0.00001	0.00010	0.00024	0.00105
ENZ	X					1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.00941	0.71244	0.26917	0.19664	0.41756	0.62295	0.82144
	A						1.00000	1.00000	1.00000	1.00000	1.00000	0.02737	0.77841	0.60591	0.50805	0.71072	0.76162	0.88586
	B							1.00000	1.00000	1.00000	0.99997	0.02343	0.83048	0.71500	0.64405	0.79254	0.83085	0.92100
	C								1.00000	1.00000	1.00000	0.02911	0.74433	0.38824	0.29816	0.55600	0.66258	0.85550
	D									1.00000	1.00000	0.00897	0.75222	0.47556	0.38277	0.58180	0.73397	0.85868
	R										1.00000	0.00991	0.63775	0.23013	0.15999	0.36792	0.54426	0.76586
NOENZ	X											1.00000	0.93222	0.96716	0.95399	0.98535	0.93527	0.91090
ALL	X												1.00000	0.99999	0.99997	1.00000	1.00000	1.00000
	A													1.00000	1.00000	1.00000	1.00000	0.99990
	B														1.00000	1.00000	0.99999	0.99977
	C															1.00000	1.00000	1.00000
	D																1.00000	1.00000
	R																	1.00000

(1: Matrix type in Table 2-2, 2: Masking type in Table 2-2)

Table 2-5 A distance Matrix of 17 ESSTs

The difference between ESSTs is measured by the distance defined in Materials and Methods. Within a pair of matrix-type (OLD, ENZ, NOENZ and ALL), two farthest distances are in a bold character. Pairs of matrix-type are shaded alternately.

Matrix Type[1]	Masking Type[2]	OLD				ENZ						NOENZ	ALL					
		X	J	B	R	X	A	B	C	D	R	X	X	A	B	C	D	R
OLD	X	0.0	170.7	**220.6**	33.8	464.0	481.6	**487.0**	465.6	481.6	460.6	**464.9**	466.1	489.2	**496.3**	476.9	474.7	459.5
	J		0.0	161.3	177.9	437.5	443.0	446.8	437.1	444.2	433.6	428.6	428.2	433.6	441.7	426.8	432.7	420.6
	B			0.0	**226.1**	430.5	427.4	426.1	425.4	432.0	425.7	435.5	435.5	432.5	438.6	427.8	437.5	424.5
	R				0.0	465.3	482.7	**488.4**	466.4	483.2	461.8	**465.5**	467.7	491.1	**498.2**	478.7	476.5	461.1
ENZ	X					0.0	145.8	133.0	124.8	74.9	84.3	**507.0**	340.4	356.1	365.5	345.0	348.2	322.6
	A						0.0	74.9	73.3	124.3	**147.6**	501.2	363.5	340.3	356.5	338.0	364.1	342.5
	B							0.0	107.2	104.7	**146.8**	492.4	344.3	325.8	334.8	324.2	344.1	324.5
	C								0.0	141.6	128.3	502.2	361.6	351.1	366.7	340.7	368.1	341.1
	D									0.0	109.6	505.3	342.2	343.8	353.4	340.6	343.0	323.7
	R										0.0	**508.3**	357.3	**367.8**	**378.9**	357.5	364.4	337.5
NOENZ	X											0.0	**310.8**	309.0	303.1	306.9	308.8	**315.9**
ALL	X												0.0	**147.0**	130.7	132.7	37.9	73.1
	A													0.0	70.8	44.3	136.2	**147.5**
	B														0.0	83.7	117.5	141.4
	C															0.0	132.9	134.6
	D																0.0	81.1
	R																	0.0

(1: Matrix type in Table 2-2, 2: Masking type in Table 2-2)

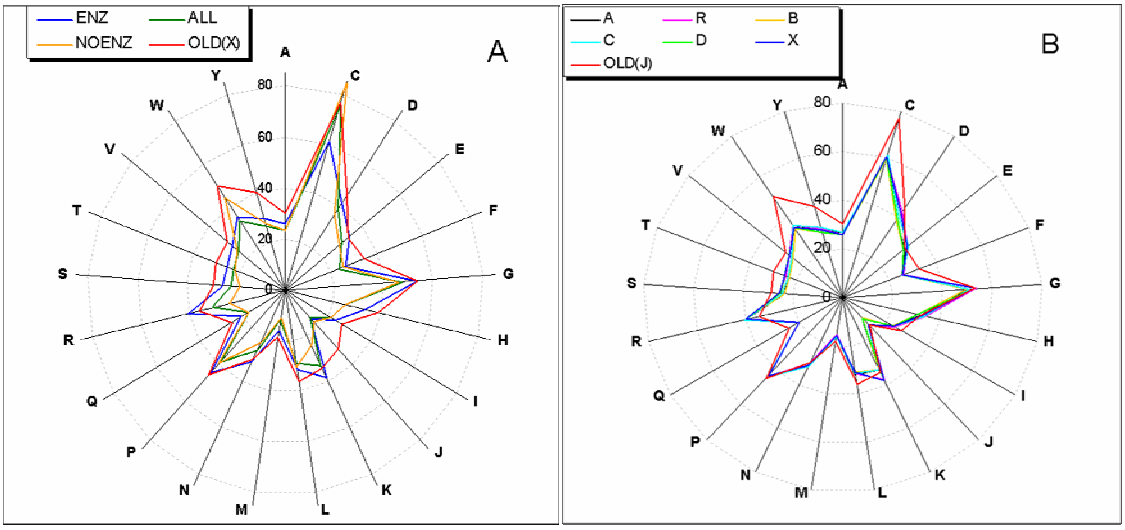


Figure 2-1 Probabilities of Residue Conservation for 21 Amino Acids

The probability of residue conservation (P_{CONS}) was averaged for the diagonal axis of substitution tables. **A**. P_{CONS} values of three matrix-types (ENZ, NOENZ and ALL) are compared with those of OLD. Non-masking models (X) were used for three matrix-types and OLD to see the effect of alignment source. (ENZ: enzyme-specific 221 SCOP families, NONENZ: non-enzymes, ALL: all the alignments, OLD: non-masking ESST of shi *et al.* [197]. See Table 2-2 for details)

B. Five masking tables and one non-masking table are compared with the ESST of Shi *et al.* [197]. Masking and non-masking tables are from the 221 enzyme-specific alignments (ENZ). Masking sources of A, B, C and D are listed in Table 2-1. (R: random-masking, X: non-masking)

Figure 2-1B shows the comparison of P_{CONS} values of amino acids from the same source of alignment (ENZ) but having different masking types (A, B, C and D) with those of non-masking (X), random-masking (R) and ESST of Shi *et al.* (OLD-J). Overall, the differences of P_{CONS} among the tables are less clear than the differences shown in Figure 2-1A. In addition, Table 2-5 shows that the distances (DIST) between tables of different masking types, but having the same alignment source, are smaller than the distances of tables from the different alignment sources. This explains why the variations of P_{CONS} and DIST between tables are more affected by the source of alignments than the masking sources. However, the relationship between P_{CONS} (or DIST) and the number of masking residues (%Mask) could be clearly understood by the Spearman's rank correlation between the two (see Table 2-6). Increasing the masking of functional residues (%Mask) from the alignments leads to smaller P_{CONS} values and greater differences as measured by DIST between the substitution tables. The

correlation between P_{CONS} and %Mask (-0.3) was not made more distinctive by removing residues involved in protein-protein interactions. A-type masks 13.4% and 16.9% many more residues than B-type in ENZ and ALL, respectively, where the discrepancies lie in the protein-protein interactions as B does not include InterPare as masking sources. However, the average P_{CONS} of A is bigger than B, although A masks much more residues than B. This becomes much clearer on looking at the P_{CONS} of A and D where the difference is in residues annotated as CSA and ACT_SITE. The P_{CONS} of D is bigger than A, although D masks many fewer residues than A. The result shows that the residues involved in protein-protein (or domain-domain) interactions are not as conserved as residues responsible for the catalytic activity of enzymes. From P_{CONS} of ENZ-D and ENZ-X (Table 2-3), which differ in active sites as the masking source, I observe that active site residues J, D, H and E are most conserved throughout enzyme families, where H is the most abundant amino acid annotated as ACT_SITE or CSA followed by D, E and J.

Table 2-6 Rank Correlation

Spearman's rank correlations were calculated between the variables of P_{CONS}, Z-score, SENS, DIST and %Mask. See Materials and Methods for the definition of Spearman's rank correlation. %Mask is from Table 2-2. Z-Score and SENS are from Table 2-8. DIST is from the first row of Table 2-5. P_{CONS} is from the bottom line of Table 2-3.

	P_{CONS}	**Z-score**	**SENS**	**DIST**	**%Mask**
P_{CONS}	1	-0.85	-0.93	-0.38	-0.30
Z-score		1	0.95	0.54	0.45
SENS			1	0.48	0.45
DIST				1	0.29
%Mask					1

(P_{cons}: average probability of residue conservation taken from Table 2-3, Z-score: average Z-score of 602 active sites, SENS: sensitivity, DIST: distance between two ESSTs, %Mask: percentage of discarded functional residues)

2.2.4 Benchmarking Design

The performance of the new ESSTs was benchmarked by using CRESCENDO [100], which is a program for predicting functional residues given a three-dimensional structure. The rationale behind CRESCENDO is to distinguish functional restraints from structural restraints, both of which give rise to the conservation of amino acids in the evolutionary process. For example, amino acids in the core region of a protein are conserved or conservatively varied in order to maintain an appropriate structure (and ultimately function) whereas the catalytic triad of a protease, such as CYS-HIS-ASP, is conserved to maintain the functional properties of the enzyme family. CRESCENDO quantifies the degree of amino acid conservation by measuring 1) the observed value based on the alignment to which a queried protein sequence belongs and 2) the expected value calculated by using ESST. Note that the first value reflects both structural and functional restraints, whereas the latter only reflects the structural restraints because ESST, by definition, only takes structural environments into account. The overall difference between the two is converted into Z-score (or CRESCENDO score) which can represent extra restraints - probably functional - on the process of evolution. Hence, the more accurate the ESST, the less good the agreement between the probabilities of conservation observed and that predicted on the basis of the structure of the protein alone. CRESCENDO can be a good benchmarking tool for the evaluation of new ESSTs, because more functional residues are masked than the old ESST. In addition, I could identify relative contributions of four masking resources on the performance of ESSTs. The benchmarking was designed to investigate the following two questions. (1) How well can a new ESST identify functional residues compared with the ESST of Shi *et al.* which is used currently as the default by CRESCENDO? (2) If there is any improvement, what makes the improvement?

From 221 enzyme-specific SCOP families for ENZ in Table 2-2, one third (73 SCOP families) were selected as a test-set and the rest were used to make benchmarking-ESSTs for ENZ. The test-set consists of 339 SCOP domains having 81,410 residues in total. Out of 81,410 residues, 602 residues are active sites (ACT_SITE or CSA), 11,917 residues are annotated by InterPare, 194 residues for nucleic-acid interactions and 1,348

residues are involved with ligand interactions. They are the true functional residues that need to be predicted using CRESCENDO in order to evaluate the performance of our new ESST. In the analysis I took only the first cluster as the predicted residues. The performance of the new ESST was compared with that of the old in terms of detecting functional residues. Note that, for both ENZ and ALL types, the 73 SCOP families in the test-set were removed from the original ESST. The benchmarking ESSTs were renamed as At, Bt, Ct, Dt, Rt and Xt to distinguish them from the original new ESSTs which are A, B, C, D, R and X, respectively. This was in order to make our benchmarking an unbiased blind test by removing sequences in the test-set which might affect the benchmarking results. In the case of OLD and NOENZ, the original masking types were used in the benchmarking process as they did not contain SCOP families in the test-sets. The test-sets and benchmark results are accessible from http://samul.org/ESST.

2.2.5 Performance of new ESSTs in Detecting Functional Residue

Table 2-7 shows the average Z-score of CRESCENDO for 602 active sites, 11,917 PPI residues, 194 residues for protein-nucleic acid interactions (PNI) and 1348 residues responsible for interaction with ligands (PLI) along with the P-values for the predicted residues. The P-value demonstrates that the Z-score of the predicted residues is different from the randomly selected residues with a 0.09 level of significance. In other words, the predicted residues of CRESCENDO are far from random within a 0.09 error rate. The Z-scores for all the residues (81,410) in the test-sets are compared with those of functional residues predicted by CRESCENDO. The average Z-score of all the residues is near zero, regardless of masking types, which means there are no differences between the probabilities of residue conservations observed in the alignments and those predicted by ESST. However, the Z-scores for 602 active sites range between 0.48 and 0.93 depending on matrix type and masking source. This observation suggests there are extra restraints that make the active sites more conserved in families of homologous proteins. The Z-scores of 1,348 PLI (Protein-Ligand Interaction, see Table 2-7) residues also imply that they are under restraints in addition to those arising from structure. On the other hand, the average Z-scores for PPI and PNI residues are much smaller than

that of 602 active sites. This may suggest that residues at protein-protein interfaces are under less strong restraints than residues responsible for the catalytic activity. However, there is strong evidence that sub-regions in protein interfaces – so called hot spots – are energetically more important and may be under stronger restraints in evolution [204,205].

Table 2-7 Z-score of CRESCENDO for Functional Residues

The average Z-scores are shown for four categories of functional residues in the test-sets: catalytic activity, protein-protein interactions, protein-nucleic acid interactions and protein-ligand interactions. The test-sets consist of 73 SCOP families, which is one third of SCOP families in ENZ (see Table 2-2).

Matrix Type	Masking Type†	Average Z-score						Ratio[g]	P-value[h]
		all[a]	predicted[b]	active site[c]	PPI[d]	PNI[e]	PLI[f]		
OLD	X	0.00063	1.396	0.480	0.0250	0.055	0.449	0.78	0.081
	R	0.00067	1.402	0.483	0.0249	0.052	0.450	0.79	0.080
	J	0.00062	1.410	0.612	0.0284	0.055	0.461	1.00	0.079
	B	0.00065	1.420	0.734	0.0274	0.059	0.490	1.20	0.078
ENZ	Xt	0.00060	1.387	0.635	0.0042	0.024	0.426	1.04	0.083
	Rt	0.00060	1.387	0.652	0.0067	0.025	0.431	1.06	0.083
	Ct	0.00063	1.413	0.734	0.0100	0.025	0.427	1.20	0.079
	Dt	0.00062	1.399	0.772	0.0078	0.051	0.428	1.26	0.081
	At	0.00063	1.423	0.858	0.0143	0.056	0.433	1.40	0.077
	Bt	0.00064	1.411	0.870	0.0086	0.068	0.447	1.42	0.079
NOENZ	X	0.00063	1.420	0.835	0.0046	0.099	0.508	1.36	0.078
ALL	Xt	0.00063	1.414	0.696	0.0085	0.068	0.489	1.14	0.079
	Rt	0.00064	1.415	0.771	0.0065	0.075	0.501	1.26	0.079
	Dt	0.00066	1.412	0.798	0.0055	0.078	0.495	1.30	0.079
	At	0.00064	1.433	0.860	0.0159	0.069	0.495	1.41	0.076
	Ct	0.00067	1.436	0.893	0.0155	0.077	0.515	1.46	0.076
	Bt	0.00068	1.435	0.936	0.0073	0.086	0.518	1.53	0.076

[a]Total number of residue from test-sets (81,410) [b]Residue predicted by CRESCENDO [c]Active-site residues (602)

[d]Protien-protein interaction sites (11,917) [e]Protein-nucleic acid interaction sites (194) [f]Protein-ligand interaction sites (1,348)

[g]Ratio of Z-score at the active site residues compared with that of OLD-J [h]P-value (right-tail) of the predicted residues

† Xt, Rt, Ct, Dt, At, Bt: bench marking ESSTs where the test-set are eliminated from X, R, C, D, A and D, respectively

In Table 2-8, the performance of 17 ESSTs is compared in terms of recognizing 602 active-site residues. SENS, SPEC and COV were measured using the ratios of TP (true positive), FP (false positive), FN (false negative) and TN (true negative) (see Materials and Methods for the definitions). The Z-score and SENS are plotted together in Figure 2-2; they are highly correlated having 0.95 Spearman's rank correlation score (see Table 2-6). As shown in Figure 2-2, the average Z-scores and SENS of non-masking (X) and random-masking (R) models are always less than those from masking-models (A, B, C and D) within the same matrix type. This clearly shows that the position of masking is significant and discarding the substitution counts of functional residues from the substitution table can increase the performance of CRESCENDO by making ESST less dependent on the substitution patterns of the residues under functional restraints. This result is clearer from the rank correlation (0.45) between %Mask and SENS in Table 2-6. In addition, the new masking models (A, B, C and D) outperform the ESST of Shi *et al.* (J) and even the non-masking model (ENZ-X, NOENZ-X and ALL-X) outperform J (see Figure 2-2 and Table 2-8) This can be explained in terms of P_{CONS} and SENS; the average P_{CONS} is highest in the order of J, followed by ENZ-X, ALL-X and NOENZ-X, but the performance (SENS) is exactly the reverse order of P_{CONS}. Figure 2-3A shows an example of predicting active sites of a SCOP domain d1evua4 (a domain in the A chain of PDB 1evu, [206] which is a cysteine proteinase containing three active site residues annotated by UniProt. Three active site residues (CYS-314, HIS-373 and ASP-396) could be identified only by ALL-type ESSTs (ALL-B and ALL-C), which are highly ranked in Figure 2-2. This is probably because P_{CONS} of ALL is lower than that of ENZ and OLD for the local environments of the three catalytic residues.

Table 2-8 Performance of 17 ESSTs on Detecting Active Sites

Out of 81,410 residues in the test-sets, 602 residues are annotated as "ACT_SITE" by UniProt [200] or CSA [199]. For those active sites, CRESCENDO [100] could either correctly predict (TP) or fail to predict (FN) (see text). Two active sites of 'd7odca1' (A chain of PDB 7odc), which is a SCOP domain in the test-sets, was discarded as of an internal error; hence, 600 active sites either in the TP or FN. The number of predicted residues is same as the sum of TP and FP for each ESST type. Note that residues only from the first cluster of predicted residues (rank 1) were considered in this analysis.

Matrix Type	Masking Type	TP	FP	FN	TN	SENS	SPEC	COV	F-measure
OLD	X	168	4832	432	75976	0.28	0.9401	0.0336	0.060
	R	168	4830	432	75978	0.28	0.9401	0.0336	0.060
	J	189	4877	411	75931	0.315	0.9395	0.0373	0.067
	B	219	4888	381	75920	0.365	0.9394	0.0429	0.077
ENZ	Xt	221	4942	379	75866	0.3683	0.9387	0.0428	0.077
	Rt	225	4968	375	75840	0.375	0.9384	0.0433	0.078
	Ct	240	4870	360	75938	0.4	0.9396	0.047	0.084
	Dt	248	4977	352	75831	0.4133	0.9383	0.0475	0.085
	At	264	4805	336	76003	0.44	0.9404	0.0521	0.093
	Bt	270	4984	330	75824	0.45	0.9382	0.0514	0.092
NOENZ	X	273	5234	327	75574	0.455	0.9351	0.0496	0.089
ALL	Xt	249	5283	351	75525	0.415	0.9345	0.045	0.081
	Dt	259	5285	341	75523	0.4317	0.9345	0.0467	0.084
	Rt	262	5246	338	75562	0.4367	0.935	0.0476	0.086
	At	273	5150	327	75658	0.455	0.9362	0.0503	0.091
	Ct	277	5136	323	75672	0.4617	0.9363	0.0512	0.092
	Bt	282	5187	318	75621	0.47	0.9357	0.0516	0.093

(TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative, SENS: Sensitivity, SPEC: Specificity, COV: Coverage)

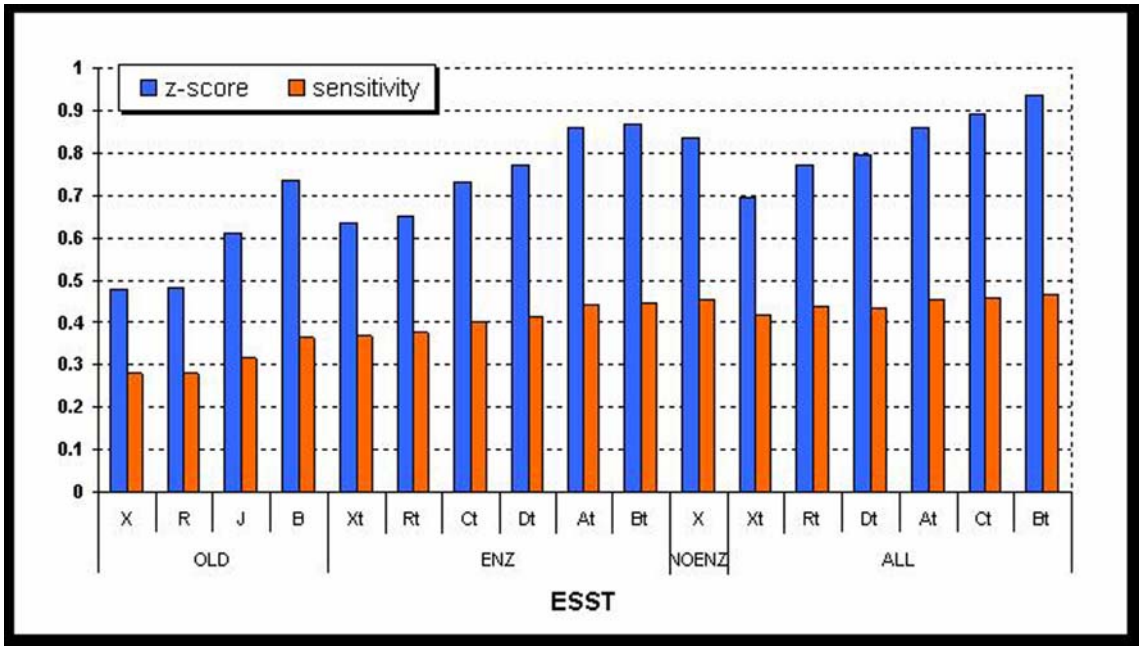


Figure 2-2 Performance of 17 ESSTs on Detecting Active Site Residues

Z-score (blue) and sensitivity (red) are plotted against 17 ESSTs. Z-score is averaged for 602 active-site residues in the test-sets (see text). Z-score and sensitivity (SENS) are highly correlated (0.95 in Spearman's rank correlation, Table 2-6). If any SCOP families in the test-sets are included in 17 ESSTs, they are removed from the ESSTs to avoid any bias. Those benchmarking ESSTs are marked by 't' (e.g. At, Bt, Ct and Dt) to distinguish from the original. Z-score and SENS of non-masking (X) and random-masking (R) tables are always lower than those of masking models (At, Bt, Ct and Dt) within the same matrix type (OLD, ENZ, ALL). All the masking-tables outperform the ESST of Shi *et al.* (J) [197].

Table 2-9 shows the recognition performance for 11,917 PPI residues with the same measurements (TP, FP, FN and TN) in Table 2-8. Four masking substitution tables of ALL-matrix could detect more PPI residues than that of Shi *et al.* (J), but not all tables in ENZ-matrix outperform J. Regardless of matrix types and masking types, the sensitivity (SENS) of detecting PPI residues is much lower than those for detecting active site residues. This probably arises from the average Z-score for PPI residues (see Table 2-7) which is close to zero, suggesting less strong evidence for extra restraints. Figure 2-3B shows an example of predicting PPI residues of a SCOP domain d1i7kb_ (B chain of PDB 1i7k, [207]) which is a ubiquitin conjugating (UBC) enzyme containing 14 residues interfacing with the A chain. Using ALL-A, CRESCENDO predicted 12 residues of which five were correct PPI residues (true positive, coloured in pink in Figure 2-3B). Among the nine missing residues (orange), PRO-30, SER-87, TYR-91, GLU-120 and LYS-121 were highly accessible (more than 50 $Å^2$) to solvent in the complex whereas five true positives had relatively small solvent accessible area (see Figure 2-3B for details). Thus, as expected, residues within the protein-protein interaction interface that are partially accessible are less conserved and more difficult to identify by CRESCENDO. Table 2-10 contains benchmark results for detecting residues interacting with nucleic acids and ligands. The sensitivity is better than the benchmarking results of recognizing PPI residues but still less than that of detecting active site residues. Figure 2-3C and Figure 2-3D show examples of predicting residues interacting with nucleic-acids and ligands, respectively (see legend to Figure 2-3 for details).

Table 2-9 Performance of ESSTs on Protein-Protein Interaction Residues

11,917 residues are annotated by InterPare [201] out of 81,410 residues in the test-sets. The definitions of TP, FP, FN, TN, SENS, SPEC, COV and F-measure are same as Table 2-8. Residues only from the first cluster of predicted residues were considered in this analysis.

Matrix Type	Masking Type	TP	FP	FN	TN	SENS	SPEC	COV	F-measure
OLD	B	931	4176	10986	65317	0.0781	0.8560	0.1823	0.1094
	R	934	4064	10983	65429	0.0784	0.8563	0.1869	0.1104
	X	939	4061	10978	65432	0.0788	0.8563	0.1878	0.1110
	J	939	4127	10978	65366	0.0788	0.8562	0.1854	0.1106
ENZ	At	906	4163	11011	65330	0.0760	0.8558	0.1787	0.1067
	Ct	908	4202	11009	65291	0.0762	0.8557	0.1777	0.1067
	Xt	921	4242	10996	65251	0.0773	0.8558	0.1784	0.1078
	Rt	925	4268	10992	65225	0.0776	0.8558	0.1781	0.1081
	Dt	960	4265	10957	65228	0.0806	0.8562	0.1837	0.1120
	Bt	973	4281	10944	65212	0.0816	0.8563	0.1852	0.1133
NOENZ	X	893	4614	11024	64879	0.0749	0.8548	0.1622	0.1025
ALL	Xt	930	4602	10987	64891	0.0780	0.8552	0.1681	0.1066
	Bt	953	4516	10964	64977	0.0800	0.8556	0.1743	0.1096
	Dt	963	4581	10954	64912	0.0808	0.8556	0.1737	0.1103
	Rt	980	4528	10937	64965	0.0822	0.8559	0.1779	0.1125
	Ct	1000	4245	10917	65248	0.0839	0.8567	0.1907	0.1165
	At	1003	4420	10914	65073	0.0842	0.8564	0.1850	0.1157

Table 2-10 Performance of ESSTs on the Residue Interacting with Nucleic-acids and Ligands

Out of 81,410 residues in the test-sets, 194 residues are annotated as DNA_BIND by UniProt [200] or BIPA and 1348 residues are annotated as either BINDING, CA_BIND, NP_BIND or METAL by UniProt (see Table 2-1 for the annotations). For those residues, if CRESCENDO [100] could correctly predict, they were counted as TP.

Matrix Type	Masking Type	PNI[a]		PLI[b]	
		TP[c]	SENS[d]	TP[c]	SENS[d]
OLD	B	20	0.1031	274	0.2033
	J	25	0.1289	261	0.1936
	R	22	0.1134	261	0.1936
	X	22	0.1134	253	0.1877
ENZ	At	24	0.1237	259	0.1921
	Bt	25	0.1289	265	0.1966
	Ct	25	0.1289	254	0.1884
	Dt	27	0.1392	261	0.1936
	Rt	22	0.1134	260	0.1929
	Xt	22	0.1134	259	0.1921
NOENZ	X	32	0.1649	279	0.2070
ALL	At	27	0.1392	281	0.2085
	Bt	34	0.1753	283	0.2099
	Ct	27	0.1392	286	0.2122
	Dt	40	0.2062	280	0.2077
	Rt	34	0.1753	277	0.2055
	Xt	35	0.1804	291	0.2159

[a]: Protein-nucleic acid interaction sites

[b]: Protein-ligand interaction sites

[c]: True Positive

[d]: Sensitivity

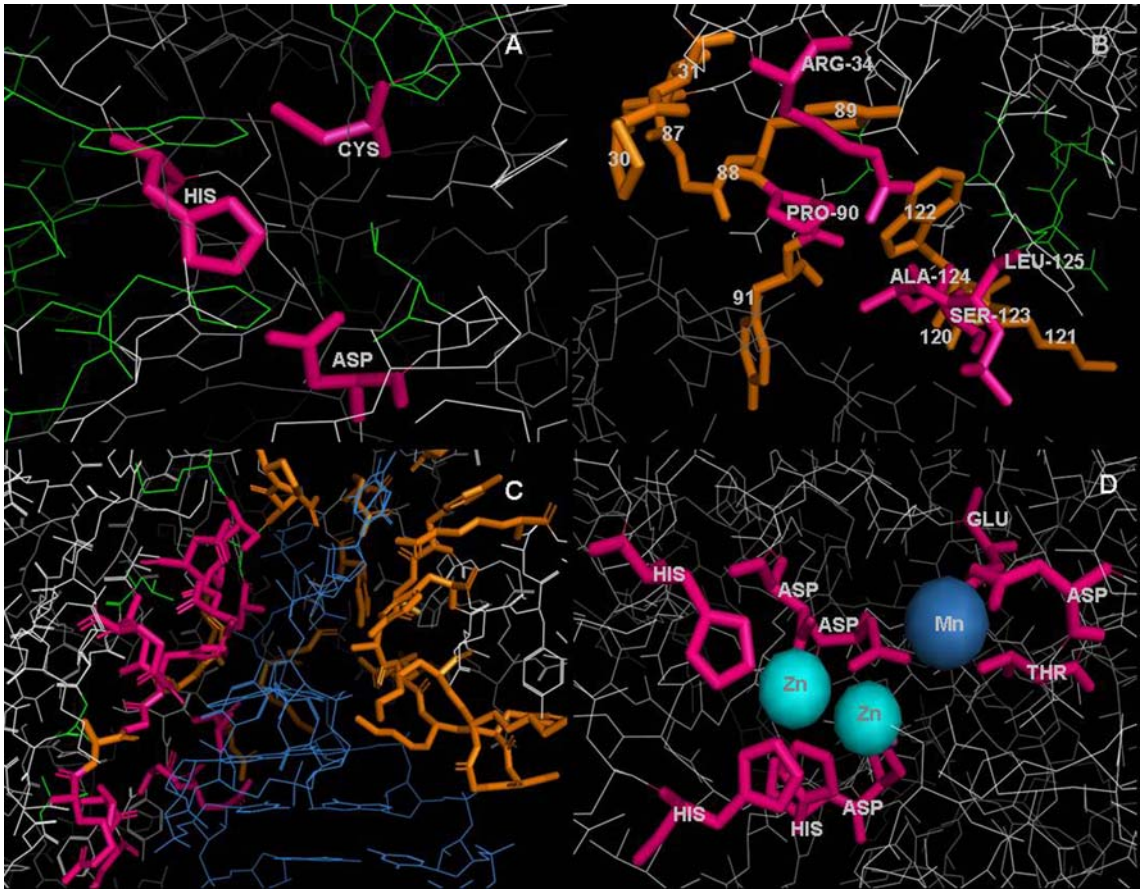


Figure 2-3 Predicting Four Categories of Functional Residues by CRESCENDO

Four case-studies of predicting functional residues are shown; A) active-sites, B) PPI (protein-protein interaction), C) PNI (protein-nucleic acid interaction, D) PLI (protein-ligand interaction). SCOP domains d1evua4 [206], d1i7kb_ [207], d1k8wa5 [208] and d1ed9a_ [209] were used for A, B, C and D, respectively. True positives (TP) are coloured in pink, false negatives (FN, missing residues) in orange and false positives (FP) in green. TP and FN are shown as sticks (bold-frame).

A. Cysteine protease. CRESCENDO predicted 27 residues as functional residues. All three (CYS-314, HIS-373 and ASP-396) catalytic residues were correctly identified. ALL-B type ESST (see Table 2-2) was used in this figure. FP (green) are clustered around the three real active sites (pink).

B. Ubiquitin conjugating (UBC) enzyme. 12 residues were predicted by CRESCENDO using ALL-A ESST. Five (coloured in pink) were correctly identified among 14 residues annotated as PPI residues. Interacting partner (A chain of 1i7k) is placed at the bottom and coloured in gray. The solvent accessible surface areas (SASA) for five TP are as follow; ARG-34 (35.64), PRO-90 (4.12), SER-123 (4.74), ALA-124 (0.55), LEU-125 (72.39). SASA for 9 FN are as follow; PRO-30 (77.26), VAL-31 (24.02), SER-87 (110.40), GLY-88 (16.05), TYR-89 (0.01), TYR-91 (58.29), GLU-120 (108.68), LYS-121 (113.96), TRP-122 (7.20). The SASA is from InterPare [201].

C. Pseudouridine synthase. BIPA [202] annotates 43 residues as PNI. 14 residues were TP (coloured in pink) among 20 residues predicted by CRESCENDO. ALL-D was used as ESST. DNA is coloured in blue.

D. Alkaline phosphatase. UniProt annotates 9 residues as metal-binding (METAL), which were all correctly identified by CRESCENDO among 30 predicted residues. ALL-B was used as ESST. ZN (zinc) and MG (magnesium) are coloured in cyan and blue, respectively.

2.2.6 The Effect of Discarding Residues Involved in the Protein-Protein Interactions

I found that the number of functional residues masked and discarded (%Mask) from the substitution table does not always guarantee the best performance (SENS) of ESST in detecting functional sites using CRESCENDO. The rank correlation between %Mask and SENS is 0.45 (see Table 2-6). Hence, it is very evident that masking-models outperform non-masking and the ESST of Shi *et al.* as described above. However the category of functional residues does matter and affects the performance. Figure 2-2 shows the performance of 17 ESSTs on the predictions of 602 active sites of the test-sets. Regardless of the alignment source, the performance (Z-score and SENS) of table B (no-PPI mask) is always better than table A (all mask), which means discarding PPI residues is not effective in the recognition performance of enzyme's active sites. In addition, OLD-B also outperforms OLD-J by 5% in the sensitivity, where the difference lies in the PPI residues as well. However, in the case of recognizing PPI residues, table A of ALL-matrix outperforms table B by 5.2% in terms of TP (Table 2-9). Interestingly, table C, which does not mask active sites, ranked as second highest and the performance of table D, which masks only active sites, is worse than the random-masking (R) substitution table (see Table 2-9). This result indicates that discarding PPI residues can increase the recognition performance of PPI residues but does not improve predictions of active sites of enzymes. This observation probably arises from the fact that the interfacial interactions differ in nature from those residues in catalytic sites and therefore masking of catalytic residues has little impact on those in interfaces.

2.2.7 Concluding Remarks

I have shown that discarding functional residues from the calculation of the substitution table improves the detection of functional residues when the new substitution table is used with CRESCENDO. I considered four categories of functional residues in this study (Table 2-1) and found that functional residues can be better predicted when the relevant category is discarded from the calculation of the substitution table. However, the performance of CRESCENDO for recognizing functional residues depends on the extent of amino acid conservation for the functional residues to be sought and how strong extra restraints – mainly non-structural – are imposed on the multiple sequence alignments from which the restraints are not considered in ESST. According to the benchmarking results studied here, enzyme active sites are under strong structural and functional restraints; hence they are relatively well predicted compared with amino acid residues responsible for protein-protein interaction, which are less conserved and very poorly predicted by CRESCENDO. Other interaction-site prediction methods using a support vector machine [210] and a random forest algorithm [211] seem to outperform CRESCENDO in terms of sensitivity and coverage (see 2.3.4), but direct comparison would not be appropriate as the benchmarking datasets are different and CRESCENDO is not only designed to predict PPI residues but functional sites in general. None the less, the new masking models outperformed non-masking, random masking and the old ESST (Shi *et al.*, [197]) not only in terms of true positives but also sensitivity.

As shown in Table 2-8 and Table 2-9, false positives (FPs) and false negatives (FNs) are relatively high compared with the number of true positives (TPs). The high FPs are expected to arise from the strict definition of functional residues. As shown in Figure 2-3A, FPs, coloured in green, are clustered around the catalytic triad (CYS-HIS-ASP) of the cysteine protease shown here. Some of these FP residues will be important for the local architecture of the active site and may even be buried; the substitutions accepted at these positions will therefore be restrained. Others will be directly involved in binding and positioning the substrate for catalysis. It has been previously shown that CRESCENDO identifies such residues in predicting the active site [100]. Furthermore the degree of residue conservation is significantly higher the closer the residues are to

the active site and that geometrical proximity to the known active sites can be considered to constitute a new environment of ESST [198]. Hence, due to the strict definition of functional residues, some of the FPs could not have been recognized as functional residues even though their structural and functional importance. A reason for some high FNs is that only the first cluster predicted by CRESCENDO were taken into account as positive results in the benchmark analysis; however CRESCENDO is expected to predict all regions under functional restraints and occasionally those critical for protein interactions, allostery, metal binding, post-translational modification and so on will be as conserved and score as high or higher than the active site residues. In addition, the annotations of functional residues might not be complete, which makes both FPs and FNs relatively high.

Other than CRESCENDO, there are several computational approaches to detecting possible functional regions of a protein in a fast and low-cost manner. Among them, the Evolutionary Trace method (ET), introduced by Lichtarge *et al.* [212], is widely used and very successful in identifying functional regions, for example of SH2, SH3, and DNA binding domains. ET differs from CRESCENDO in that it identifies conserved residues only on the protein surface and exploits the use of a phylogenetic tree to identify local patterns of conservation unique but distinct amongst different branches which constitute protein subfamilies. Hence, the performance of ET highly depends on the quality of a phylogenetic tree which is determined by a set of sequences to which a query protein belongs. If the sequences were recently diverged, the branch-specific conservation could not be detected because the substitutions were not accumulated enough to construct a reasonable phylogenetic tree. CRESCENDO does not explicitly use the phylogenetic tree (although it could well do so), but will also not work well if the degree of divergence is low. It will, however, gain from local conservation of buried residues in the active site, for example the threonine of the aspartic proteinase catalytic triad. It also gains from a careful definition of the expected substitution patterns in any local environment and for this the proper treatment of functional residues when deriving substitution tables is of critical importance.

2.3 Materials and Methods

2.3.1 Structure Alignments

New ESSTs were derived from the structure alignments of SCOP families [37]. Baton (D.F. Burke, unpublished, see Table 2-11), which is a successor of COMPARER [213], was used as a structure alignment program. The domain boundary and classification scheme of protein families were adopted from SCOP 1.71 as of this writing. PDB [214] was used as a source for protein three-dimensional structures. SCOP class F, which contains membrane and cell surface proteins, was not included in the alignment process as their amino acids can be in environments which differ from those in the cytoplasm. Also, non-canonical SCOP classes, H, I, J, and K, which are coiled-coil proteins, low resolution protein structures, peptides, and designed proteins, respectively, were removed from the alignment sources.

Table 2-11 Lists of Computer Programs and Databases used in this Study

Category	Name	Description	URL
Software	BATON	Structure alignments	http://www-cryst.bioc.cam.ac.uk/COMPARER
	CRESCENDO	Detecting functionally important residues	http://www.bioinf.manchester.ac.uk/crescendo
	SUBST	ESST calculation	http://www-cryst.bioc.cam.ac.uk/~kenji/subst
	JOY	Protein structure and alignment analysis	http://www-cryst.bioc.cam.ac.uk/~joy
	Kin3DCont	Making contour maps in kinemage format	http://kinemage.biochem.duke.edu/software/kincon.php
	EXONERATE	A generic tool for sequence alignment	http://www.ebi.ac.uk/~guy/exonerate/
	BL2SEQ	This tool produces the alignment of two given sequences	http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi
	CD-HIT	A program for clustering large protein database at high sequence identity threshold	http://bioinformatics.ljcrf.edu/cd-hi/
Database	CSA	Catalytic Site Atlas	http://www.ebi.ac.uk/thornton-srv/databases/CSA
	HOMSTRAD	Homologous Structure Alignment Database	http://tardis.nibio.go.jp/homstrad
	InterPare	A database server for protein interaction interfaces	http://interpare.net
	SCOP	Structural Classification of Proteins	http://scop.mrc-lmb.cam.ac.uk/scop
	UniProt	A comprehensive protein sequences and annotations	http://uniprot.org

To guarantee the best alignment quality, the following three filtering conditions were applied. (1) Filtering by resolution: NMR structures and structures having resolution worse than 2.5Å were not included in the alignment procedures. (2) Filtering by sequence identity: For each SCOP family, protein domains were clustered by running CD-HIT [215] with sequence identity of 80% or more. Within a cluster, a protein structure having the best resolution was selected as the representative. This is to remove any bias arising from the majority sequences of proteins in a SCOP family. (3) Filtering by sequence length: Within a SCOP family, the average sequence length is maintained by removing any domains having sequence below of (1-0.3)*mean-length and above of (1+0.3)*mean-length. Single member SCOP families were removed as they can not provide multiple alignments for the substitution calculation.

2.3.2 Mapping UniProt and PDB at Residue Level

UniProt [216] is a central hub for protein sequences, providing rich annotation on function and cross-references. However, it does not explicitly provide any three-dimensional structure information of proteins at the amino acid residue level. Hence, in order to harness both UniProt and PDB information, sequences in UniProt have been mapped to their corresponding structures in the PDB [55,217,218,219,220,221,222]. In January 2007[12], UniProt decided to reintroduce the initiation methionine (INIT_MET) into the full length sequence of UniProt proteins. This is a major change which gives rise to an increase in residue serial numbers by one. However at the time of this study (2007), no mapping methods, mentioned above, reflected changes of UniProt sequence into their mapping procedures, which lead to incorrect mapping between UniProt sequences and their corresponding proteins in PDB.

To take UniProt's upates into account in sequence-structure mapping, I developed a mapping protocol, "double-map", which aligns a sequence of UniProt with that of PDB at residue level. Three sequences are required for every PDB chain; 1) one from SEQRES record of a PDB file, 2) another from the residue (SEQ) in ATOM record of a PDB file, and 3) the third (SP) from the corresponding UniProt entry of a PDB chain.

[12] http://www.uniprot.org/news/2007/01/23/release

Double-map makes two alignments from the three sequences (so the name "double-map"). The first is an alignment between SEQ and SEQRES and the second is between SEQRES and SP. Using SEQRES as a reference, SP can be aligned with SEQ and the locations of UniProt residues can be mapped onto three-dimensional structures. Ideally, the alignment between SEQ and SP is enough to locate UniProt residues in PDB. However, residues in the sequence (SEQRES) can be absent and sometimes different from the coordinate section (SEQ) for various reasons (e.g. the position in space is undetermined) and this makes the direct alignment between SEQ and SP incomplete. Double-map uses two sequence alignment programs; EXONERATE [223] and BL2SEQ of NCBI blast package [64]. If EXONERATE fails to run for a short sequence around 10-15 amino acids, BL2SEQ succeeds to complete the alignment. To share the mapping data, I developed a web site which is further described in Chapter 6.

2.3.3 Calculation of Substitutions and Distance of Substitution Table

The program SUBST (http://www-cryst.bioc.cam.ac.uk/~kenji/subst), written by Dr Kenji Mizuguchi (unpublished software, see Table 2-11), was used in the calculation of substitution table. SUBST takes structural templates as inputs which can be generated by JOY [60], a program to identify the local structural environments of amino acids in the structure alignment files. The Euclidean distance between two ESSTs, X and Y, ($\mathrm{DIST}(\mathrm{X}\cdot\mathrm{Y})$) was calculated as;

$$\mathrm{DIST}(\mathrm{X}\cdot\mathrm{Y}) = \left(\sum_{i=1}^{64} \left(\sum_{j=1}^{21} \sum_{k=1}^{21} \left(X_{j\to k}^{i} - Y_{j\to k}^{i} \right)^2 \right) \right)^{1/2}$$

, where $X_{j\to k}^{i}$ and $Y_{j\to k}^{i}$ is the probability of amino acid *j* to be substituted by *k* from the ESST of X and Y under the structure environment of *i*. Note that there are 64 structure environments (4*2*8 from the secondary structures, solvent accessibility and H-bonds, respectively) and 21 amino acids (Cysteine and half-cystine using one-letter code J and C, respectively).

2.3.4 Benchmarking

CRESCENDO [100] was used to benchmark new ESSTs based on the predictions of four categories of functional residues: 1) catalytic residues of enzyme active sites, 2)

residues involved in protein-protein interactions, 3) protein-nucleic acid interactions and 4) protein-ligand interactions (see Table 2-1 for the source). The divergent score was used as it is more sensitive to the environments and it better discriminates functionally conserved residues from structurally conserved residues. The CRESCENDO scores (Z-score) were smoothed and contoured using Kin3Dcont [224]. CRESCENDO returns several clusters of predicted residues based on the size of grid points contoured using the Z-score. Residues only in the first cluster were used as the predicted residues of functional residues in the analysis. The details of the equation can be found in the original paper [100]. The P-value of the predicted residues is calculated using a one-tailed test under the standard normal distribution.

The performance ESSTs were assessed by measuring sensitivity (SENS), coverage (COV) and F-measure. These measurements were calculated based on the ratios derived from TP (true positives), FP (false positives), FN (false negatives), and TN (true negatives), which are defined as follow.

$$\mathrm{SENS} = \frac{\mathrm{TP}(ESST)}{\mathrm{TP}(ESST) + \mathrm{FN}(ESST)} \quad , \quad \mathrm{SPEC} = \frac{\mathrm{TN}(ESST)}{\mathrm{TN}(ESST) + \mathrm{FP}(ESST)} \quad ,$$

$$\mathrm{COV} = \frac{\mathrm{TP}(ESST)}{\mathrm{TP}(ESST) + \mathrm{FP}(ESST)} \text{ and } \mathrm{F-measure} = 2\frac{\mathrm{SENS} * \mathrm{COV}}{\mathrm{SENS} + \mathrm{COV}}$$

TP is the number of residues correctly predicted by CRESCENDO. If the residues predicted by CRESCENDO are the same as those annotated by the reference database, they are counted as being correct. FN is the number of real functional residues where CRESCENDO failed to predict. FP is the number of false hits that CRESCENDO predicted as functional residues but not actually annotated by the references. TP, FP, FN, and TN are exclusively determined by the ESST used in CRESCENDO.

The Spearman's rank correlation (ρ) was calculated as follows;

$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, where d_i is the difference between each rank of corresponding values and n is the number of pairs of values

Chapter 3

Three-Dimensional Structural Determinants of Amino Acid Conservation in Proteins

Neutral evolution of proteins occurs through the establishment of amino acid substitutions in the population at rates that depend on type, local tertiary environment and functional interactions of each amino acid. ESSTs (Environment Specific Substitution Tables) describe the way that amino acids are substituted as a function of their local environments, often defined by secondary structure, solvent accessibility and the existence of hydrogen-bonds from side-chains to main-chains or other side-chains. In this chapter, I quantify and rank the determinants of amino acid substitutions in the three-dimensional structures of proteins by the way they affect the rate of accepted substitutions. I show that solvent accessibility is the most important determinant, followed by the existence of hydrogen-bonds from the side-chain to main-chain functions and the nature of the element of secondary structure to which the amino acid contributes. Some of the material in this chapter has been published in Nature Review Molecular Cell Biology[13] and Biochemistry Society Transactions[14].

[13] *Worth CL, Gong S, Blundell TL (2009) Structural and functional constraints in the evolution of protein families. Nat Rev Mol Cell Biol 10: 709-720.*
[14] *Gong S, Worth CL, Bickerton GR, Lee S, Tanramluk D, et al. (2009) Structural and functional restraints in the evolution of protein families and superfamilies. Biochem Soc Trans 37: 727-733.*

3.1 Introduction

Although amino acid sequence determines protein three-dimensional structure — sometimes with a little help from chaperones — tertiary structure tends to be better conserved in evolution than sequence [1,227]. Thus, in homologous families of proteins, functions are often retained, and structures are usually very similar, even though sequences have diverged. The mantra becomes even more evident in protein superfamilies, in which overall sequence similarity can be insignificant but structural and functional similarities still provide evidence of distant common ancestry.

Comparisons of homologous proteins show that interaction sites that mediate important functions by binding regulatory proteins, nucleic acids and other ligands also provide strong evolutionary restraints on amino acid substitutions [69,205,212,228]. These cannot be understood at the level of an isolated protein; rather, different proteins and sometimes other macromolecules associate to form a multicomponent system that serves as a functional unit and provides significant restraints on evolutionary change. In insulin, for example, comparative analysis of family members have revealed that amino acid substitutions at the interfaces involved in dimer, hexamer and receptor complex formation have been under strong restraints since the evolution of bony fishes — only the rodent sub-order of histricomorpha, such as the guinea pig and coypu, have monomeric insulins [69]. Although the amino acid substitutions leading to the loss of ability of insulin to hexamerize in histrocomorpha were first thought to be selectively neutral, it is now thought that they were probably selectively advantageous, providing a stable storage form, possibly in an environment with a shortage of zinc that prevented the use of zinc insulin hexamers as found in other mammals.

For enzymes, it is clear that the local environment of catalytic residues in reaction intermediates and transition states must be considered. Strong restraints arise on recognition sequences at sites of post-translational modification, of protein-protein interactions in adaptor and template interactions and of allosteric effector binding. Recently, it has become evident that these restraints can extend to substitution of amino

acid residues in the vicinity of protein binding sites but not in immediate contact with a ligand [100].

Such comparative analyses of proteins can throw light on these observations by focusing on substitutions at topologically equivalent amino acid positions in families and superfamilies, and integrating the information into local environment-dependent amino acid substitution tables (see section 1.2.2 and Chapter 2). These show that identical amino acids are substituted in different ways, depending on the role of an amino acid in maintaining protein structure and functional interactions in the protein. What then is the nature of the restraints on amino acid substitutions that give rise to distinct patterns of protein evolution? In this chapter, I wish to investigate how local environments affect the substitution of amino acids and which environments are the major determinants of distinct patterns of amino acid substitution.

3.2 Results

An ESST (Environment Specific Substitution Table) describes the substitution of amino acids as a function of structural environments which restrict the allowable substitutions [88]. The combination of environmental descriptors for solvent accessibility, secondary structure and side-chain hydrogen-bonding gives 64 matrices for each amino acid in this model and each is associated with a distinct pattern of amino acid substitution (see section 1.2.2 and Chapter 2 for details). First of all, I measure distances amongst the 64 ESSTs and then cluster them using the UPGMA algorithm (Unweighted Pair Group Method with Arithmetic mean) [229] in order to identify which matrices give rise to similar substitution patterns. I also carry out Principal Component Analysis (PCA) [230] based on 1) the distance matrix (64*64) and 2) a matrix of substitution profiles for all 64 environments over 441 (21*21) possible substitutions (note that cysteine (J) with a free sulphydryls group is distinguished from half-cystine (C) which participates in a disulfide bridge). Figure 3-1 and Figure 3-2 show the results of clustering and PCA analysis, respectively.

3.2.1 Solvent accessibility has a major role

It has long been understood that residue conservation in the solvent inaccessible regions is much higher than those that are solvent accessible [88]. Figure 3-1 shows clustering of 64 local structural environments (ENVs) with the UPGMA algorithm [231], based on distances amongst 64 substitution tables (64*64 distance matrix) to identify the structural constraints that determine similar substitution patterns of amino acids. The distance between two substitution tables was measured by summing the differences in the probability of amino acid substitutions (see section 3.3). In Figure 3-1, the matrices for the 64 environments form three distinct clusters: two are distinguished by solvent accessibility (clusters 1 and 2 in Figure 3-1), whereas the third is characterized by the presence of a positive φ mainchain torsion angle (cluster 3 in Figure 3-1). PCA, in Figure 3-2, also divides the 64 ENVs by solvent accessibility, which corresponds to the primary principal component (PC1). From a neutral evolutionary point of view, substitutions of amino acids that change hydropathy do not in general favour protein stability, so they are selected against in evolution. As expected, for all 21 amino acids, it is observed that the degree of residue conservation in the solvent inaccessible regions is much higher than that of solvent accessible regions (see Figure 3-3).

Even within the cluster of environments having positive φ mainchain torsion angles (see section 3.2.3), solvent accessibility divides the environments into two: accessible and inaccessible. Solvent inaccessibility thus puts constraints on the acceptance of selectively neutral amino acid substitutions during evolution, although it should be noted that thermodynamically stable proteins are much more tolerant to mutations [232,233]. Based on the clustering pattern of 64 ENVs and other evidence mentioned earlier, it is evident that solvent accessibility is the primary structural constraint on amino acid substitutions and mutation rates during protein evolution.

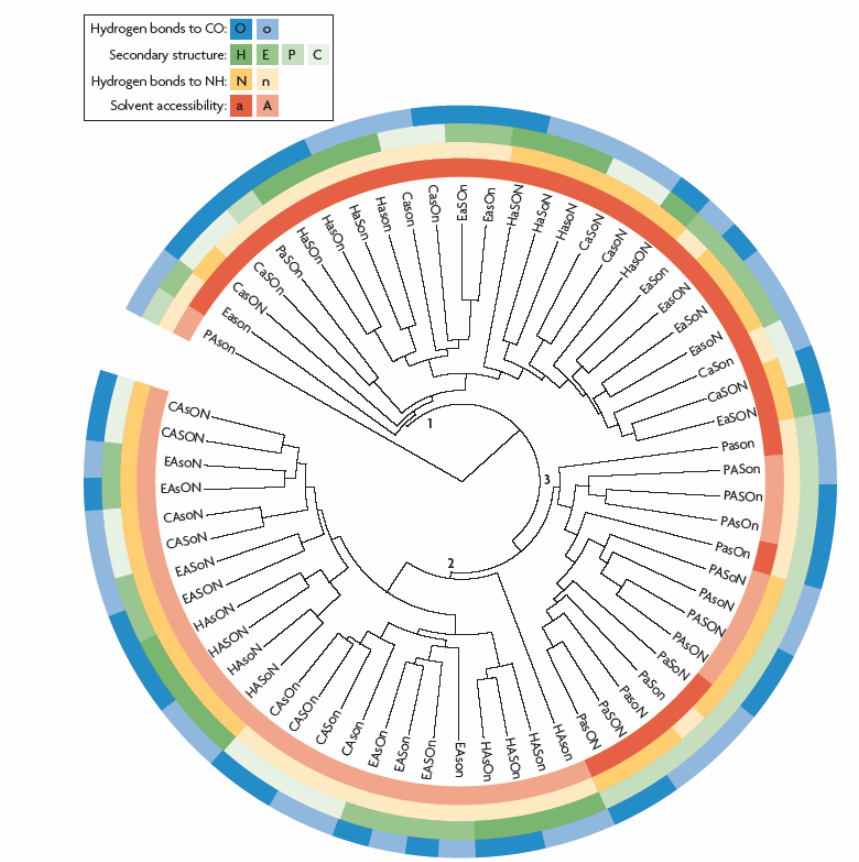


Figure 3-1 Results of hierarchical clustering of 64 environments

Trees are constructed on the basis of the 64*64 distance matrix. Environments are shown using five-letter code representation: the first letter defines the secondary structure (α-helix (H), β-strand (E), positive φ main-chain torsion angle (P) and coil (C)), the second defines solvent accessibility (accessible (A) and inaccessible (a)) and the remaining three letters define the existence (upper case) or absence (lower case) of hydrogen bonds from a side chain to another side chain (S and s, third letter), to a main-chain carbonyl group (O and o, fourth letter) and to a main-chain amide group (N and n, fifth letter) (see also section 1.2.2 for details). Three major clusters are numbered as 1, 2 and 3 on the nodes from which they branch. Around the tree there are four concentric rings, each of which represents a particular structural parameter: the first ring represents solvent accessibility, the second ring represents the existence or absence of hydrogen bonds from a side chain to a main-chain amide group, the third ring represents the type of secondary structure and the fourth ring represents the existence or absence of hydrogen bonds from a side chain to a main-chain carbonyl group. The 4 concentric rings highlight the hierarchical clustering of the 64 environments by showing which amino acid substitution matrices are similar and which local environments are the major determinants of the substitution patterns. The trees were drawn using iTOL[15] [234]. The figure is taken from the reference [225] by Worth et al., which I co-authored with.

[15] http://itol.embl.de/

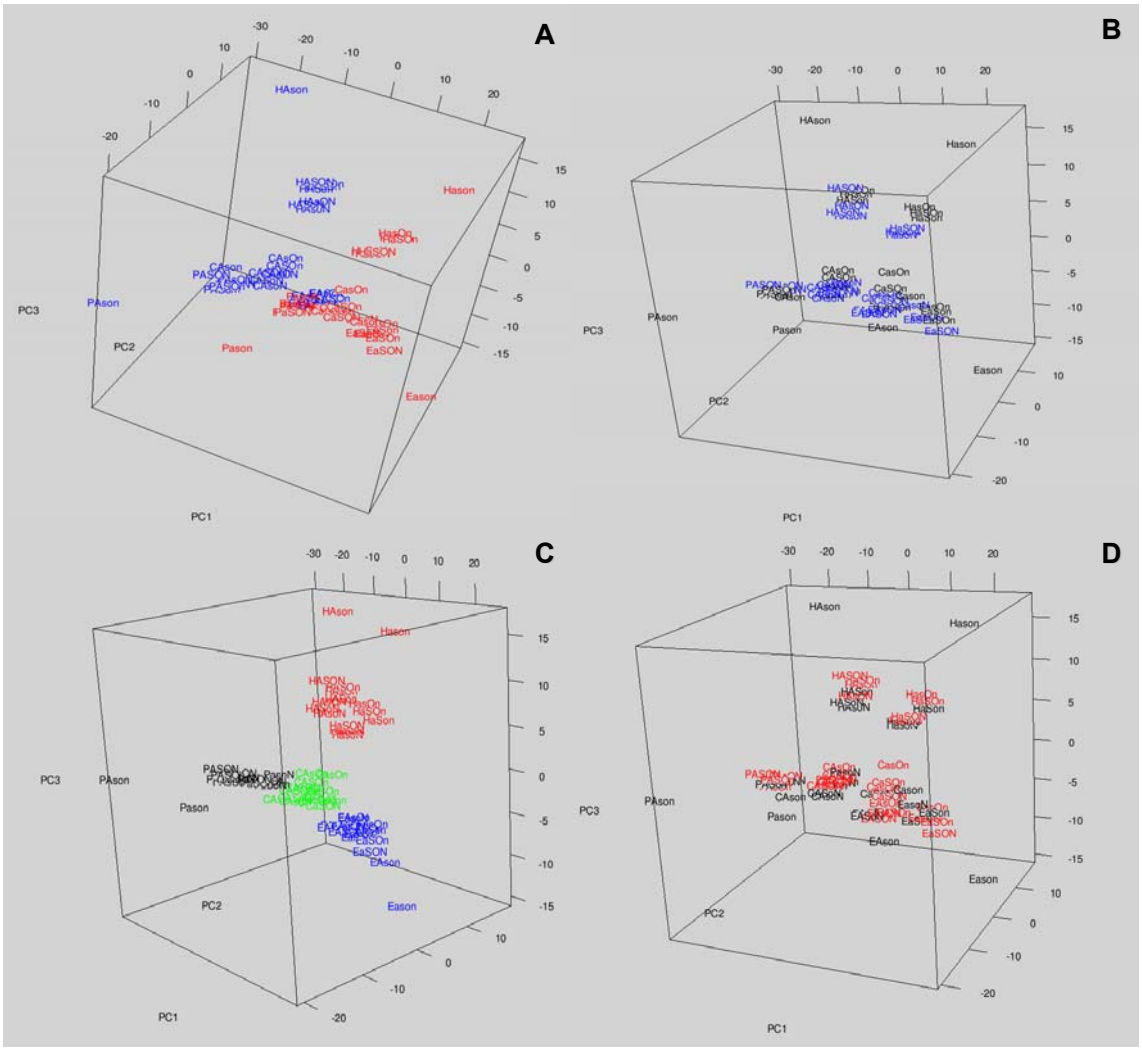


Figure 3-2 64 Environments Projected into the Axis of Three Major Principal Components

A matrix of substitution profiles (64*21*21) was used for the PCA (Principal Component Analysis). Each of the ENVs are coloured by A) the solvent accessibility (red: inaccessible, blue: accessible), B) the presence (blue) or absence (black) of hydrogen-bond from side-chain to main-chain amides, C) the element of secondary structures (red: α-helix, blue: β-strand, black: positive φ main-chain torsion angle, green: coil), and D) the existence (red) or absence (black) of hydrogen-bond from side-chain to main-chain carbonyls. The first, second and third principal component are responsible for 31%, 13%, and 8% of the total variance. See Appendix I for the coordinates of 64 environments projected on the PC1, 2 and 3. Figures were drawn by RGL package of R software[16].

[16] http://www.r-project.org/

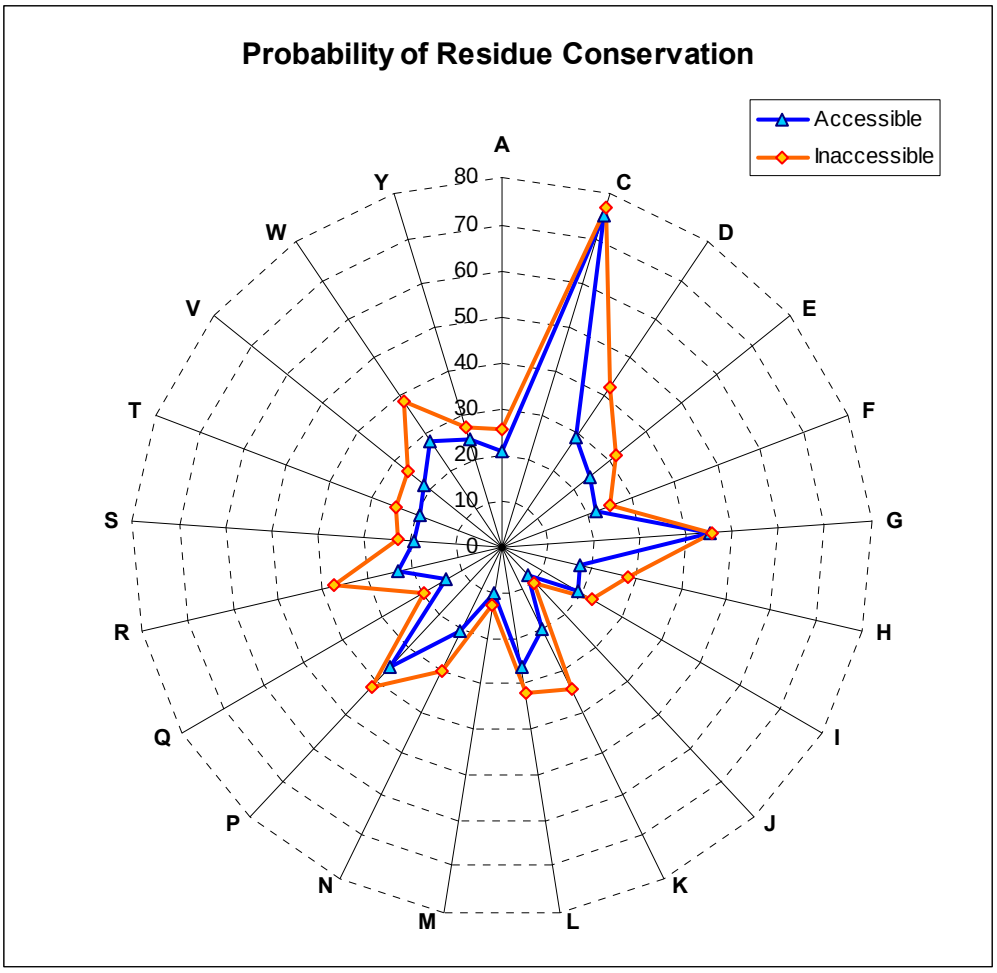


Figure 3-3 Probabilities of Residue Conservation by Solvent Accessibility

The probabilities of residue conservation in the solvent accessible area (blue) are compared with those in the solvent inaccessible region for 21 amino acids. From the 64 substitution tables, the probabilities on the diagonal axis were averaged for each of the two groups; solvent accessible and inaccessible. Note that cysteine and half-cystine are distinguished using one-letter codes J and C, respectively.

3.2.2 Influence of hydrogen bonds on amino acid substitutions

Each of the three major clusters discussed above is further divided by the presence or absence of hydrogen bonds from sidechains to mainchain NH (shown as the second concentric ring in Figure 3-1). Hence, in either solvent accessible or inaccessible environments, the establishment of hydrogen bonds from sidechains to mainchain NH restricts the substitution of amino acids, regardless of the local secondary structure. Interestingly, secondary structure (third concentric ring) defined as helix, extended

strand, positive φ torsion angle, or coil leads to the formation of clusters within each of those defined by mainchain NH.

Amino acids with hydrogen bonds to mainchain CO groups (outermost concentric ring) are grouped together within the secondary structure cluster, but the clustering pattern is weaker than that of mainchain NH groups. This suggests that the different types of hydrogen bonds have hierarchical effects on the substitution patterns of amino acids; hydrogen bonds between sidechain and mainchain NH groups are most influential, followed by mainchain and mainchain, and then sidechain and mainchain CO groups. I further investigated this pattern by averaging the effect of the solvent accessibility and then both solvent accessibility and the type of secondary structures. When the effects of solvent accessibility and then both solvent accessibility and the type of secondary structure are averaged, the clustering retains the same order of hierarchy (see Figure 3-4A and Figure 3-4B). Especially, it is evident that there is a hierarchy in the influence of the eight types of hydrogen bonds from sidechains on amino acid substitutions within homologous proteins; 8 ENVs are divided by the existence of a hydrogen-bond from a side chain to main-chain amide (N/n) followed by main-chain carbonyl (O/o) (see Figure 3-4C and Figure 3-4D).

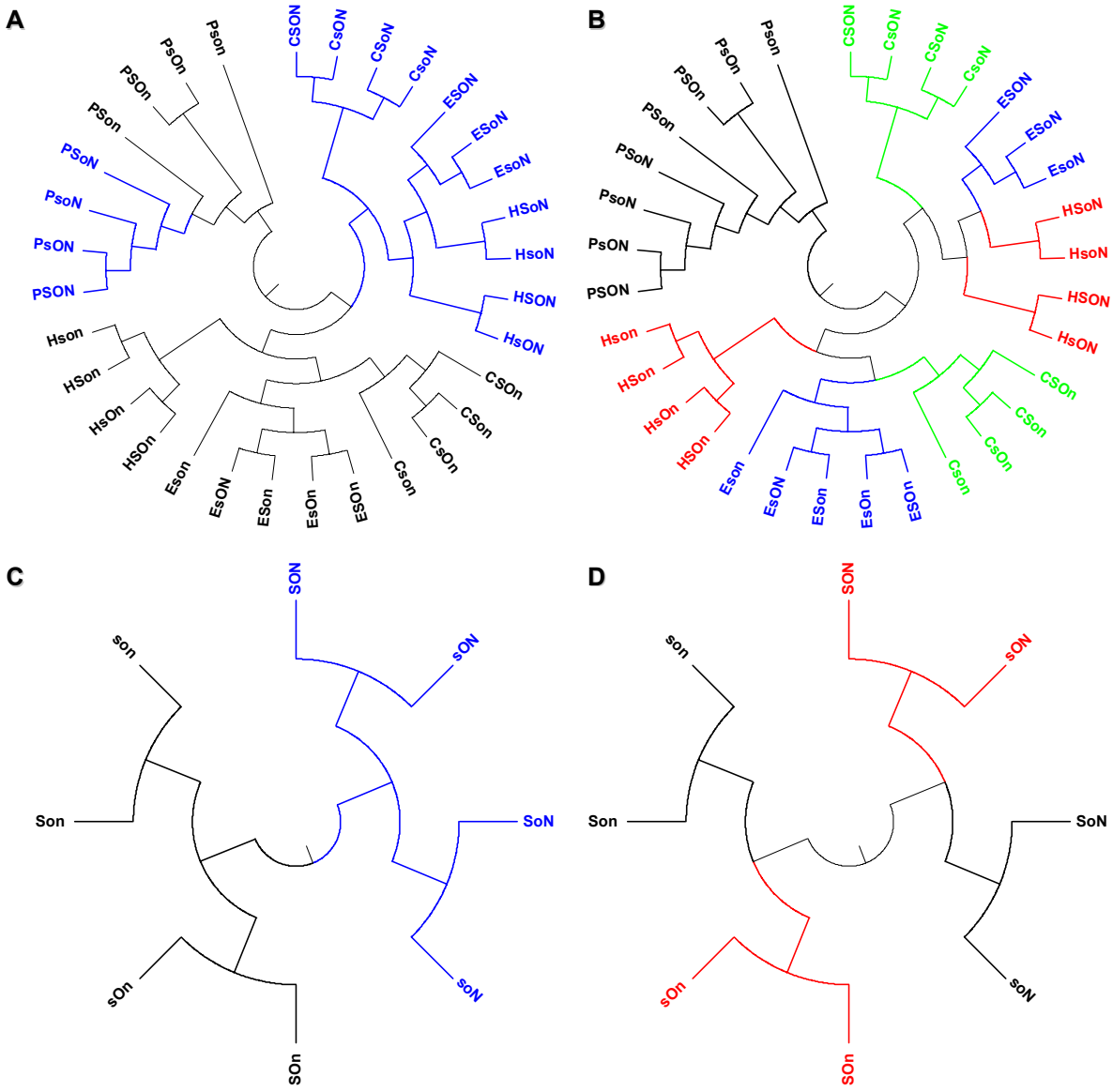


Figure 3-4 Results of hierarchical clustering of 32 and 8 environments.

A, B | Hierarchical clustering for 32 Environments whereby 64 tables are aggregated into 32 tables by averaging the effect of solvent accessibility (A/a). Hence, the tree was constructed based on the 32*32 distance matrix. **C, D** | Hierarchical clustering for 8 types of hydrogen bonds from sidechains where 62 tables are aggregated into 8 tables by averaging the effect of solvent accessibility (A/a) and the elements of secondary structure (H/E/P/C). Hence, the distance matrix reflects only the effect of hydrogen bonds from sidechains. **A** | Coloured by the existence (blue) or absence (black) of hydrogen bonds from sidechain to mainchain NH. **B** | Coloured by the element of secondary structures (red: α-helix, blue: β-strand, black: positive φ mainchain torsion angle, green: coil). **C** | Coloured by the existence (blue) or absence (black) of hydrogen bond from sidechain to mainchain NH. **D** | Coloured by the existence (red) or absence (black) of hydrogen bond from sidechain to mainchain CO.

3.2.3 Positive φ torsion angles constrain protein evolution

In Figure 3-1, matrices for the 64 environments with positive φ torsion angles constitute a distinct cluster, whereas other elements of secondary structure are divided by solvent accessibility. A positive φ torsion angle can be accommodated by a Gly, which has no sidechain, but for most other L-amino acids it leads to disallowed interactions between sidechain and mainchain atoms. However, for L-amino acids such as Asp or Asn, interactions between the sidechain CO group with the CO of the mainchain peptide bond can give rise to relative stabilization of a conformation with a positive φ angle [235]. Indeed, Gly represents 63% of total amino acids that have a positive φ torsion angle, followed by Asn (8%) and Asp (5%) (see Table 3-1). In addition, within a positive φ class, solvent accessible amino acids occur five times more frequently than inaccessible residues, whereas the average ratio of accessible to inaccessible residues falls within 2.2 for all classes of secondary structure. Hence, the predominance of Gly and polar residues in the set of amino acids with a positive φ torsion angle makes a distinct substitution pattern and eventually a distinct cluster.

Table 3-1 Propensity of Amino Acids within a Positive φ Torsion Angle

Amino Acids	Solvent Accessible			Solvent Inaccessible			Total		
	NO.	ratio	log odd ratio[1]	NO.	ratio	log odd ratio[2]	NO.	ratio	log odd ratio[3]
G	33674	0.604	0.906	8537	0.768	0.993	42211	0.631	0.920
N	5093	0.091	0.243	296	0.027	0.140	5389	0.081	0.284
D	3385	0.061	-0.093	181	0.016	-0.097	3566	0.053	-0.037
K	2245	0.040	-0.314	34	0.003	-0.312	2279	0.034	-0.238
E	1580	0.028	-0.507	63	0.006	-0.421	1643	0.025	-0.438
R	1566	0.028	-0.380	59	0.005	-0.347	1625	0.024	-0.314
S	1487	0.027	-0.377	244	0.022	-0.334	1731	0.026	-0.354
Q	1227	0.022	-0.327	65	0.006	-0.338	1292	0.019	-0.272
A	1103	0.020	-0.520	407	0.037	-0.528	1510	0.023	-0.569
H	1009	0.018	-0.176	83	0.007	-0.288	1092	0.016	-0.152
Y	699	0.013	-0.443	125	0.011	-0.497	824	0.012	-0.453
L	680	0.012	-0.697	277	0.025	-0.795	957	0.014	-0.800
F	554	0.010	-0.461	241	0.022	-0.478	795	0.012	-0.528
T	366	0.007	-0.948	74	0.007	-0.864	440	0.007	-0.924
M	321	0.006	-0.473	84	0.008	-0.644	405	0.006	-0.564
V	252	0.005	-1.014	78	0.007	-1.270	330	0.005	-1.170
C	213	0.004	-0.313	161	0.014	-0.286	374	0.006	-0.403
W	163	0.003	-0.598	56	0.005	-0.528	219	0.003	-0.608
I	136	0.002	-1.157	42	0.004	-1.454	178	0.003	-1.336
P	33	0.001	-1.955	3	0.000	-2.038	36	0.001	-1.931
Total	55786	1		11110	1		66896	1	

[1]: log odd ratio over total accessible amino acids

[2]: log odd ratio over total inaccessible amino acids

[3]: log odd ratio over total amino acids

3.2.4 On the frequency of occurrence of local environments

Analysis of representative structures [193] of protein families shows that ~80% of all the amino acids belong to one of 11 (out of 64) local environments (see Table 3-2). However none of these 11 local environments includes any hydrogen bonds from sidechains to mainchain NH, as expected from the observation that 68.6% of amino acids are non-polar and therefore cannot take part in any hydrogen bonds from sidechains. Only 8.5% of amino acids have a sidechain with a proton acceptor group and can therefore make hydrogen bonds from sidechains to mainchain NH groups, the second most important local environmental determinant of substitutions after solvent accessibility (See Table 3-3). The 8.5% of amino acids include 10 amino acids (Asp, Ser, Asn, Thr, Glu, Gln, Tyr, Met, Cys, His), and among them only Asp, Asn and Ser are over-represented compared to their background propensities in the protein dataset. This shows that the distribution of amino acids taking part in hydrogen bonds from sidechain to mainchain follows the power law distribution – only a small proportion of amino acids have an important role in the substitution pattern.

Table 3-2 The occurrence of amino acids by 64 local structural environments

The dataset was downloaded from:

http://samul.org/ESST/esst/Result.SCOP/ALL/MaskB.tgz

Rank	ENV	NO. of amino acids	Occurrence (%)	Cumulative percentage
1	CAson	187,506	18.05	18.05
2	HAson	162,809	15.67	33.72
3	Hason	92,903	8.94	42.66
4	Eason	84,693	8.15	50.81
5	EAson	71,783	6.91	57.72
6	Cason	55,480	5.34	63.06
7	PAson	47,996	4.62	67.68
8	HASon	43,702	4.21	71.89
9	CASon	35,158	3.38	75.27
10	HAsOn	26,983	2.60	77.87
11	EASon	22,009	2.12	79.99
12	CAsOn	19,333	1.86	81.85
13	CAsoN	16,454	1.58	83.43
14	CASoN	14,863	1.43	84.86
15	HASOn	12,381	1.19	86.05
16	Pason	9,837	0.95	87.00
17	EaSon	8,636	0.83	87.83
18	CASOn	8,479	0.82	88.65
19	HasOn	8,311	0.80	89.45
20	HaSon	7,243	0.70	90.14
21	EAsOn	6,376	0.61	90.76
22	HaSOn	5,949	0.57	91.33
23	CAsON	5,575	0.54	91.87
24	CaSon	5,381	0.52	92.38
25	CasOn	4,832	0.47	92.85
26	EasOn	4,452	0.43	93.28
27	HASoN	4,166	0.40	93.68
28	CaSoN	4,072	0.39	94.07
29	CasON	3,803	0.37	94.44
30	EASOn	3,775	0.36	94.80
31	CaSOn	3,678	0.35	95.15
32	HAsoN	3,589	0.35	95.50
33	EaSOn	3,557	0.34	95.84
34	PASon	3,430	0.33	96.17
35	CASON	3,203	0.31	96.48
36	CaSON	2,824	0.27	96.75
37	HAsON	2,651	0.26	97.01
38	HasON	2,622	0.25	97.26
39	EAsoN	2,466	0.24	97.50

40	EaSoN	2,349	0.23	97.72
41	EASoN	2,276	0.22	97.94
42	HaSoN	2,245	0.22	98.16
43	PAsOn	1,968	0.19	98.35
44	EasON	1,940	0.19	98.53
45	CasoN	1,874	0.18	98.71
46	HaSON	1,868	0.18	98.89
47	EaSON	1,668	0.16	99.06
48	HASON	1,658	0.16	99.21
49	EasoN	1,417	0.14	99.35
50	HasoN	1,320	0.13	99.48
51	EAsON	1,098	0.11	99.58
52	PASOn	770	0.07	99.66
53	EASON	659	0.06	99.72
54	PAsoN	602	0.06	99.78
55	PASoN	499	0.05	99.83
56	PAsON	352	0.03	99.86
57	PaSon	320	0.03	99.89
58	PaSOn	237	0.02	99.91
59	PasOn	180	0.02	99.93
60	PASON	169	0.02	99.95
61	PaSoN	166	0.02	99.96
62	PasON	148	0.01	99.98
63	PaSON	128	0.01	99.99
64	PasoN	94	0.01	100.00
Total		1,038,965	100	

Table 3-3 The occurrence of eight types of hydrogen bonds from sidechains

(F: False, T: True, see Figure 1-1 for the ENV code).

ENV	Hydrogen-bonds from sidechains			NO. of amino acids	Occurrence (%)
	to other sidechain	to mainchain CO	to mainchain NH		
son	F	F	F	713,007	68.63
Son	T	F	F	125,879	12.12
sOn	F	T	F	72,435	6.97
SOn	T	T	F	38,826	3.74
SoN	T	F	T	30,636	2.95
soN	F	F	T	27,816	2.68
sON	F	T	T	18,189	1.75
SON	T	T	T	12,177	1.17
Total				1,038,965	100

The dataset was downloaded from:

http://samul.org/ESST/esst/Result.SCOP/ALL/MaskB.tgz

3.2.5 Discussion

In this Chapter, I have shown that the degree of amino acid conservation is most affected by the solvent accessibility followed by the presence of hydrogen bonds from sidechains to mainchains and between mainchains. However, there are other types of non-conventional interactions, which are highly conserved and have important roles in protein structures and binding regions [74,97,236]. A further consideration is the extent to which the local environment is conserved in homologous families and therefore can provide constraints on amino acid substitutions. Analyses of families and superfamilies of proteins show that the most crucial packing arrangements of individual sidechains begin to differ when two proteins have less than 30% sequence identity due to relative movements of equivalent secondary structural elements, but some crucial hydrogen-bonding interactions are retained at much greater levels of sequence divergence.

It has long been understood that hydrogen bonds play a very important role in the stability of a protein structure, and provide restraints on the substitutions of amino acids during evolution by neutral drift. Recently, Worth et al. addressed the importance of hydrogen-bond potentials from side-chains in the stability of protein structures [97].

They showed that the formation of hydrogen bonds to main-chain amide atoms influences conservation of amino acids, with those satisfied buried polar residues that form two hydrogen bonds to main-chain amides being significantly more conserved than those that form only one or none. Their evidence and my findings provide new insights into the roles of networks of hydrogen bonds within the three-dimensional structures of proteins.

3.3 Methods

3.3.1 Environment Specific Substitution Tables

The Environment Specific Substitution Table [88,89] was derived from the alignments of homologous proteins whose three-dimensional structures have been determined. The PDB [214] was used as a source for the three-dimensional structures of proteins and SCOP [37] for the definition of protein families and domains. For each SCOP family, domains were clustered with sequence identity of 80% or more, after pre-processing the structure data. SUBST[17] (Dr Kenji Mizuguchi, unpublished software) was used to calculate the ESST. The detailed procedures for making the ESST are explained in our recent paper and the web site[18] [193].

3.3.2 Calculation of Structural Environments of Amino Acids

JOY[19] was used to identify the local structural environments of amino acids [60]. JOY consists of three supporting programs − SSTRUC, PSA, and HBOND − to annotate 1) the elements of secondary structure, 2) solvent accessibility, 3) hydrogen-bonds from side chains, respectively. SSTRUC calculates torsion angles within a main-chain to assign secondary structure. For the threshold of solvent accessibility, a cut-off of 7.0% relative total side-chain accessibility has been applied. HBOND identifies all possible hydrogen bonds based on a distance criterion; 3.5Å between donor and acceptor except for interactions involving sulphur atoms where 4.0Å is used.

[17] http://mordred.bioc.cam.ac.uk/~kenji/subst
[18] http://samul.org/ESST
[19] http://tardis.nibio.go.jp/joy/

3.3.3 Hierarchical Clustering and Principal Component Analysis (PCA)

The hierarchical clustering analysis is based on the Euclidean distances amongst 64 environments. The Euclidean distance ($\mathrm{DIST}(\mathrm{X}\cdot\mathrm{Y})$), between two environments, X and Y, defined as;

$$\mathrm{DIST}(\mathrm{X}\cdot\mathrm{Y}) = \left(\sum_{j=1}^{21} \sum_{k=1}^{21} \left(X_{j\to k}^{i} - Y_{j\to k}^{i} \right)^2 \right)^{1/2}$$ where $X_{j\to k}^{i}$ and $Y_{j\to k}^{i}$ are the probabilities of amino acid *j* to be substituted by *k* under the environment of X and Y, respectively. Hence, the distance matrix is an upper (or lower) triangular matrix of 64*64 dimensions with 0 in the diagonal entries. PCA was performed based on either the distance matrix or a matrix of substitution profiles for all 64 environments over 441 (21*21) possible substitutions. For an ESST, I used ALL-B type (see Table 2-2), which turns out to be the best in our benchmarking process in Chapter 2 [193]. For the hierarchical clustering, I used the PHYLIP package with UPGMA method as a clustering algorithm [237]. For the PCA analysis, "prcomp" function, in stat package of standard R software[20], has been used. The source code for R is available from http://samul.org/ESST/R.tar.gz.

[20] http://www.r-project.org/

Chapter 4

Structural and Functional Restraints on the Occurrence of Single Amino Acid Variations in Human Proteins

Human genetic variation is the incarnation of diverse evolutionary history, which reflects both selectively advantageous and selectively neutral change. In this chapter, I catalogue structural and functional features of proteins that restrain genetic variation leading to single amino acid substitutions. The variation dataset used in this study is divided into three categories: i) Mendelian disease-related variants, ii) neutral polymorphisms and iii) cancer somatic mutations. I characterize structural environments of the amino acid variants by the following properties: i) side-chain solvent accessibility, ii) main-chain secondary structure, and iii) hydrogen bonds from a side chain to a main chain or other side chains. To address functional restraints, amino acid substitutions in proteins are examined to see whether they are located at functionally important sites involved in protein-protein interactions, protein-ligand interactions or catalytic activity of enzymes. I also measure the likelihood of amino acid substitutions and the degree of residue conservation where variants occur. I show that various types of variants are under different degrees of structural and functional restraints, which affect their occurrence in human proteome. An initial report of this work has been published as the same title of this chapter in Plos One[21].

[21] *Gong S, Blundell TL (2010) Structural and functional restraints on the occurrence of single amino Acid variations in human proteins. PLoS One 5: e9186.*

4.1 Introduction

The evolution of orthologous proteins occurs through the establishment of amino acid substitutions in the population at rates that depend on restraints arising from the need to maintain proper three-dimensional structure and to retain functional interactions of each amino acid within or between molecules [9,10,225,226]. For example, amino acids in the cores of proteins are relatively conserved compared to those in the solvent accessible regions [16,97] and catalytic amino acids responsible for enzymatic reaction are also well conserved throughout evolution. Hence, mutations tend to be accepted in amino acid residues where evolutionary pressure is relatively relaxed and where they can remain in the population without selective disadvantage (or advantage). Recently, high-throughput DNA sequencing technology has begun to have a major impact on this field and is shedding light on genomic sequence variations between human individuals [62,239,240,241]. Single nucleotide polymorphisms (SNPs) in protein coding regions are of special interest as they may be non-synonymous (nsSNPs), resulting in changes in the types of amino acid in the protein products. Indeed, recent analysis of human nsSNPs shows that the majority are commonly found and appear to be functionally neutral [242]. Thus, it is of interest to examine whether the occurrence of coding variations in the human population is equally affected by the factors that restrain the substitutions of amino acids observed in divergent evolution of proteins.

One of the consensus agreements from molecular analyses of coding variants is that, although most of them are selectively neutral, their occurrence is restrained by various factors such as solvent accessibility, type of secondary structure, and presence of side-chain hydrogen bonding. Compared with benign and neutral variants, disease-related variants are more likely to be located in solvent inaccessible regions and tend to change the physicochemical properties from those of the wild type amino acids [110,113]. In addition, disease-related variants are more likely to be located at conserved residues, which are believed to be functionally important [176,243]. However, previous analyses have been based on relatively small sub-sets of sequence variants, and have not fully taken advantage of the rapidly growing information on protein structure and function. Hence, in this era of information deluge from high-speed genome sequencing, high-

resolution protein structure determination, and enriched annotation on protein functions, it is desirable to have large-scale cataloguing of coding variants in the light of structure and function of proteins. This will help us understand not only the nature of deleterious mutations, but also the evolutionary nature of the occurrence of single amino acid variations.

In this chapter, I address structural and functional restraints that shape the occurrence of single amino acid variations, which I categorise them into three categories: i) Mendelian disease-related variants, ii) neutral polymorphisms and iii) cancer somatic mutations. Structural environments of amino acid variants are further characterised by mapping sequence positions onto their corresponding three-dimensional structures if available. I confirm earlier analyses [110,113] that report nsSNPs occur less frequently at the solvent inaccessible region of proteins, whereas disease-related mutations occur much more frequently than the average. I also find that cancer somatic mutations and disease-related variants occur more frequently at amino acids making hydrogen bonds from side chains than neutral polymorphisms. Substitution scores and the degree of sequence conservation at the variant positions are measured and differences amongst the variant datasets are compared.

4.2 Results and Discussion

4.2.1 Compilation of Amino Acid Variant Dataset

The variant dataset was compiled from the following sources: 1) Swiss-Prot human variants [244], 2) Ensembl human variation database [245], and 3) COSMIC (Catalogue Of Somatic Mutation In Cancer) database [140] (see Materials and Methods for details). The Swiss-Prot variants are further classified by Mendelian disease-related variants (SVD) and polymorphic variants (SVP) according to the original annotations from the source. For Ensembl human variations (SAP), only verified SNPs (see section 4.3.1) were used in order to ensure an accurate and reliable polymorphic dataset. The COSMIC dataset (CSM) differs from the others in that it contains somatic mutations observed in various cancer types. The sequence positions of variants from the source

data were transferred to UniProt protein sequence level [216] and further mapped onto their corresponding locations in terms of three-dimensional structures if available in PDB [214]. Table 4-1 shows the number of variants from the source data, variants mapped onto UniProt protein level, and PDB level. SVD does not share variants with SVP, but does share 232 and 104 variants with CSM and SAP respectively, which are less than 1.4% of SVD (see Figure 4-1 for details). CSM shares less than 0.9% either with SAP (15/4476) or SVP (31/4476). However, SVP and SAP share ~51% (16863/32748) and ~57% (16863/29541) with each other, which is not surprising because both represent polymorphic variants. Considering the low percentage of overlaps amongst Mendelian disease (SVD), cancer somatic (CSM) and neutral polymorphic variants (SAP and SVP), those overlaps are not removed in the analysis which I now describe.

Table 4-1 Four types of sequence variants and their numbers

Sources	Types	Abbreviations	NO. of distinct variants		
			from the source	mapped to UniProt	mapped to PDB
UniProt	Disease	SVD	16,776	16,776	4,942
	Polymorphism	SVP	32,748	32,748	2,895
Ensembl	verified SNPs	SAP	29,541	28,702	2,024
COSMIC	cancer mutations	CSM	5,260	4,476	2,016

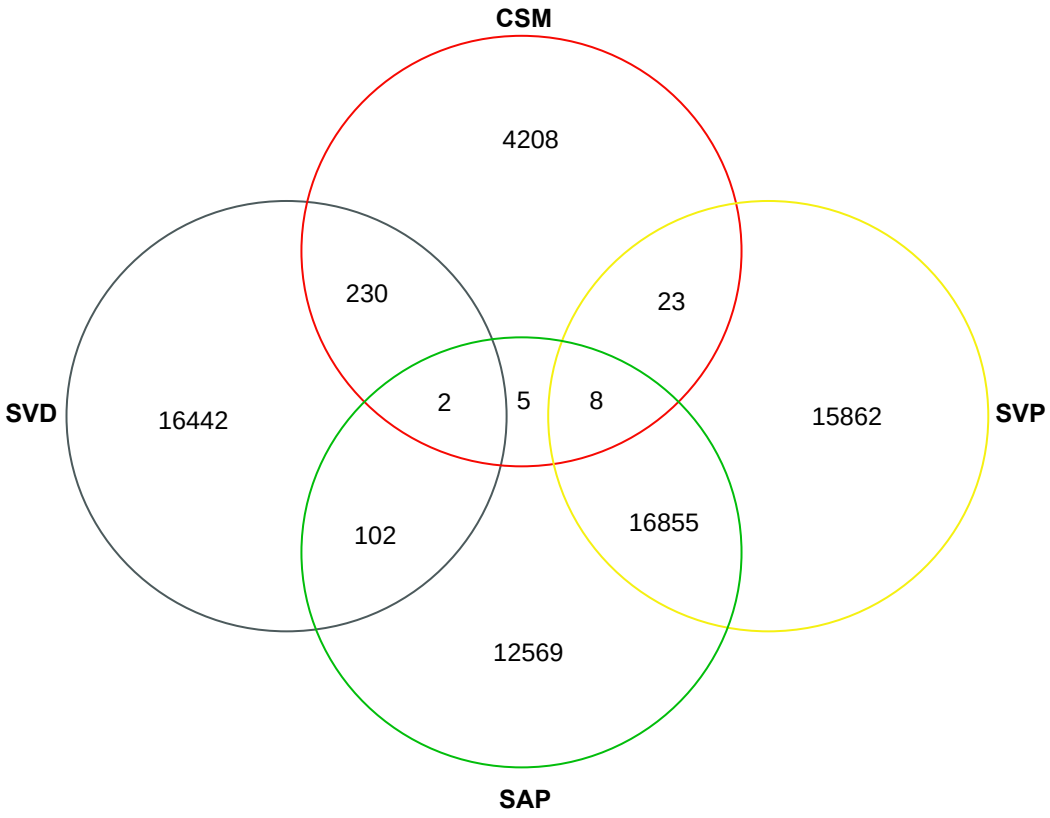


Figure 4-1 A Venn diagram showing the number of overlaps amongst variant datasets

Four variant datasets (SVD, SVP, SAP and CSM) are from Table 1. (SVD: Mendelian disease-related variants, CSM: Cancer somatic mutations, SVP and SAP: Polymorphic variants, see 'Compilation of amino acid variant dataset' of Results and Discussion section)

4.2.2 Local Structural Environments of Sequence Variants

In order to characterize the local structural environments of amino acid variants where three-dimensional structures of proteins are known, the local structural environments of amino acids were first defined as suggested by Overington and colleagues [88,89]: 1) main-chain conformation and secondary structure, 2) solvent accessibility and 3) hydrogen bonding between side chains and main chains. In this framework, there are 64 distinct environments for a residue from the combination of structural features: four from secondary structures (α-helix, β-strand, coil and residue with positive φ main-chain torsion angle), two from solvent accessibility (accessible and inaccessible), and eight (2^3) from hydrogen bonds to main-chain carbonyl (CO) or amide (NH) or to another side chain. Four types of variants were mapped onto PDB structures and characterized by their local structural environments (see Supplementary Dataset S1, S3 and S5 in [238]). In Table 4-2, I quantified the proportions of variants that belong to each environmental category and compared them among four variant classes. To give background proportions of amino acids for each environmental feature, amino acids from representative domains (see Materials and Methods) of SCOP families [37] are counted and their proportions are given in Table 4-2. I investigated whether the ratio of variants for each environment category could result from the structural restraints that shape the occurrence of variants in proteins.

Table 4-2 Occurrence (%) of variants by structural environments

Structural environment			Types of variants				Background
Categories		**types**	**SVD[7]**	**SVP[8]**	**CSM[9]**	**SAP[10]**	**SCOP[11]**
solvent accessibility		a[1]	42.25	18.45	26.45	19.48	31.21
hydrogen bonds from side chains	to main-chain amides	T[2]	10.69	5.79	8.44	5.69	8.55
	to main-chain carbonyls	T	19.50	13.01	13.27	13.36	13.63
	to other side chains	T	25.58	19.31	21.93	17.04	19.97
secondary structure		H[3]	27.98	32.98	22.14	31.58	36.61
		E[4]	23.25	20.23	20.26	20.13	21.09
		P[5]	9.71	6.40	10.26	6.60	6.45
		C[6]	39.06	40.39	47.34	41.69	35.85

[1]: inaccessible [2]: True (hydrogen bonded)

[3]: α-helix [4]: β-strand [5]: positive φ main-chain torsion angle [6]: coil

[7]: see Supplementary DatasetS1 of [238], [8]: see Supplementary DatasetS3 of [238]

[9]: see Supplementary DatasetS7 of [238], [10]: see Supplementary DatasetS5 of [238]

[11]: see 'Representative SCOP domains' of Materials and Methods

4.2.2.1 By solvent accessibility

I observed that Mendelian disease-related variants (SVD) occur twice as often as polymorphic variants (SVP and SAP) at solvent inaccessible positions. For cancer mutations (CSM), the proportion of variants in solvent inaccessible regions is more than that of SVP but less than SVD. If a sequence variant occurs randomly in proteins, the probability of being located in a solvent inaccessible region would be close to 31.21%, which is the proportion of solvent inaccessible amino acids from the representative SCOP domains. As shown in Table 4-2, SVD occur 35% (42.25/31.21 - 1) more than expected ($P < 10^{-6}$) [22], whereas polymorphic variants (SVP and SAP) occur 40% (1 - 18.45/31.21) less often than expected ($P < 10^{-6}$) [23]. This observation is inline with one of the early analyses of the frequency of disease mutations, which showed that 35% of 551 disease-causing mutations affect buried sites, whereas only 9% of 225 substitutions between species do [110]. This also agrees with the finding that for most monogenic diseases a single DNA variant, resulting in an amino acid substitution, is responsible for the disease by affecting protein stability rather than damaging the protein's specific function directly [112]. Presumably, the differences in the frequency of occurrence by mutation types may arise from evolutionary pressure, which restricts the occurrence of variants in the core regions of proteins in order to minimize the effects on the stabilities of proteins. The mechanism of protein stability studied by using thermodynamics measurements (on mutants created by site-directed mutagenesis) revealed that the degree of free-energy changes (i.e. $\Delta\Delta G$) is highly correlated with the location where the mutation occurs within three-dimensional proteins – $\Delta\Delta G$ is negatively correlated with solvent accessibility [131]. Hence, these indicate that disease-causing mutations often affect intrinsic structural features of proteins.

[22] P-value is obtained by an approximation via the normal distribution because the total number of observations is quite large (n=4942). The exact calculation of P-values is based on the binomial distribution; the probability of observing 2088 solvent-inaccessible mutants (2854 within accessible) out of 4942 disease-related mutants mapped to PDB, under the null-hypothesis which states the occurrence of mutants follows the proportion of inaccessible residues (31.21%).

[23] P-value is obtained by the same method stated above, but observing 534 solvent-inaccessible mutants (2361 within accessible) out of 2895 polymorphic variants mapped to PDB (see Table 4-1).

4.2.2.2 By hydrogen-bond capacity

For three categories of hydrogen-bond types, SVD occur more frequently at amino acids making hydrogen bonds ('T' in Table 4-2) than do the other variants. CSM also occur more frequently than polymorphic variants, but the difference is smaller than that of SVD. This observation, together with the ratios of occurrence in the interior/surface regions of proteins, clearly shows that amino acid variants are under strong restraints, resulting in the observation that they occur less frequently in regions maintaining the architectures of protein structures.

4.2.2.3 By element of secondary structure

As shown in Table 4-2, compared with the ratios of residues from representative SCOP domains and other polymorphic variants (SVP and SAP), SVD and CSM occur less in residues in α-helices (H), but more often at residues with positive φ main-chain torsion angles (P). Interestingly, almost half of CSM (47.34%) occur in coil regions, distinguishing them from other variant datasets (~41.69%). However, this observation is probably skewed towards well characterised cancer proteins such as p53 and various types of kinase proteins which are dominantly found in the COSMIC dataset. Indeed, only 10 UniProt proteins, out of 188 known three-dimensional structures, are responsible for 80% of cancer mutations mapped to the PDB. p53 (UniProt accession: P04637) alone takes up 27% of 2016 cancer mutations shown in Table 4-1. Therefore, it is not reasonable to base any statistical interpretation on the preference of secondary structure for cancer mutations on this observation. However for disease-causing mutations, my results agree with those of Ferrer-Costa and colleagues [113] who showed disease-related SNPs occur less in α-helices but more frequently in β-strands than neutral nsSNPs, although differences in the percentages may arise from the methods used for defining secondary structure.

4.2.3 Amino Acid Substitution Scores

Amino acid substitution models such as PAM [80] and BLOSUM [82] describe the degree of substitutions as log-odd ratio values where the positive scores suggest commonly occurring and preferred substitutions, whereas the negative scores imply very rare substitutions which are disfavoured in nature. Those substitution tables were

widely used to assess and predict the effects of nsSNPs [101,113]. An ESST (Environment Specific Substitution Table, http://samul.org/ESST) also describes the degree of substitution of amino acids, but differs from PAM or BLOSUM by taking into account structural environments which restrict the possible and allowable substitutions [88,89]. Hence, ESSTs provide more accurate and discriminating measures of substitution probabilities in a particular environment in a three-dimensional protein structure. Figure 4-2A and Figure 4-2B show box plots of substitution scores from four types of variants in the dataset using BLOSUM62 and ESST, respectively. From both models, the median substitution scores for SVD and CSM are lower than those of SVP and SAP. Substitution scores are further investigated by the local structural environments of the variants where they occur in three-dimensional structures of proteins.

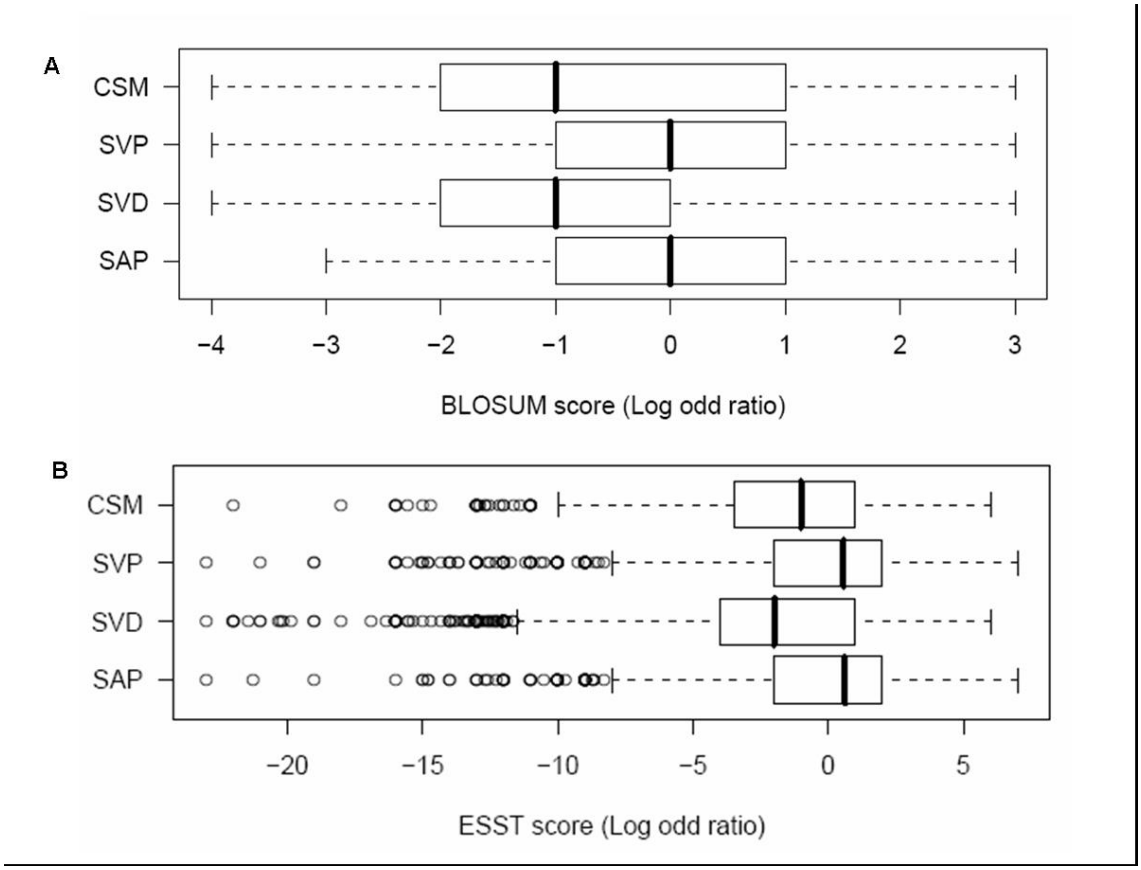


Figure 4-2 Box plots of substitution scores from four types of variants in the dataset

Each box plot is derived from the four variant datasets (see Table 4-1) and data are plotted against the BLOSUM62 substitution table and ESST in A and B, respectively. The median value is represented as a bold vertical line within a box, which represents the interquartile range (IQR) where lower quartile (cut-

off at the lowest 25% of the data) and upper quartile (cut-off at the highest 25% of the data) are the left and right edges of the box. Two vertical lines extended from the left and right hand sides of a box represent the smallest (left whisker) and largest (right whisker) non-outlier observations, respectively. Any data observation that lies more than 1.5*IQR lower than the lower quartile or 1.5*IQR higher than the upper quartile is considered an outlier which is shown as a circle.

4.2.3.1 By solvent accessibility

Figure 4-3 shows box plots of substitution scores by solvent accessibility for the four types of variant dataset. Except for SVP, the median values of substitution scores in the core regions of proteins are always smaller than those from the surface regions. The difference in substitution scores between core and surface region is highly significant for both SVD and CSM ($P < 10^{-12}$) and significant for SVP ($P < 10^{-4}$), whereas it is not significant for SAP ($P < 0.78$). This suggests that, although variants occur less frequently at solvent inaccessible regions, their effect would be detrimental if they occurred at the solvent inaccessible regions. In addition, the average proportions of variants having negative values of substitution score are 63% and 55% for SVD and CSM respectively, whereas the average proportions are less than 40% for SVP and SAP (see Table 4-3).

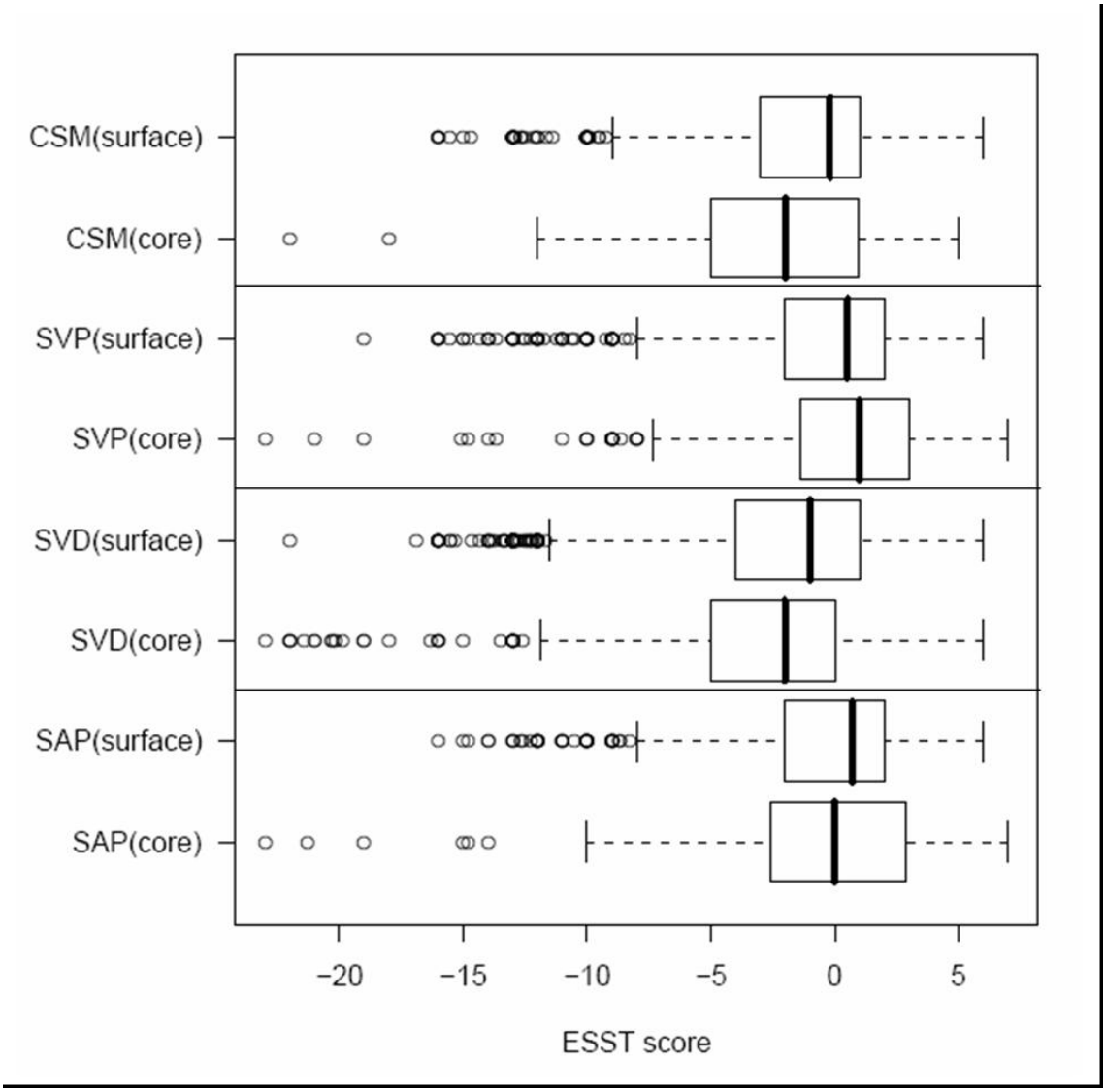


Figure 4-3 Box plots of substitution scores by solvent accessibility

Each of the four datasets is divided into solvent accessible (surface) and inaccessible (core) datasets. The representation scheme of a box plot is the same as shown in Figure 4-2.

Table 4-3 Ratios of variants having negative and non-negative substitution scores

Structural environment			Types of variants							
categories		types	SVD		SVP		CSM		SAP	
			<0	>=0	<0	>=0	<0	>=0	<0	>=0
Solvent accessibility		A[1]	0.58	0.42	0.38	0.62	0.51	0.49	0.37	0.63
		a[2]	0.70	0.30	0.36	0.64	0.67	0.33	0.41	0.59
Hydrogen bonds from sidechains	to main-chain amide	F[3]	0.62	0.38	0.37	0.63	0.54	0.46	0.38	0.62
		T[4]	0.7	0.3	0.45	0.55	0.67	0.33	0.43	0.57
	to main-chain carbonyl	F	0.63	0.37	0.37	0.63	0.53	0.47	0.37	0.63
		T	0.63	0.37	0.38	0.62	0.64	0.36	0.43	0.57
	to other side chains	F	0.63	0.37	0.37	0.63	0.54	0.46	0.38	0.62
		T	0.62	0.38	0.39	0.61	0.56	0.44	0.37	0.63
secondary structure		H[5]	0.59	0.41	0.4	0.6	0.52	0.48	0.4	0.6
		E[6]	0.65	0.35	0.3	0.7	0.58	0.42	0.35	0.65
		P[7]	0.79	0.21	0.62	0.38	0.68	0.32	0.62	0.38
		C[8]	0.61	0.39	0.35	0.64	0.52	0.48	0.35	0.65
All			0.63	0.37	0.37	0.63	0.55	0.45	0.38	0.62

[1]: accessible [2]: inaccessible [3]: False (no hydrogen bonds) [4]: True (hydrogen bonded)
[5]: α-helix [6]: β-strand [7]: positive φ main-chain torsion angle [8]: coil

4.2.3.2 By hydrogen-bond capacity

Figure 4-4 shows box plots for the distributions of substitution scores by existence or absence of hydrogen bonds from a side chain to a main-chain amide (Figure 4-4A), main-chain carbonyl (Figure 4-4B), and other side chains (Figure 4-4C). Overall, most of the median substitution scores for the residues making hydrogen bonds (NH/CO/SC) are smaller or equal to those from non-hydrogen bonding residues (nh/co/sc), which implies it would be more deleterious if variants were to occur at amino acids making hydrogen bonds. Indeed, the median values of SVD and CSM are negative for all three types of hydrogen bonds, although the difference is significant ($P < 10^{-3}$) only for amide (NH/nh) and carbonyl (CO/co) types of CSM dataset.

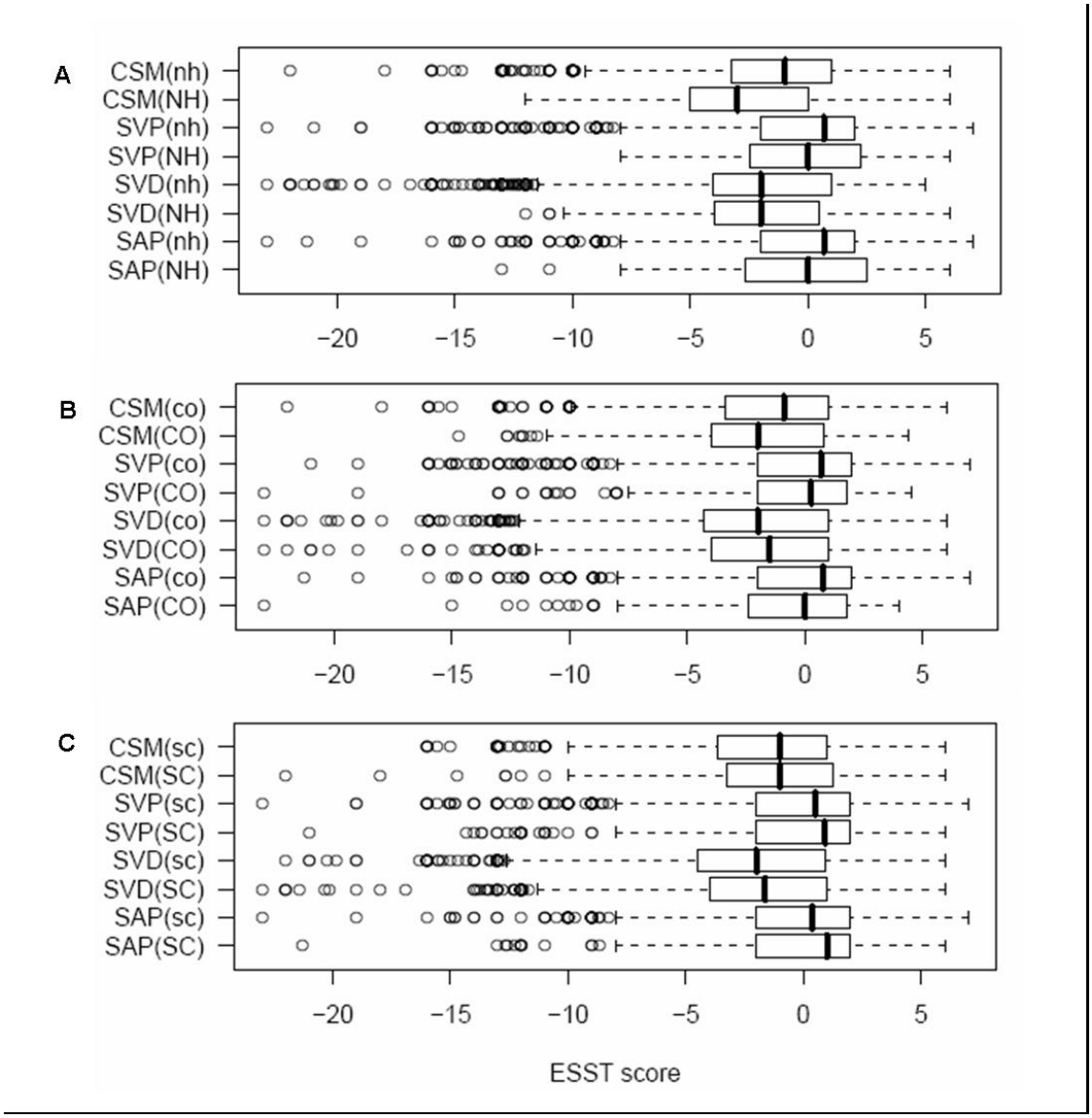


Figure 4-4 Box plots of substitution scores by hydrogen-bond types

A-C show box plots of substitution scores for the three hydrogen-bond types from a side chain: hydrogen bonds to amides (NH/nh), to carbonyls (CO/co), and to other side chains (SC/sc). The existence and absence of hydrogen bonds are shown in upper and lower case, respectively. The representation scheme of a box plot is the same as shown in Figure 4-2.

4.2.3.3 By elements of secondary structure

In Figure 4-5, substitution scores are plotted by class of secondary structure at the position where the variants occur. For SVD (Figure 4-5C) and CSM (Figure 4-5D), the median values are less than zero, regardless of secondary structures. Interestingly, for all variant types, those that occur at positive φ main-chain torsion angles (P) are always negative and they are significantly different ($P < 10^{-5}$) from the distributions of substitution scores for helix (H), beta (E) and coil (C). A positive φ torsion angle can be accommodated by a Gly, which has no side chain, but for most other L-amino acids it leads to disallowed interactions between side-chain and main-chain atoms. However, for L-amino acids such as Asp or Asn, interactions between the side-chain carbonyl group with the carbonyl of the main-chain peptide bond can give rise to relative stabilisation of a conformation with a positive φ angle [235]. Hence, sequence variants occurring at the residues within a positive φ torsion angle could be very deleterious and affect the native structures. For a positive φ torsion angle, I found that 55-57% of polymorphic variants (SVP and SAP) involve substitutions of amino acids from Gly, Asp and Asn, compared to 65-68% of SVD and CSM. This suggests that disease-causing mutations affect the native structure more frequently than neutral polymorphic variants (see Table 4-4).

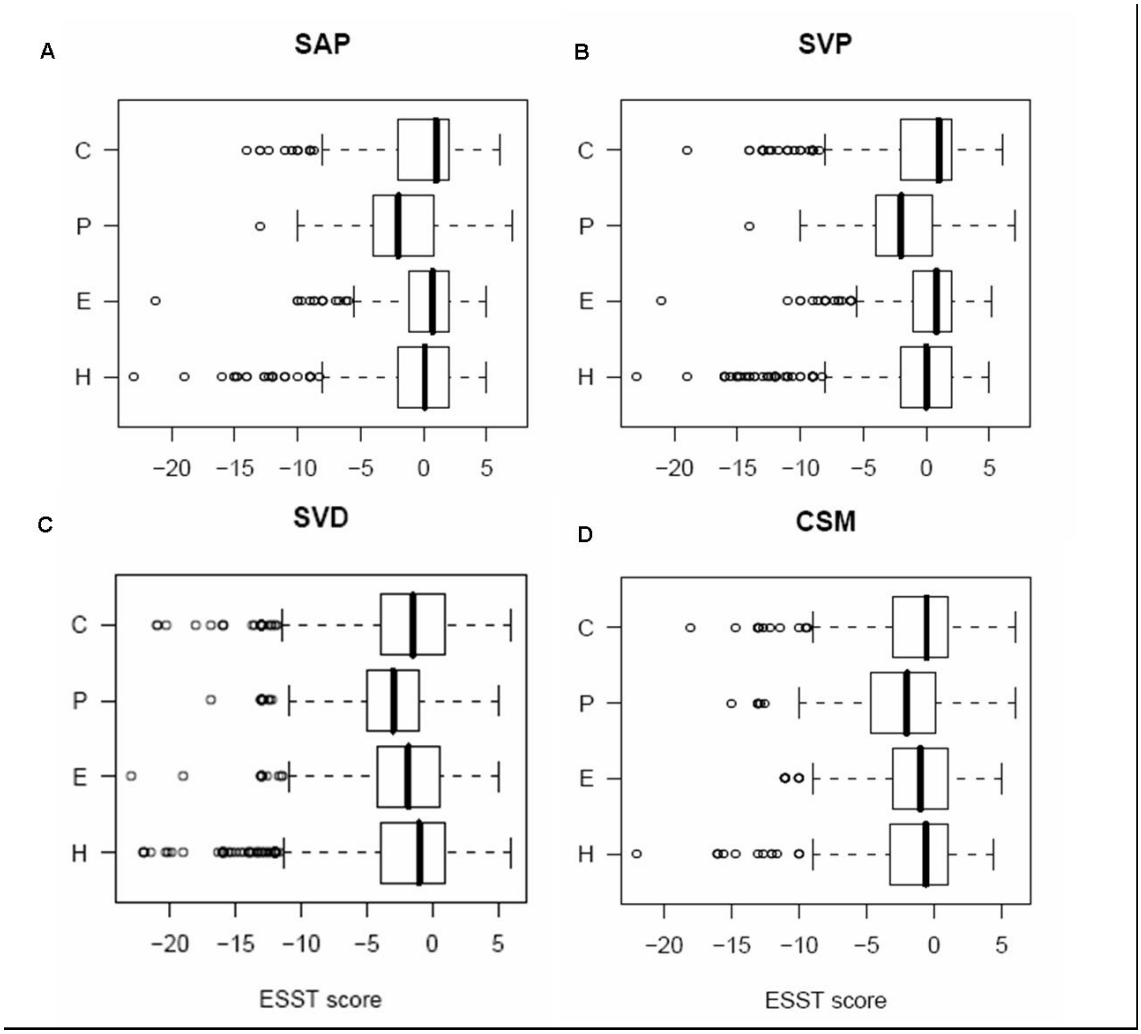


Figure 4-5 Box plots for the substitution scores by the class of secondary structure

A-D show box plots of substitution scores from four variant dataset (see Table 4-1) which are further divided by the element of secondary structures; α-helix (H), β-strand (E), coil (C) and residue with positive φ main-chain torsion angle (P). The representation scheme of a box plot is same as shown in Figure 4-2.

Table 4-4 Percentage (%) of amino acid variants occurring at positive φ main-chain torsion angle

Wild type Amino acids	SVD	SVP	SAP	CSM
G	58.59	42.64	44.06	55.65
R	6.11	11.68	13.29	6.09
N	4.20	7.11	7.69	2.17
A	4.01	3.55	1.40	0.43
D	3.24	6.60	7.69	6.96
S	2.86	5.58	4.90	5.22
C	3.63	1.02	3.50	0.00
E	2.29	3.55	3.50	4.35
F	1.91	1.02	1.40	0.00
M	2.67	0.00	0.00	1.74
L	2.48	2.54	2.10	1.30
Y	1.72	1.02	1.40	1.30
K	0.76	3.05	2.10	3.91
Q	1.72	2.54	2.10	1.30
T	1.53	1.02	0.70	0.43
H	0.95	3.05	2.10	3.48
V	0.76	1.52	0.70	0.87
I	0.38	0.51	0.00	3.04
W	0.19	1.52	0.70	0.00
P	0.00	0.51	0.70	1.74

4.2.4 Amino Acid Property Substitution Matrix

Substitution scores could be a proxy for the effect of variants, but do not provide any details of amino acid substitution types. To investigate this, 20 amino acids are classified into six types on the basis of physicochemical properties of amino acids (see Material and Methods) and 6 * 6 amino acid property substitution matrices are generated by counting the number of substitutions of amino acid by their types. Figure 4-6 shows amino acid property substitution matrices for the four types of variants in which the probability of substitutions is represented as heat maps. Aliphatic amino acids (Ala, Ile, Leu, Val and Met) from SVD (Figure 4-6C) and CSM (Figure 4-6D) are relatively less conserved than those observed from SAP (Figure 4-6A) and SVP (Figure 4-6B). In addition, amino acid substitutions from usually negatively charged (Asp and Glu) to positively charged (Arg, His and Lys) and aromatic (Phe, Trp, and Tyr) to polar non-charged (Cys, Asn, Gln, Ser and Thr) types are more frequently observed in SVD and CSM than those observed in SAP and SVP. In terms of substitution patterns, SVP and SAP are most similar, followed by SVD and CSM, whereas SVP and SVD are most different (see Table 4-5).

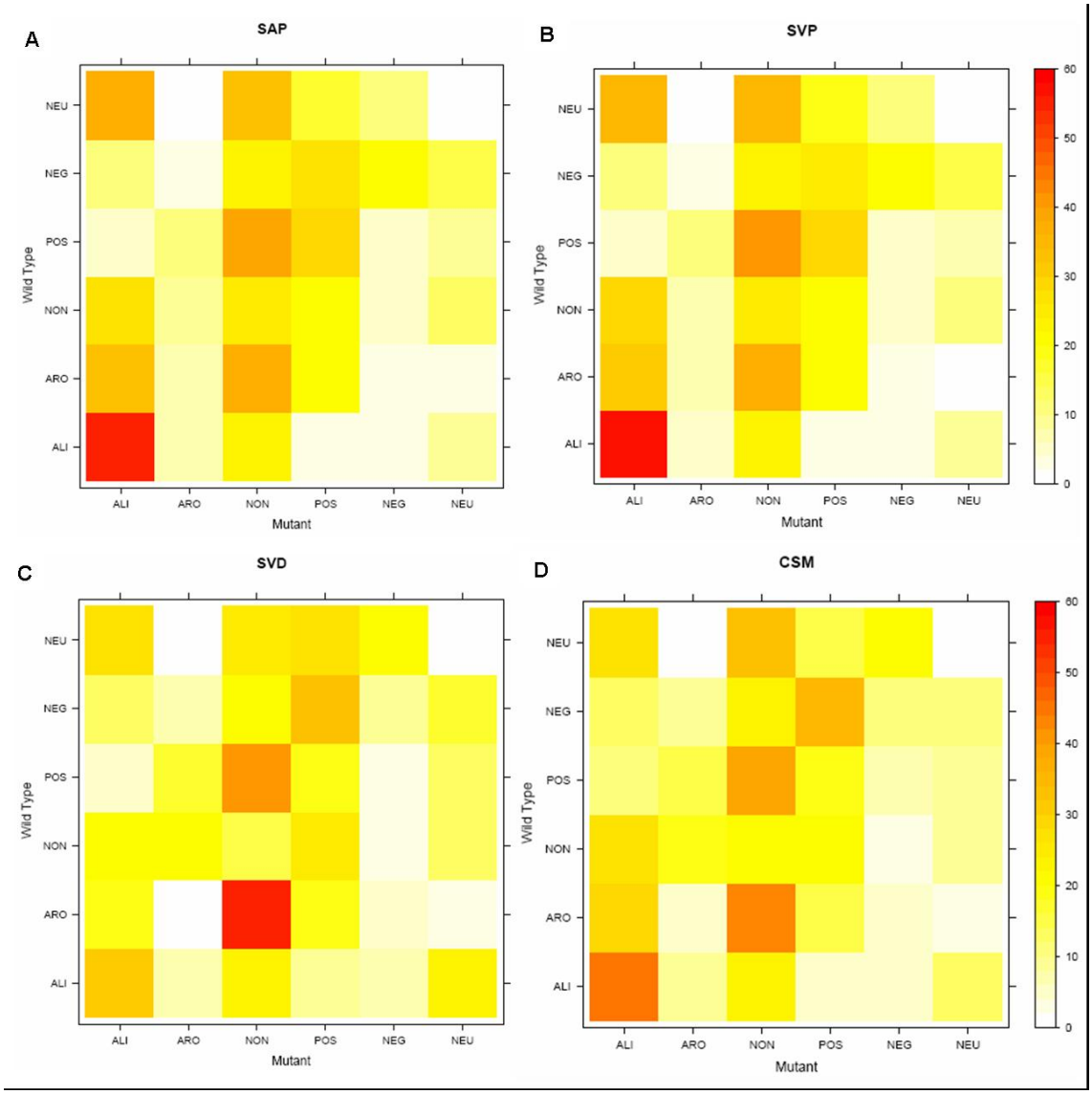


Figure 4-6 Amino acid property substitution matrices represented by heat maps

20 amino acids are classified into six types based on their physicochemical properties (see Materials and Methods) and the substitution probabilities among the six types are represented as heat maps. A-D are from the four variant datasets in Table 4-1. (ALI: aliphatic, ARO: aromatic, NON: polar non-charged, POS: positively charged, NEG: negatively charged, and NEU: neutral)

Table 4-5 Distance matrix of amino acid mutations from the four types of variants

	CSM	SAP	SVD
SAP	31.11		
SVD	26.72	43.86	
SVP	32.21	6.54	45.26

4.2.5 Degree of Sequence Conservation at the Variant Locations

I investigated the relationship between the variant types and the degree of sequence conservation at the locations where variants occur. Figure 4-7 shows box plots for the degree of sequence conservation measured by the Shannon's entropy (see Materials and Methods) from the four types of variants. In Figure 4-7A, it is very clear that Mendelian disease-related variants (SVD) occur at positions where amino acids are relatively conserved compared with those from polymorphic datasets (SVP and SAP) and cancer somatic mutations (CSM) with significant differences in the distribution ($P < 10^{-11}$). From Table 4-2, it is observed that the frequency of solvent inaccessible residues is much higher for SVD than those from SVP, CSM and SAP. Hence, the lower sequence entropy of SVD might arise from the relatively larger fraction of solvent inaccessible residues compared with the other variants, as solvent inaccessible residues are more conserved than solvent accessible residues. To address this issue, variants are classified into either solvent accessible (Figure 4-7B) or inaccessible environments (Figure 4-7C) and their sequence entropies were measured differently. I found that, regardless of their solvent accessibility, SVD occur at relatively conserved regions compared with variants from SVP, SAP and CSM ($P < 10^{-7}$ and $P < 0.0496$ from Figure 4-7B and Figure 4-7C, respectively). Interestingly, as shown in Figure 4-7B and Figure 4-7C, the median entropy value of CSM is higher than that of SVP and SAP, even though the distribution is not significantly different from that of polymorphic variants (*P-values* are <0.8071, <0.7032 and <0.1240 from Figure 4-7A, B and C, respectively). This observation contrasts with a current report that cancer-related mutations are frequently found at evolutionarily conserved amino acid residues whereas polymorphic variants occur in

relatively less conserved regions [246]. The conflict in this observation probably arises from the following reasons; i) differences in the nature of the 'cancer datasets' – in this study the COSMIC database was used whereas the report is based on curated lists of cancer mutations selected from the literature, ii) the use of different conservation measurements – Shannon's sequence entropy in this study whereas combinations of percentage identity and sequence-entropy by Talavera *et al.* iii) differences in the source and method of multiple sequence alignment – SCOP and Baton in this study whereas Ensembl-Compara [247] and MUSCLE [248] by Talavera *et al.*.

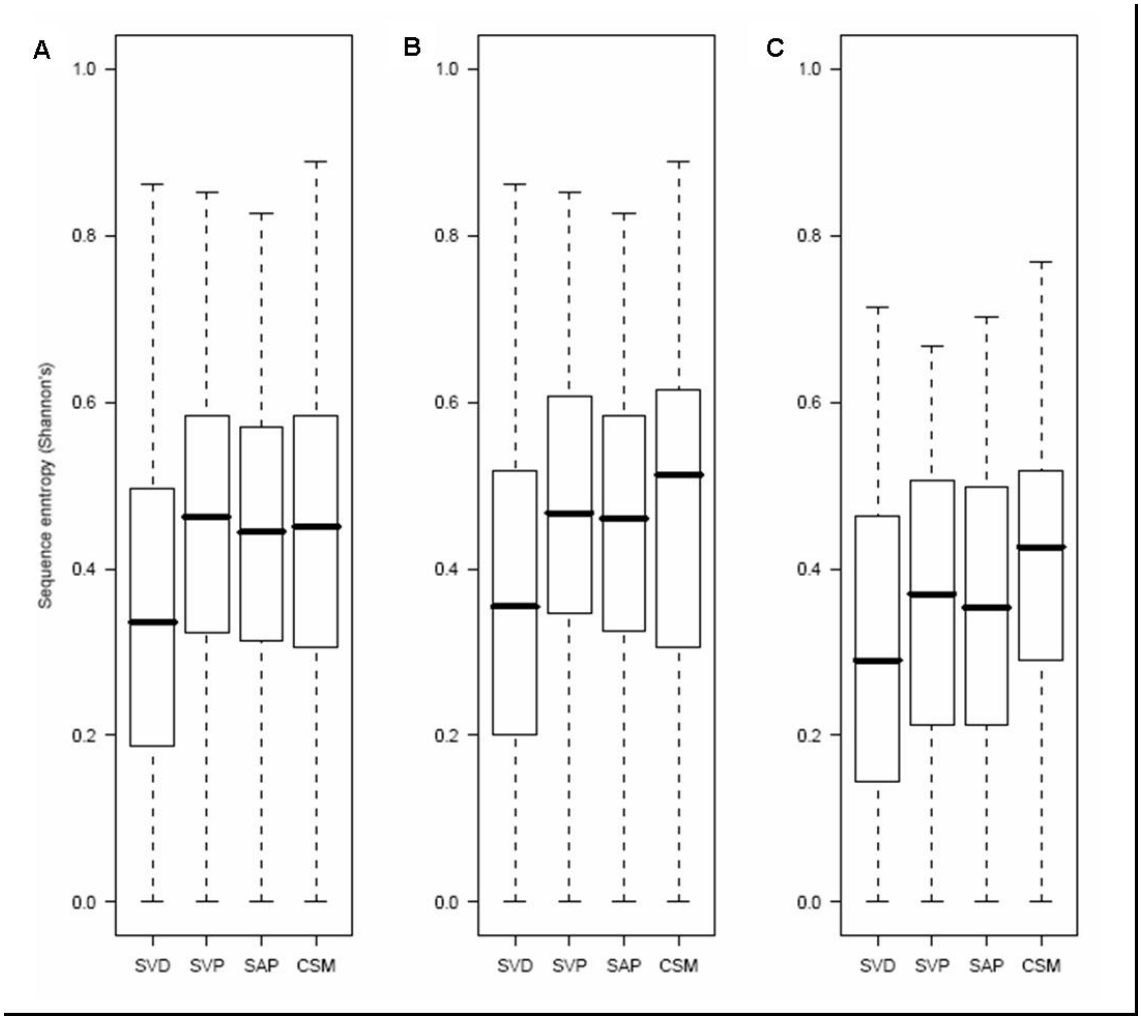


Figure 4-7 Box plots for the degree of sequence conservation measured by Shannon's entropy

Sequence entropies (see Material and Methods) from the four variant datasets (Table 4-1) are shown as box plots in A. Sequence entropies are calculated separately according to solvent accessibility of the

variants defined by where they occur in three-dimensional structures: solvent accessible (B) and inaccessible (C). The representation scheme of the box plots is the same as shown in Figure 4-2.

4.2.6 Functional Restraints

Amino acids responsible for specific functions of proteins tend to be conserved throughout evolution and are likely to be under strong restraints. Hence, mutations that do not improve or change function in a way that confers any selective advantage to the organism would likely be deleterious. To test this, I investigated variants occuring at amino acid residues responsible for protein function. Eight functional feature types are used, defined by UniProt annotations − ACT_SITE, BINDING, CA_BIND, DISULFID, DNA_BIND, LIPID, METAL, and NP_BIND (see Material and Methods for details) − and protein-protein interaction information from the PICCOLO database, http://mordred.bioc.cam.ac.uk/piccolo/piccolo.php (Bickerton GR, Higueruelo AP, and Blundell TL (2010) PICCOLO: comprehensive, atomic-level characterization of structurally characterized protein-protein interactions. *manuscript in preparation*). Table 4-6 shows frequencies of functional residues having a sequence variant at such a position. Polymorphic variants (SVP and SAP) occur in less than 1% of functional residues, whereas Mendelian disease-related variants (SVD) occur from 1.47% for calcium-binding residues (CA_BIND) up to 10.47% for residues interacting with a metal ion (METAL). Cancer somatic mutations (CSM) occur less frequently than SVD for all functional categories, but more frequently than polymorphic variants except for two categories: BINDING (binding sites for chemical groups) and CA_BIND (calcium-binding regions).

Table 4-6 Proportion (%) of functional residues having at least one sequence variant

Functional categories[1]	Types of variants			
	SVD[2]	SVP[3]	CSM[4]	SAP[5]
DNA_BIND	4.65	0.31	2.00	0.29
DISULFID	6.52	0.10	0.20	0.13
NP_BIND	3.91	0.25	1.39	0.32
METAL	10.47	0.21	1.16	0.18
BINDING	10.43	0.52	0.29	0.63
ACT_SITE	7.24	0.30	0.72	0.36
CA_BIND	1.47	0.54	0.22	0.51
PPI	3.53	0.83	2.15	0.51

[1]:see Materials and Methods for definitions

[2]: see Supplementary DatasetS2 of [238], [3]: see Supplementary DatasetS4 of [238]

[4]: see Supplementary DatasetS8 of [238], [5]: see Supplementary DatasetS6 of [238]

In order to illustrate these features, I examined a number of specific cases. As an example, Figure 4-8 exemplifies amino acid variants occurring at functional residues mentioned above from the following four UniProt entries: O14832, P00533, P24941, and O00204 for A-D, respectively. In Figure 4-8A, there are 17 sequence variants annotated by UniProt, one of which (VAR_050528) is annotated as polymorphic (SVP) and the rest are disease-related variants (SVD) responsible for Refsum disease (RD) [249,250,251]. Amongst 16 disease-related variants, two occur at metal-binding (METAL) and two at ligand-binding (BINDING) residues, which are directly responsible for the disease by inducing the loss of activity for the protein [28,249,251]. Figure 4-8B illustrates the locations of cancer somatic mutations occurring at the kinase domain of EGFR (Epidermal Growth Factor Receptor). There are 10 ATP-binding sites and one active site residue of which 8 ATP-binding sites are reported amongst somatic mutations responsible for lung cancer. Figure 4-8C and Figure 4-8D show variants in a protein kinase 2 (CDK2) and an alcohol sulfotranferase (SULT2B1), respectively. Two polymorphic variants (Y15S and V18L) occur amongst 19 ATP-binding residues in Figure 4-8C and only one polymorphic variant (V225I) out of 53 adenosine diphosphate binding residues in Figure 4-8D. The full list of all individual variants mentioned above

is available as Supplementary DatasetS2, S4 and S6 in an initial report of this work in PLoS ONE [238].

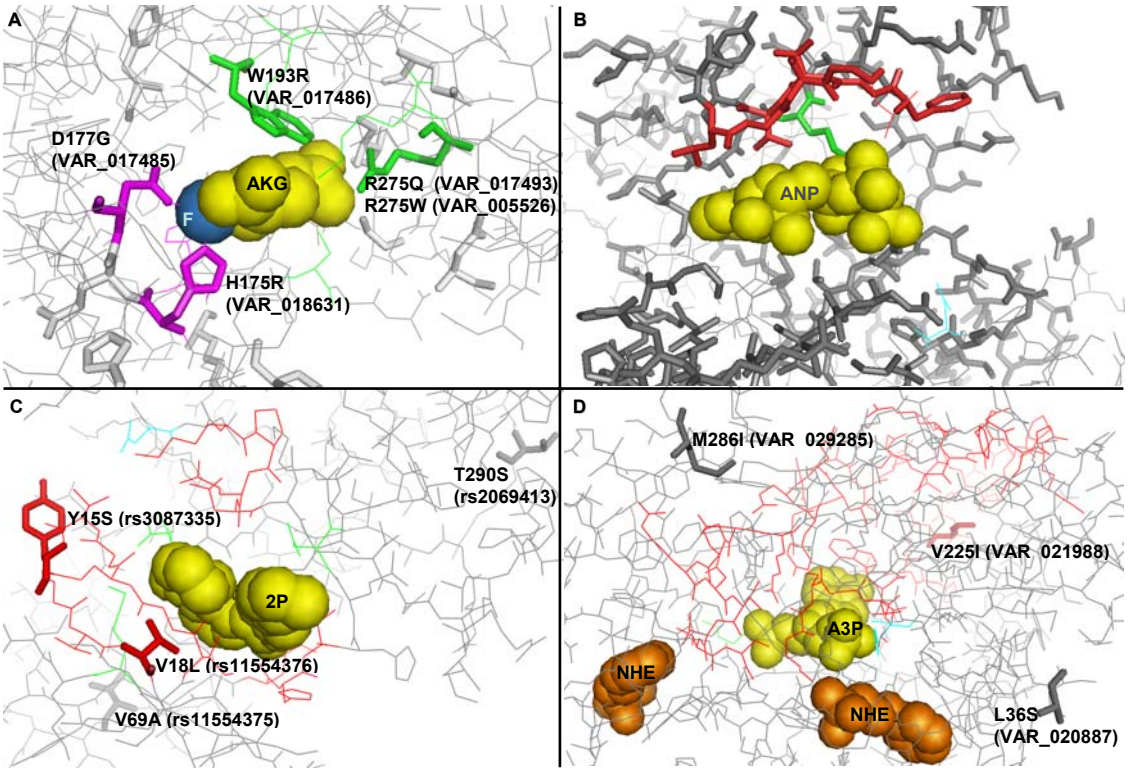


Figure 4-8 Examples of amino acid variations from the four datasets

UniProt feature annotations are transferred onto three-dimensional structures of proteins by aligning UniProt sequences with their corresponding PDB sequences using double-map method [193] (see Materials and Methods): O14832 with 2a1x in **A**, P00533 with 2itv in **B**, P24941 with 1gij in **C**, and O00204 with 1q1q in **D**. The regions not shown in the alignments are indicated with blue arrows. Amino acid variants are shown within boxes of grey background in the alignments and as bold-frame in the structure images. Metals and ligands are illustrated as spheres. Metal-binding (METAL), ligand-binding (BINDING), nucleotide phosphate-binding (NP_BIND), and active sites (ACT_SITE) residues are coloured in magenta, orange, red and cyan, respectively, both in the alignments and structure images. All structure images and alignments are drawn using PyMOL [94] and Jalview [252], respectively. (AKG: 2-Oxyglutaric acid, Fe: Iron ion, ANP: Phosphoaminophosphonic acid-adenylate ester, 2PU: 1-(5-oxo-2,3,5,9b-tetrahydro-1h-pyrrolo[2,1- a]isoindol-9-yl)-3-(5-pyrrolidin-2-yl-1h - pyrazol-3-yl)-urea, A3P: Adenosine-3'-5'-diphosphate, NHE: 2-[n-cyclohexylamino] ethane sulfonic acid)

4.2.7 Concluding Remarks

In this chapter, I have shown that the occurrence of amino acid variants is affected by the structural and functional restraints. Based on the frequency of their occurrence in particular structural environments, disease-related variants occur more often at solvent inaccessible regions, and at amino acid residues making hydrogen bonds compared with polymorphic variants. Overall, substitution scores of Mendelian disease and cancer somatic mutations are lower than those of polymorphic variants, suggesting deleterious and harmful effects when they occur. However, I observe that there are polymorphic variants that have very low substitution scores, especially variants changing the physicochemical properties of amino acids. Indeed, the presence of polymorphic variants (SVP and SAP) in the dataset does not necessarily mean they are neutral with respect to the phenotypes. There are likely to be variants related to a certain disease type, which have not been identified yet. However, I have not attempted to predict sequence variants causing deleterious effects on protein structures and depriving functions, which eventually lead to a specific disease, as this has been addressed extensively by others [165,167,168,170,175]. See section 1.3.3 for computational methods predicting disease-related mutations. Rather, I focused on the distributions and occurrences of amino acid variants in terms of structural and functional features of proteins.

In terms of amino acid conservation score, I showed that variants responsible for Mendelian disease are more frequently observed at rather conserved regions compared with cancer mutations and polymorphic variants. To quantify conservation score, Shannon's sequence entropy, which basically measures relative frequency of symbols (amino acids), was used to measure conservation score in this study. However, there are many other measurements and even there are some variations within the same entropy-based scoring method. See [253] for in-depth review on various residue conservation methods. One of the short comings of entropy-based methods is that most of these scoring schemes do not take account of gaps (e.g. columns dominated by gaps would score as more conserved). In this study, I did not measure entropy if gaps occur in more than 50% of the sequences at the alignment position. Even with this drawback, I believe

Shannon's sequence entropy method can reveal the relative degree of amino acid conservation amongst the four variation data sets analysed in this study.

4.3 Materials and Methods

4.3.1 Variants Data Source

SVD and SVP are defined by annotations of UniProt human sequence variations (http://www.uniprot.org/docs/humsavar.txt, release: 57.5) where types of amino acids variants are classified either disease, polymorphism or unclassified [244]. For SVD, variants are further filtered out by removing non-Mendelian diseases which have not been assigned any MIM number from the OMIM (http://www.ncbi.nlm.nih.gov/omim/) database and any disease names related with cancers from the following key tokens: cancer, tumor, neoplasia, leukaemia, lymphoma, melanoma, carcinoma, blastoma, and cytoma. CSM is taken from the COSMIC (Catalogue of Somatic Mutation in Cancer, http://www.sanger.ac.uk/genetics/CGP/cosmic/, version: 42) database [140] from which mutations result in amino acid changes were taken and SAP is from the Ensembl human variation database (http://www.ensembl.org, database version: 54_36p) [245] which compiles SNPs (Single Nucleotide Polymorphisms) mainly from dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) [128]. From Ensembl human variations, only verified SNPs have been used; those genotyped and validated by the international HapMap project [151]. Amino acid variants of CSM and SAP were transferred onto the positions of their corresponding UniProt sequence using the sequence alignment program, BL2SEQ, of NCBI blast package [64] if necessary.

4.3.2 Representative SCOP Domains

SCOP 1.71 was used to define representative domains by applying the following conditions:

1) NMR structures and proteins having resolution worse than 2.5Å were excluded.
2) Protein domains were clustered for each SCOP family by running CD-HIT [215] with sequence identity of 80% or more.

3) Within a SCOP family, the average sequence length is maintained by removing any domains having sequence below of (1-0.3)*mean-length and above of (1+0.3)*mean-length.
4) Within a cluster, a protein structure having the best resolution was selected as a representative.

Non-canonical SCOP classes (H, I, J, and K,) and membrane and cell surface proteins (F) were not included in the process described above.

4.3.3 Mapping the Location of Variants onto 3D Structure

To locate the position of a sequence variant in the three-dimensional structure, variants mapped onto UniProt sequences were further transferred onto three-dimensional structures using double-map [193] which aligns a sequence of UniProt to its corresponding PDB structure at residue level. In short, double-map makes two alignments from the three sequences. The first alignment is between a sequence in atomic coordinate record (SEQATM) and SEQRES record of a PDB file. The second is between SEQRES and its corresponding UniProt sequence (SP). Using SEQRES as a reference SP can be aligned with SEQATM and the locations of UniProt residues can be mapped onto three-dimensional structures. Detailed description of the mapping procedure is available from section 2.3.2 and an online database is implemented to share pre-run data, which is described in Chapter 6. In parallel, there are public resources which also provide genetic variations mapped onto three-dimensional structures of proteins [142,184,254,255].

4.3.4 Identifying Local Structural Environment of Amino Acids

JOY [60] was used to identify the local structural environments of amino acids. JOY consists of three supporting programs − SSTRUC, PSA, and HBOND − to annotate 1) the elements of secondary structure, 2) solvent accessibility, 3) hydrogen bonds from side chains, respectively. SSTRUC, a successor of DSSP [256], calculates torsion angles within a main chain to assign secondary structure. For the threshold of solvent accessibility, a cut-off of 7.0% relative total side-chain accessibility was used. HBOND

identifies all possible hydrogen bonds based on a distance criterion; 3.5Å between donor and acceptor except for interactions involving sulphur atoms where 4.0Å is used.

4.3.5 Amino Acid Substitution Scores

For variants at the UniProt protein sequence level, BLOSUM62 [82] was used to get the substitution score for a corresponding variant. However, substitution scores for the variants mapped onto three-dimensional structures were from an Environment Specific Substitution Table (ESST) [88,89], which corresponds to the local structural environment for a variant. I used ALL-B types of ESST, which has proved to be the best approach in previous benchmarking tests [193]. The detailed procedure of making ESSTs is explained in Chapter 2 and the ESST web site (http://samu.org/ESST). ESST can be generated in an automatic fashion by the recently developed computer software, Ulla [93].

4.3.6 Statistical Analysis

The Wilcoxon rank sum test was used to calculate significant differences in the distribution of substitution scores between two groups. I used *wilcox.test* of *stats* package of R [257] with a two-sided test option.

4.3.7 Classification of Amino Acid Types

20 amino acids are classified into 6 classes by their physicochemical properties as follows:

1) Aliphatic (ALI): Ala, Ile, Leu, Val and Met
2) Aromatic (ARO): Phe, Trp, and Tyr
3) Polar non-charged (NON): Cys, Asn, Gln, Ser and Thr
4) Positively charged (POS): Arg, His and Lys
5) Negatively charged (NEG): Asp and Glu
6) Neutral (NEU): Gly and Pro

4.3.8 Measuring Distances from Substitution Matrices

The Euclidean distance ($\mathrm{DIST}(\mathrm{X}\cdot\mathrm{Y})$), between two amino acid property substitution matrices, X and Y, defined as;

$$\mathrm{DIST}(\mathrm{X}\cdot\mathrm{Y}) = \left(\sum_{j=1}^{6} \sum_{k=1}^{6} \left(X_{j\to k} - Y_{j\to k} \right)^2 \right)^{1/2}$$ where $X_{j\to k}$ and $Y_{j\to k}$ are the probabilities of amino acid category *j* to be substituted by category *k* from the variant dataset X and Y, respectively.

4.3.9 Sequence Entropy

To measure the degree of sequence conservation, sequence entropy was calculated for each alignment position within a protein family having at least three sequences. Entropy was not measured if gaps occur in more than 50% of sequences at the alignment position; otherwise, gaps were treated as another symbol. Shannon's entropy equation [258] was formulated as below:

$$\text{Sequence entropy} = \frac{-\sum_{i}^{21} p_i \log_2 p_i}{\log_2 21}$$

where p_i is the frequency of symbol *i* (either an amino acid or a gap) at the alignment position.

4.3.10 Definitions of Functional Residues

Variants taken from the four types of dataset were examined to see whether they occur at protein residues responsible for specific functions. Functional residues were defined if they were annotated by UniProt functional features (from 'FT' lines) or known to maintain protein interactions detected by PICCOLO (http://mordred.bioc.cam.ac.uk/piccolo/piccolo.php; Bickerton et al. 2010; In Preparation) which is an in-house database of protein-protein interactions between every pair of chains from protein structures in the PDB. Eight types of UniProt functional features were used:

1) ACT_SITE: amino acid(s) involved in the activity of an enzyme

2) BINDING: binding site for any chemical group (e.g. co-enzyme, prosthetic group, etc.)
3) CA_BIND: extent of a calcium-binding region
4) DISULFID: disulfide bonds
5) DNA_BIND: extent of a DNA-binding region
6) LIPID: covalent binding of a lipid moiety
7) METAL: binding site for a metal ion
8) NP_BIND: extent of a nucleotide phosphate-binding region

Chapter 5

Structural and Functional Analysis of Amino Acid Variants identified in Type 1 Diabetes Genome-Wide Association Studies

Understanding the genetic basis of a phenotype has long been an attractive, yet challenging, subject of study for molecular biologists. Indeed, interrogation of the genetic make-up responsible for a certain disease phenotype has been a major focus in the post-genomic era. Recent advances in next-generation sequencing technologies are producing large amounts of genetic data in a very fast and large-scale manner, and this is revolutionizing the way we study genotype-phenotype relationship. With the help of a statistical framework comprising linkage disequilibrium and genome-wide association studies, we can now start to understand disease aetiology underlying common diseases such as cancer and diabetes. However, such statistical analyses do not provide molecular and physiological details of disease susceptibility, which is required in order to confirm associations between genetic make-up and disease aetiology. In the previous chapter, I characterized structural and functional features of amino acid variants in human proteins from various data sources comprising neutral polymorphic variations, somatic mutations and disease-associated variants. In this chapter I focus on a specific example of a complex disease, type 1 diabetes, and describe structural and functional analyses of amino acid variants identified from genome-wide association studies of type 1 diabetes.

5.1 Introduction

Early analyses of protein structure showed that single amino acid substitutions or mutations are often disease associated [111]. Most monogenic diseases, such as sickle cell disease and severe combined immunodeficiency (SCID), appear to result from a single DNA variant resulting in an amino acid substitution, which affects protein stability rather than impairing protein function directly [112] (see 1.3.1 for details). Therefore, methods that predict the effects of mutations on protein stability are useful for identifying possible disease associations [115,160]. Indeed, several computer programs successfully identify protein mutations that affect protein stability [161,162,163,165,167,170,174,259]. However, for most common diseases, such as cancers, heart diseases, and diabetes, where multiple genes and alleles play a role in complex phenotypes or traits, pinpointing the genetic loci underlying diseases has never been easy and has become even harder, especially when genetic variants responsible for disease aetiology need to be identified. With the help of recent advances in sequencing technologies [123,260] and analytical frameworks (see [126,261] for review), we are now beginning to see successful case studies, identifying the genetic loci underlying the aetiology of complex diseases such as type 1 [262,263] and 2 diabetes [264,265], asthma and coronary heart disease [124,266]. More recently, systematic resequencing of the cancer genome has revealed genetic changes that may be responsible for lung, breast and colorectal cancer [122,158,159,267]. Lists of genetic loci associated with disease susceptibility from the published studies are deposited in databases such T1Dbase[24] [149], COSMIC[25] [140], EGA[26], and a Catalog of Published Genome-Wide Association Studies [27] of NHGRI (National Genome Research Institute). Therefore, our understanding of the genetic basis of complex diseases is beginning to improve with the help of large-scale genome-wide association studies (GWAS) and high-throughput sequencing technologies, although more molecular and physiological studies of genetic variants need to follow in order to confirm association with disease aetiology.

[24] http://www.t1dbase.org
[25] http://www.sanger.ac.uk/genetics/CGP/cosmic/
[26] http://www.ebi.ac.uk/ega
[27] http://www.genome.gov/gwastudies/

In Chapter 4, I described structural and functional restraints that shape the occurrence of single amino acid variations in neutral polymorphisms, cancers and Mendelian diseases. In this chapter, I focus on an example of a complex disease—type 1 diabetes (T1D)—and present functional and structural analyses of genetic variations related to the disease. The genetic variations, which are presumably responsible for T1D, are provided by the research group of Professor John Todd[28], Cambridge Institute of Medical Research (visit http://www.t1dbase.org for details). Many genetic regions (e.g. chromosomal loci) associated with T1D have been identified through genome-wide association analysis; testing a number of common SNPs to see if different alleles show different frequencies in a large number of cases and controls. All regions contain many variants, of which only a minority will show association with T1D; those that lie on the same ancient haplotypes as the causal variant(s). Seven of these regions were chosen for sequencing in a selection of 80 samples (a mixture of cases and controls) to assemble a more complete catalogue of variation, and were further assessed statistically for association with T1D using an imputation method.

Here, I present an analysis of 355 SNPs—two lead to base 'deletion', so are omitted from this analysis—from which I characterize functional and structural environments of the amino acid variants by mapping their locations within UniProt and PDB, respectively, using Ensembl API[29] and double-map [238] introduced in Chapter 2. Sequence variants and their analyses described in this chapter are available from http://samul.org/T1D/353snps.

5.2 Results and Discussions

5.2.1 Overview

The 353 SNPs are from 51 genes (or contigs) spanning six chromosomes of which chromosome 12 and 16 account for almost 60 % (210/353) (see Table 5-1). Among 353 SNPs, 192 and 34 SNPs are mapped onto 129 UniProt and 225 PDB entries, respectively. Not all the 353 SNPs could be mapped onto their equivalent amino acid

[28] http://www-gene.cimr.cam.ac.uk/todd/index.html
[29] http://www.ensembl.org/info/docs/api/index.html

positions within their corresponding proteins and further to the known three-dimensional structure for the following two reasons; i) 161 SNPs (353 – 192) are located within non-protein coding regions, and ii) 158 SNPs (192 – 34) are within the UniProt proteins which do not have their three-dimensional structures available from the PDB at the time of this analysis. Comparative modelling can help increase the number of SNPs that can be analysed within the structure space, but those SNPs were only analysed in term of their equivalent positions within close homologs having known three-dimensional structures.

Table 5-1 353 T1D-related SNPs from 51 genes[30]

Chromosome	Gene or contig name	NO of distinct SNP within		
		Ensembl gene (ENSG)	UniProt	PDB
4	KIAA1109	34	34	0
16	CLEC16A	28	9	0
16	CIITA	28	18	0
12	ANKRD52	22	10	0
2	IFIH1	15	13	6
10	IL2RA	14	8	3
12	ERBB3	13	10	4
12	SUOX	10	4	0
12	STAT2	10	5	1
12	IKZF4	9	2	0
2	KCNH7	9	8	0
10	PFKFB3	8	5	4
18	CD226	8	4	0
12	PAN2	8	4	0
10	RBM17	8	2	0
12	RAB5B	8	0	0
12	ESYT1	8	7	0
12	SLC39A5	7	6	0
2	CTLA4	7	1	0
12	SMARCC2	7	3	0
12	OBFC2B	7	0	0
12	CDK2	7	2	2
12	RNF41	6	1	1
2	FAP	6	6	6
12	COQ10A	6	2	0
4	ADAD1	5	2	0

[30] see http://samul.org/T1D/353snps for details

18	DOK6	5	1	0
12	APOF	5	4	0
12	CS	5	3	0
10	RP11-414H17.1	5	0	0
2	ICOS	4	1	0
12	CNPY2	3	2	0
12	AC034102.1	3	0	0
12	MYL6	3	3	1
2	GCA	3	1	1
4	IL21	3	1	1
4	AC097533.2	3	0	0
10	7SK	3	0	0
4	IL2	3	1	1
12	SILV	2	2	0
12	PA2G4	2	2	2
12	RPS26P20	2	1	0
16	DEXI	2	0	0
12	ZC3H10	2	1	0
2	GCG	2	2	0
2	5S_rRNA	1	0	0
10	AL137186.1	1	1	1
12	IL23A	1	1	1
2	AC007750.1	1	1	1
12	DGKA	1	0	0
4	AC097533.1	1	0	0
Total		353	192	34

Table 5-2 shows the number of SNPs classified by the nature of their consequences. 40% (142/353) are located within intronic regions, and 29% (102/353) in non-synonymous (ns) coding, while two result in stop-gained codon responsible for premature forms of gene products from IL2RA (interleukin-2 receptor alpha chain) and COQ10A (coenzyme Q-binding protein COQ10 homologue A); from these I selected several interesting SNPs, which are further described below. From the 100 nsSNPs in Table 5-2, 41 SNPs are mapped onto UniProt protein residues where the same locations are already identified as variation sites by dbSNP [128]. Interestingly, it is reported that two of them are associated with T1D according to UniProt annotations; these two are further analysed.

Table 5-2 Numbers of SNPs grouped by their consequence types[31]

SNP types	Number of distinct SNPs		
	Ensembl transcript (ENST)	UniProt	PDB
Intronic	142	0	0
3' UTR[32]	138	0	0
Non synonymous coding (nsSNPs)	102	100	13
Synonymous coding	90	90	21
5' UTR	34	0	0
Coding region not found	27	0	0
Error[33]	7	0	0
Stop gained	2	2	0
Total	353[34]	192	34[35]

[31] See http://samul.org/T1D/353snps/gene/all/enst for details
[32] Untranslated Region
[33] Coding sequence does not seem to start with the initiation codon (AUG)
[34] Note that a SNP could result in more than one consequence type mainly by alternative splicing (one gene many transcript relationship)
[35] 34 SNPs are successfully mapped onto their corresponding location within the known three-dimensional structures; the remaining variants (192 - 34) have no structure information available from PDB

In order to characterise functional features of the amino acid variants from 100 nsSNPs, UniProt annotations[36] were investigated, of which 26 feature types have been used in this analysis (see section 5.3.3 for details). I found 45 SNPs are within the amino acids annotated as 'VAR_SEQ', three for 'REGION', two for 'TRANSMEM' and one for 'ZN_FING' (see Table 5-3); these variants are also investigated further.

Table 5-3 Functional annotations of 100 non-synonymous SNPs

(for details, visit http://samul.org/T1D/353snps/gene/all/uniprot/NON_SYNONYMOUS_CODING)

Annotation	Definition	NO. of SNP
N/A	Annotations not available	73
VAR_SEQ	Description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting	45
REPEAT	Extent of an internal sequence repetition	6
COMPBIAS	Extent of a compositionally biased region	4
REGION	Extent of a region of interest in the sequence	3
TRANSMEM	Extent of a transmembrane region	2
SIGNAL	Extent of a signal sequence (prepeptide)	2
PEPTIDE	Extent of a released active peptide	1
ZN_FING	Extent of a zinc finger region	1
PROPEP	Extent of a propeptide	1
CARBOHYD	Glycosylation site	1

Amino acid substitution models such as PAM [80] and BLOSUM [82] describe the degree of substitutions as log-odd ratio values where the positive scores suggest commonly occurring and preferred substitutions, whereas the negative scores imply very rare substitutions which are disfavoured in nature. An ESST, which I have described in Chapter 2, also addresses the degree of substitution of amino acids, but differs from PAM or BLOSUM by taking into account structural environments, thus conferring a more detailed description of substitution patterns. Figure 5-1 shows box plots of substitution scores from the 100 nsSNPs mapped onto UniProt (see Table 5-2), measured by ESST, PAM and BLOSUM matrices. The median substitution score is 0 for both PAM and BLOSUM, whereas it is one for ESST. 23% (3/13), 41% (41/100),

[36] http://www.expasy.ch/sprot/userman.html#FT_line

and 47% (47/100) of substitutions are negative, according to ESST, PAM and BLOSUM, respectively; these are further analysed below.

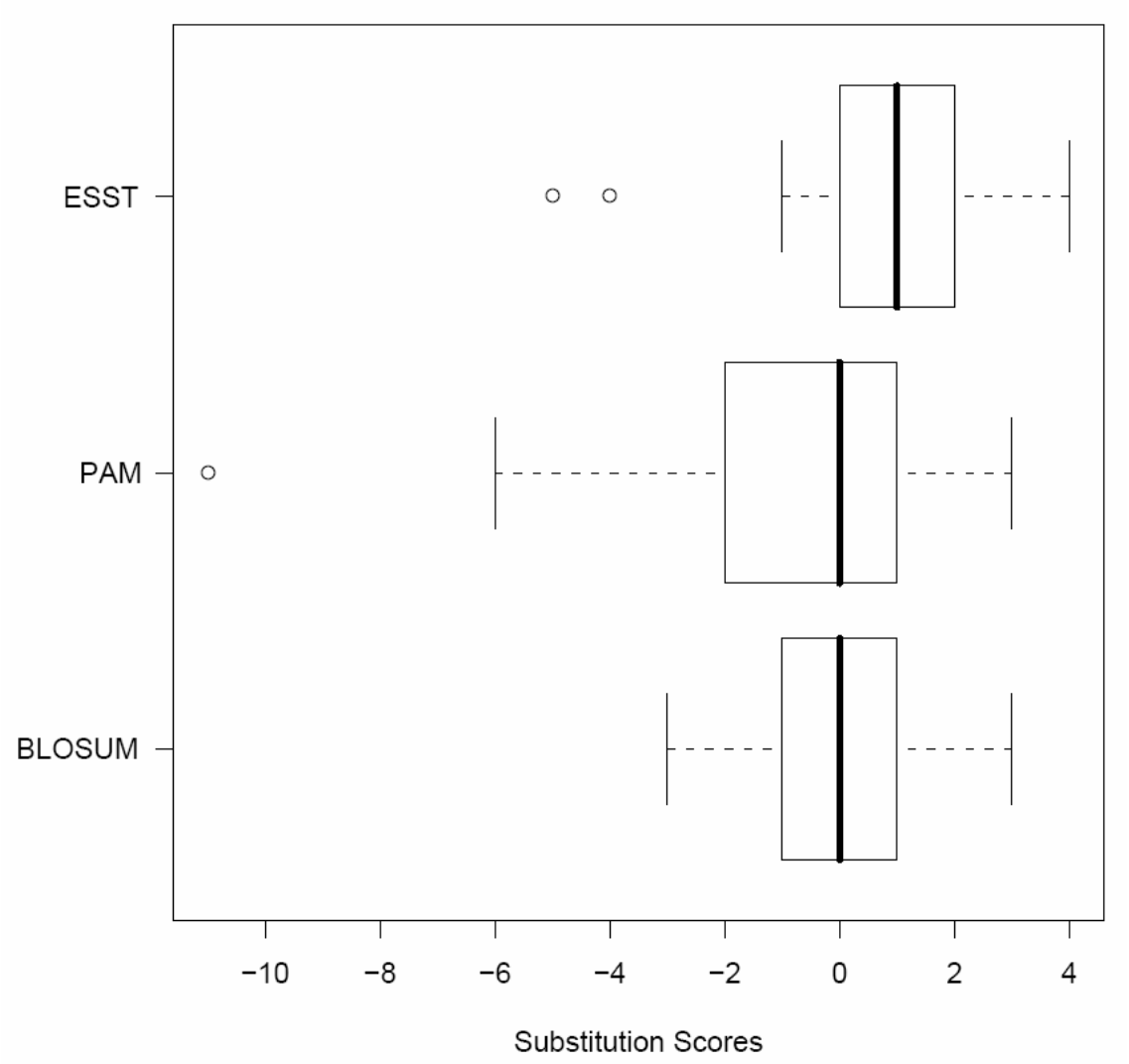


Figure 5-1 Box plots of substitution scores for the 100 non-synonymous SNPs

Substitution scores for the 100 nsSNPs (see Table 5-2) are estimated by ESST, PAM70 and BLOSUM62, shown in the Y-axis. The representation scheme of a box plot is the same as for Figure 4-2. ALL-B types of ESSTs were used; this has proved to be the best approach in previous benchmarking tests as described in Chapter 2 and [193]. Note that not all nsSNPs have their ESST scores; only 13 nsSNPs are assigned with their corresponding ESST scores due to limited numbers of three-dimensional structures.

I now describe two stop-gained SNPs and a selected number nsSNPs, which are likely to be related to T1D based on the functional and structural assessments of amino acid

residues where the variants occur in their corresponding proteins. Equivalent positions of variants in homologues are interrogated instead if the three-dimensional structures are not available for some cases. All the SNPs described in this chapter are listed in Appendix II with their gene name, equivalent Ensembl identifiers, chromosome locations and their 5' and 3' sequences altogether. SNPs, both in this chapter and Appendix II, are given with their special placeholder names starting with the prefix 'jtt1d_' followed by a numeric identifier. Appendix III lists substitution scores described in Figure 5-1.

5.2.2 Two Stop-gained SNPs

Two stop-gained SNPs—jtt1d_102 (Y239*) and jtt1d_250 (E243*)—are found in the C-terminal region of an interleukin-2 receptor subunit alpha (IL2RA) and a coenzyme Q-binding protein COQ10 homolog A (CQ10A), respectively (see Figure 5-2). From the 353 SNPs, eight genetic variants[37] are within the coding region of IL2RA, of which four are synonymous, three non-synonymous—T91M (jtt1d_107), M113I (jtt1d_105) and M113R (jtt1d_106)—and one stop-gained SNP. The dbSNP reports that seven non-synonymous SNPs are in the coding region of IL2RA, of which one nsSNP (rs41290331) corresponds to jtt1d_107. Two genetic variants[38], out of the 353 SNPs, are within the coding region of CQ10A; one synonymous and one stop-gained. IL2RA contains a potential transmembrane region and a cytoplasmic domain at the C-terminus, which would be lost by truncation if the stop-gained SNP occurs. This would prevent signal transduction by interleukin-2, typically observed in immune cells such as lymphocytes. Hence, this stop-gained SNP could contribute to susceptibility to T1D. Indeed, the John Todd group already reported that genetic variations in IL2RA are associated with susceptibility to insulin-dependent diabetes mellitus type 10 (IDDM10[39]) and T1D from the genome-wide association studies. [268,269].

[37] http://samul.org/T1D/353snps/gene/IL2RA/enst
[38] http://samul.org/T1D/353snps/gene/COQ10A/enst
[39] http://www.ncbi.nlm.nih.gov/omim/601942

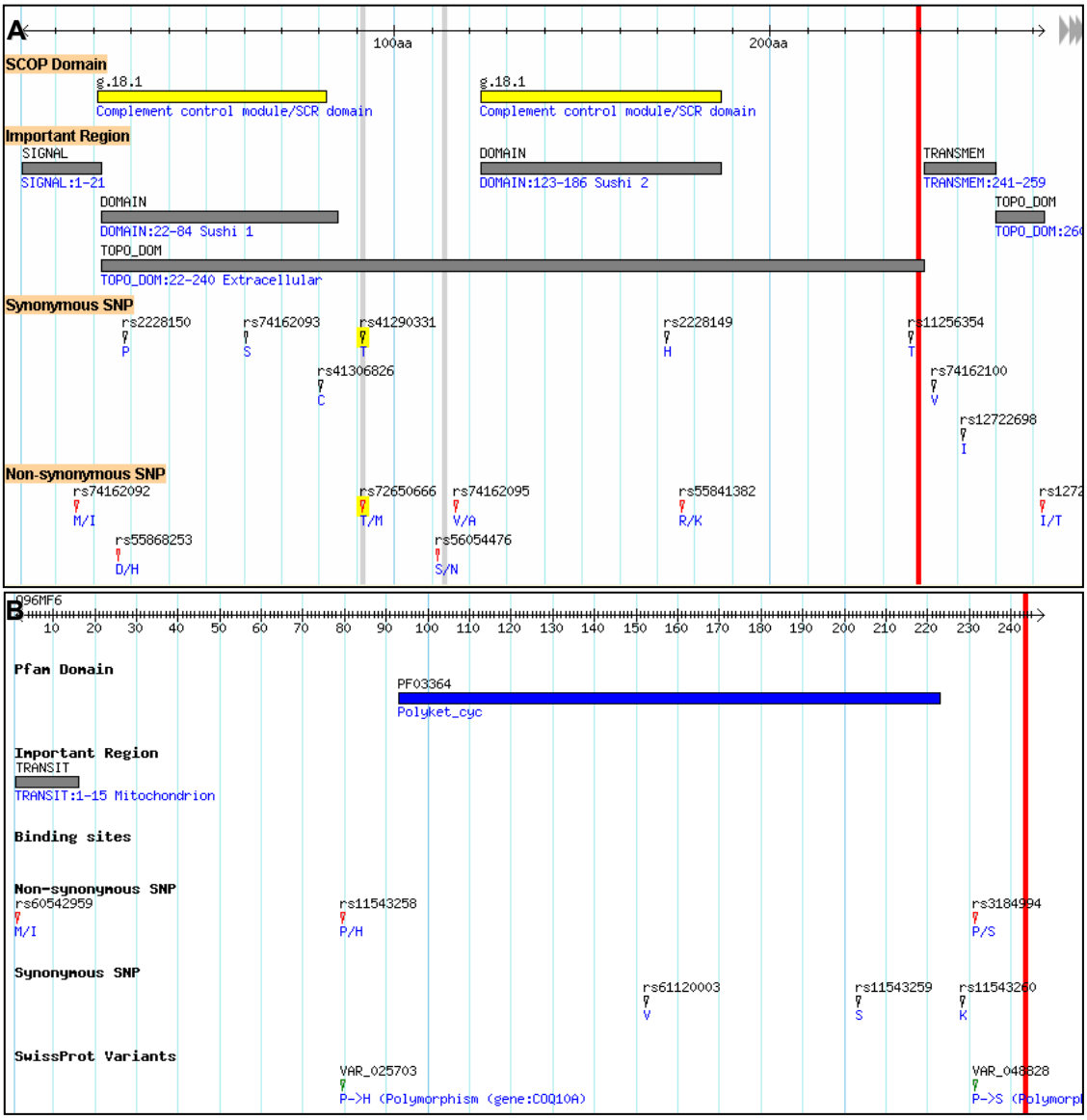


Figure 5-2 Schematic diagrams highlighting positions of two stop-gained SNPs

A and **B** illustrate the locations of two stop-gained variants (red vertical lines) within IL2RA and CQ10A, respectively. Two nsSNPs are indicated in grey vertical lines with their equivalent dbSNP identifiers highlighted in yellow. Arrows, spanning horizontally in the upper region for each picture, indicate the length of each protein. Cyan-coloured vertical lines are overlaid for every 10 amino acids across various annotations tracks with their titles in the left. Important regions, SCOP domains, and Pfam domains are indicated in grey, yellow and blue boxes, respectively. Figures are drawn using Gbrowse [270] and accessible from http://samul.org/T1D.

5.2.3 Analysis of non-synonymous SNPs

5.2.3.1 Variants in interferon-induced helicase C domain-containing protein 1 (IFIH1)

15 genetic variants[40] are found within DNA regions encoding interferon-induced helicase C domain-containing protein 1 (IFIH1). Among them, two are within introns and 13 in the coding region, of which 11 variants are non-synonymous SNPs (see Figure 5-3). Seven variants, indicated in yellow in Figure 5-3, are already deposited in dbSNP, of which rs1990760 (jtt1d_11, A946T) is reported to be associated with susceptibility to insulin-dependent diabetes mellitus [271,272,273]. Hence, only three are novel: jtt1d_10, jtt1d_19, and jtt1d_22. The variant (A946T) is located in the C-terminal region, as shown in Figure 5-3, with four metal (zinc) binding residues (residue 907, 910, 962 and 964) located nearby. The geometric distance between one of the zinc ions and Ala^{946} was measured to see whether the variation (jtt1d_11) could affect zinc binding physically, but this seems unlikely; the distance is approximately 20 Å (see Figure 5-4A). Substitution scores of the variation are also non-negative; 1 by ESST and PAM, 0 by BLOSUM. However, substitution scores of variant jtt1d_22 (V340G), located at the helicase ATP-binding domain (residues from 316 to 509), is very low; -3 both from PAM and BLOSUM, -5 by ESST. Based on the three-dimensional structure (PDB: 3B6E), Val^{340} is found in a solvent inaccessible region buried between two helices (see Figure 5-4B). Removal of two methyl groups can be very deleterious by removing the hydrophobic nature found in the wild-type amino acid residue.

[40] http://samul.org/T1D/353snps/gene/IFIH1

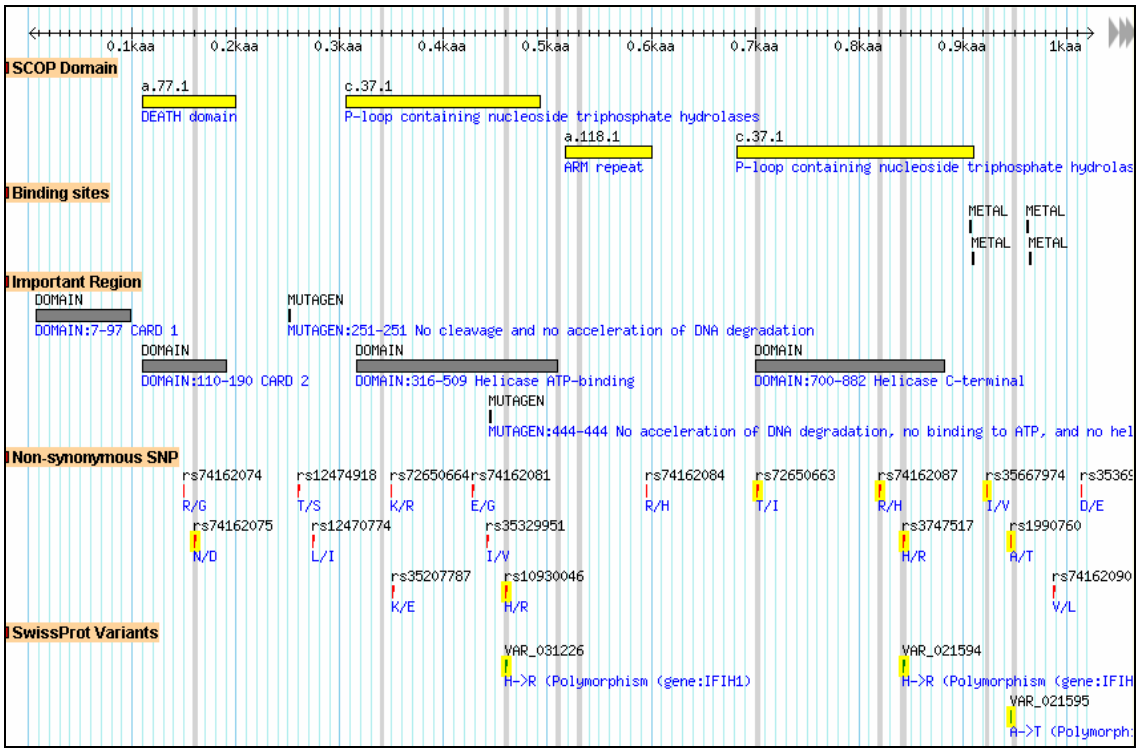


Figure 5-3 11 non-synonymous SNPs found within IFIH1

The positions of 11 amino acid variants are indicated by grey vertical lines. Note that there are two consecutive variants at residue 842 and 843 – so they are coloured together. Seven dbSNP identifiers are highlighted in yellow. Other representations and colour schemes are the same as in Figure 5-2.

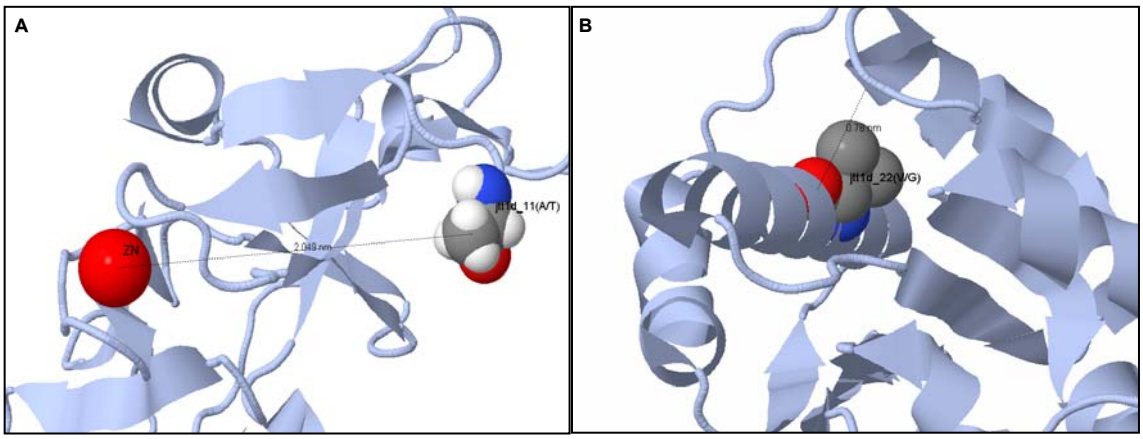


Figure 5-4 Three-dimensional structure of IFIH1 highlighting two wild-type amino acids of variant jtt1d_11 and jtt1d_22

A. Ala^{945} (jtt1d_11) and zinc ion are coloured in CPK and red, respectively, and both are represented in a space filing model. The main-chain backbone is illustrated as a cartoon. The three-dimensional structure is from PDB (2RQB), which crystallises the C-terminal region (residues from 896 to 1025) of IFIH1. The distance between the zinc ion and Ala^{945} is approximately 20 Å. **B**. Val^{340} (jtt1d_22) is coloured in CPK. The distance between Val^{340} and its nearby helical region is 7.8 Å. Other representations and colour schemes are the same as shown in **A**. Both figure **A** and **B** are drawn using Jmol [274] and accessible from http://samul.org/T1D.

5.2.3.2 Variant in Cytotoxic T-lymphocyte protein 4 (CTLA4)

There are seven genetic variants[41] in DNA regions coding for cytotoxic T-lymphocyte protein 4 (CTLA4), of which variant T17A (jtt1d_36) is annotated as "increased risk for Graves disease, insulin-dependent diabetes mellitus, thyroid-associated orbitopathy, systemic lupus erythematosus and susceptibility to hepatitis B virus infection [275,276,277,278,279]" by UniProt with an equivalent dbSNP identifier rs231775. Thr^{17} is located within the N-terminal region of a cytotoxic T-lymphocyte protein 4 where a potential signal sequence is located (see Figure 5-5). Therefore, the amino acid variant might interrupt the signal that directs where the native protein should be transported. However substitution scores are non-negative; 0 by BLOSUM and 1 by PAM.

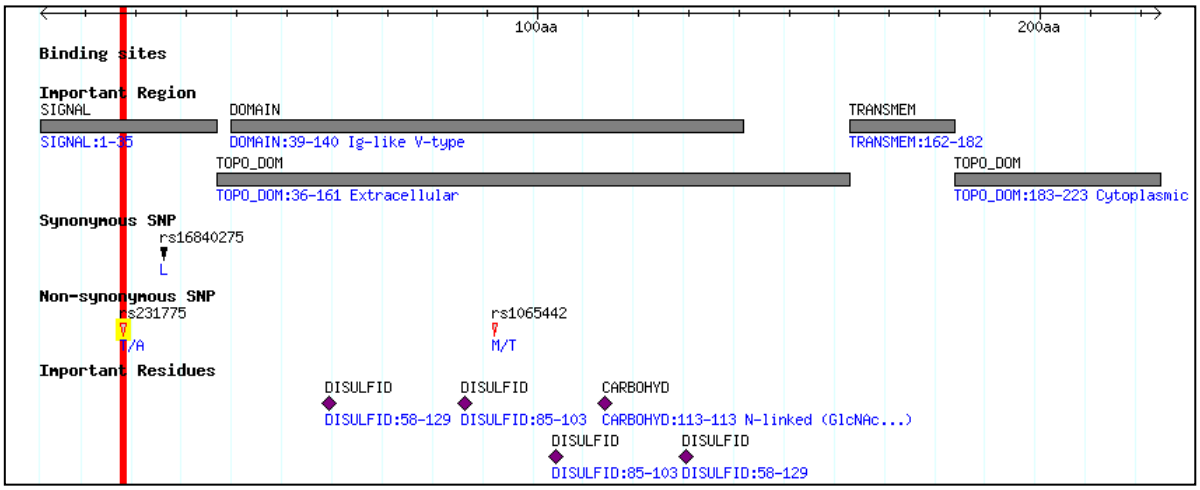


Figure 5-5 A schematic diagram highlighting the position of jtt1d_36 within CTLA4

The position of jtt1d_35 (Thr^{17}) is indicated with a red vertical line with its dbSNP equivalent (rs231775) highlighted in yellow at the same position. Other representations and colour schemes are same as shown in Figure 5-2.

[41] http://samul.org/T1D/353snps/gene/CTLA4

5.2.3.3 A variant within zinc-finger CCCH domain-containing protein 10 (ZC3HA)

Two genetic variants are found within the genetic region of ZC3H10[42] which encodes a zinc-finger CCCH domain-containing protein 10 (ZC3HA); one (jtt1d_186) in the 3' UTR and the other nsSNP—jtt1d_185 (E135Q)—within the protein coding region indicated as a vertical line in Figure 5-6A. There are three zinc-finger domains, of which the nsSNP occurs in the third domain. The three-dimensional structure of ZC3HA was not available in the PDB, but a close homologue (33% sequence identity)—RNA-binding domain in the human muscleblind-like protein 2 (PDB: 2E5S)—was found and the equivalent position was investigated (see Figure 5-6B). The homologue contains the last two zinc-finger CCCH domains out of four in total. The equivalent position (Asn^{54}) of jtt1d_185 is located in the loop region between two zinc finger domains. Asn^{54} does not seem to take part in the zinc-binding motif directly, but appears to act as a scaffolding residue by making a close contact (5.6Å) with Cys^{23} which is one of CCCH motif as shown Figure 5-6B (see [280,281] for review papers on zinc-biding sites). Considering a common qualitative feature of a metal-binding sites [282], the variant appears to be a hindrance to the zinc-finger binding motif. In addition, the secondary structure of Asn^{54} corresponds to a positive φ main-chain torsion angle, which is stabilized by establishing an interaction between a side-chain carbonyl (CO) and a main-chain carbonyl (CO) (see section 3.2.3) [235]. Interestingly, Glu^{135}, which is the wild-type amino acid of jtt1d_185, also retains a carbonyl group in its side chain and is more frequently observed in a positive φ torsion angle class than Gln, the mutated residue (see Table 3-1). The substitution score, both from PAM and BLOSUM, from Glu to Gln is 2, suggesting it would not be so much deleterious. However deprivation of an acidic carboxyl group could possibly affect the stability of zinc finger motif. There are no reported amino acid variants associated with this protein from dbSNP.

[42] http://samul.org/T1D/353snps/gene/ZC3H10

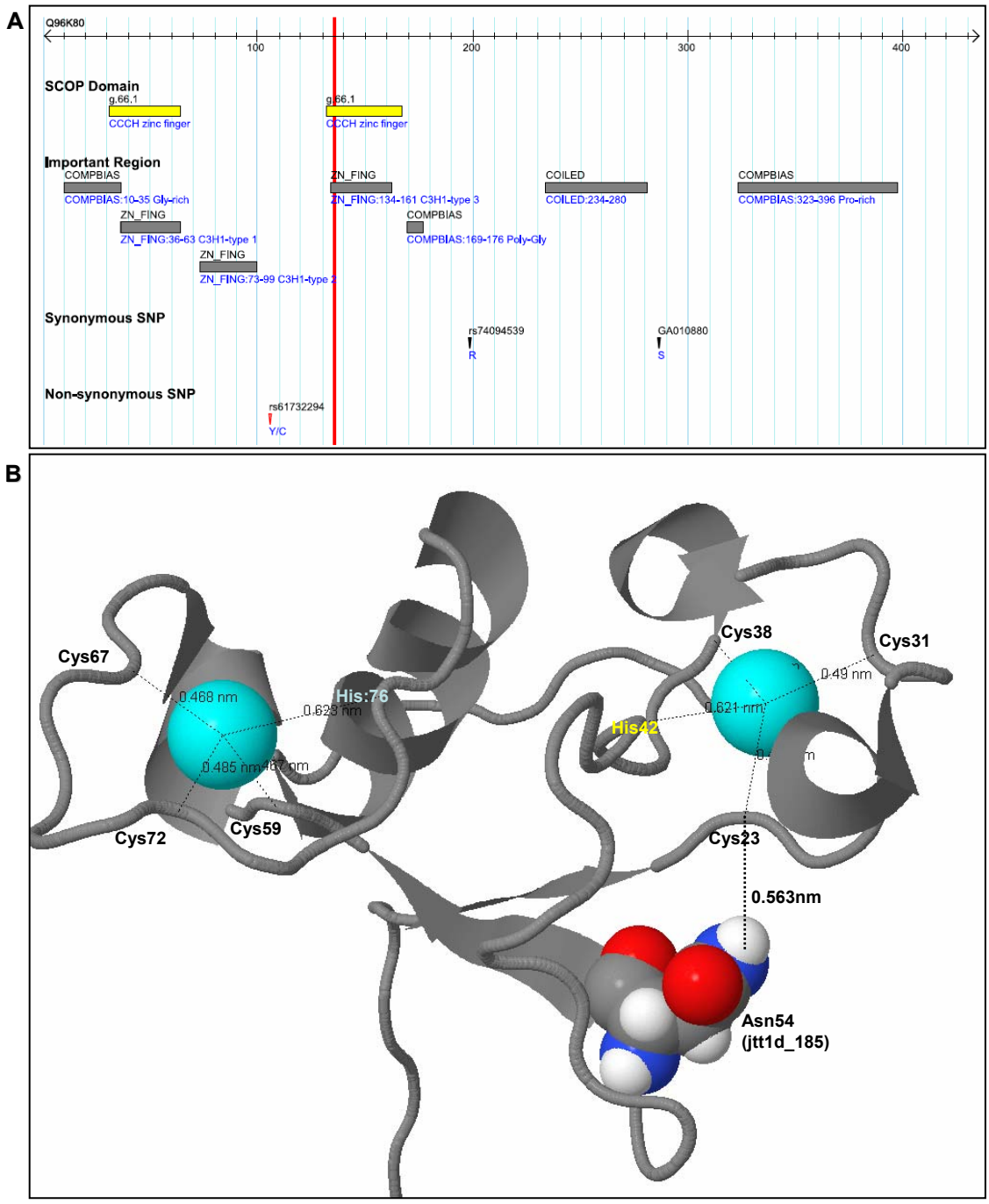


Figure 5-6 A schematic diagram highlighting the position of jtt1d_185 and its equivalent position within a homologue

A. A schematic diagram of ZC3HA showing UniProt annotations with the location of jtt1d_185. The position of jtt1d_185 (E135Q) is indicated with a red vertical line. Two structural domains, assigned by the SCOP database, are indicated in yellow boxes. Other representations and colour schemes are the same as shown in Figure 5-2. **B**. A solution structure of the two zinc finger domains (CCCH) of muscleblind-like protein 2, which is a structural homologue of ZC3HA. Asn^{54}, the equivalent position of jtt1d_185, is

represented in a space filling model and coloured in CPK. Two zinc ions are coloured in cyan with their binding motif (CCCH) annotated.

5.2.3.4 Three variants within sulphite oxidase (SUOX)

There are 10 variants[43] in the genetic region coding for mitochondrial sulphite oxidase (SUOX), of which three—jtt1d_155 (P212S), jtt1d_156 (Y392S), and jtt1d_158 (G453D)—are non-synonymous SNPs. SUOX catalyzes the conversion of sulphite (SO3) to sulphate (SO4), the terminal step in the oxidative degradation of cysteine and methionine. There are three SCOP domains within this enzyme, of which the molybdenum (Mo) pterin domain, a ligand-binding domain, contains two amino acid variants (P212S and Y392S) and the E set domain, which belongs to Ig-like fold families, contains the last (G453D) (see Figure 5-7A). Deficiency of this enzyme in humans leads to a Mendelian disease known as isolated sulphite oxidase deficiency (ISOD), characterized by neurological abnormalities including multicystic leukoencephalopathy with brain atrophy [283,284,285]. 11 amino acid variants are known to be associated with the disease (see 'SwissProt Variants' track of Figure 5-7A), but none of them overlaps with the location of the three novel nsSNPs. There is the three-dimensional structure (PDB: 1MJ4) of this enzyme, but the crystal resolves only one domain (cytochrome b5-like heme-biding domain; residue 79 to 160) which does not contain any novel nsSNPs. Its chicken homologue (SUOX_CHICK), however, has the full-length protein crystallised and its structure solved in a dimeric state, so the equivalent positions of the three nsSNPs were investigated instead (see Figure 5-7B). First, I inspected the geometrical distances between the variants and their adjacent ligands (SO4 and Mo) to see whether the variants are close enough to impair ligand bindings physically, but it is unlikely based on the distance alone (>10Å). However, I found that Gly^{375} (jtt1d_158), which is in the E set domain, makes a close contact (<5.5Å) with Ser^{435} of the other protomer coloured green in Figure 5-7B. Therefore, this variant could disturb dimerization of this enzyme, which is active only in the dimeric state. Indeed, introduction of carboxyl side-chain (Asp) in place of wild-type side chain

[43] http://samul.org/T1D/353snps/gene/SUOX

(Gly) could clash with neighbouring residues (Ser). The substitution scores for all the three nsSNPs are negative (-1 for both BLOSUM and PAM), suggesting deleterious effects if they occur.

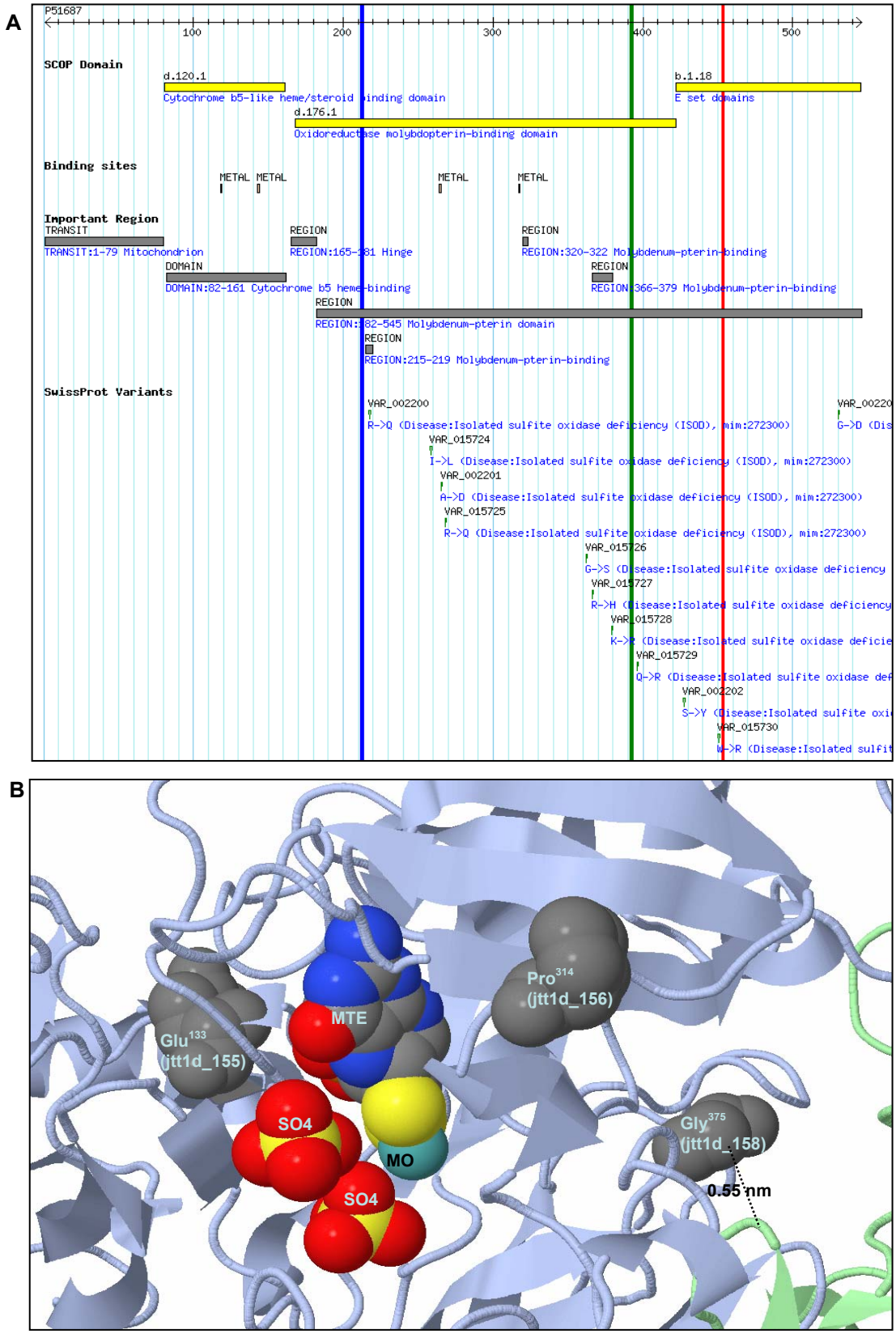


Figure 5-7 A schematic diagram highlighting the positions jtt1d_155, jtt1d_156 and jtt1d_158 and their equivalent positions within a chicken sulfate oxidase (a homologue of Human sulfate oxidase).

A. A schematic diagram of SUOX showing UniProt annotations and the locations of three nsSNPs (jtt1d_155, jtt1d_156 and jtt1d_158); they are indicated by blue, green and red vertical lines, respectively. In the "Binding Site" track, the first two metal-binding (METAL) residues (118 and 143) are responsible for interaction with an iron (part of heme) and the remaining two (264 and 317) for molybdenum (Mo). Other representations and colour schemes are the same as shown in Figure 5-2. **B**. The crystal structure of a chicken sulfate oxidase (a homologue of SUOX_HUMAN). The equivalent positions of three nsSNPs are coloured in grey with a space-filling model. The mainframe structure is represented in a cartoon and coloured by chain; chain A in blue and chain B in green. The two homologues share 64% sequence identity.

5.2.3.5 Two variants within a transmembrane region

Jtt1d_31 (R539G) and jtt1d_225 (L446P) occur in the transmembrane region of a potassium voltage-gated channel subfamily H member 7 (KCNH7) and a zinc transporter ZIP5 protein (S39A5), respectively (see Figure 5-8A and Figure 5-8B). The substitution scores, according to PAM, are -6 and -5 for jtt1d_31 and jtt1d_225, respectively, suggesting these substitutions would be very deleterious and very unlikely observed in nature. Indeed, jtt1d_31 replaces the large (174.2 g/mol), positively charged side-chain of Arg^{539}, with the small (75.07 g/mol) non-polar sidechain of Gly. Hence, the significant differences in size and conformation preferences would likely disturb the local structure. There are several structural homologues of KCNH7, but they do not contain the transmembrane domain region where the variant actually resides; this is a reflection of the fact that membrane proteins are under represented in the PDB due to difficulties in producing crystals.

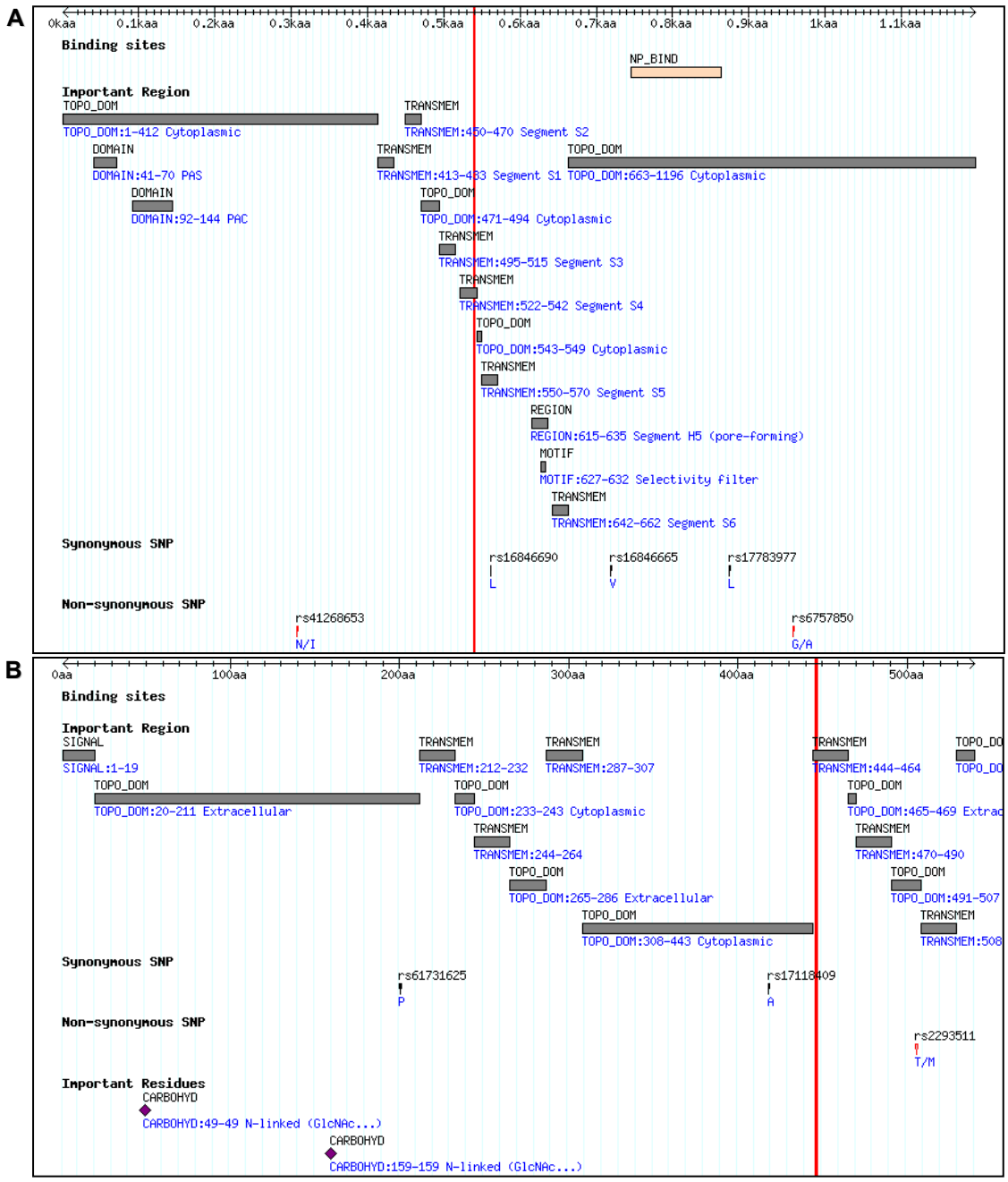


Figure 5-8 A schematic diagram highlighting the position of jtt1d_31 and jtt1d_225

A and **B** illustrate the positions of jtt1d_31 and jtt1d_225, indicated in red vertical lines, within the UniProt protein KCNH7 and S39A5 respectively. Other representations and colour schemes are the same as shown in Figure 5-2.

5.2.3.6 Variants in ErbB3 and its binding protein (PA2G4)

The receptor tyrosine protein kinase erbB-3 (ErbB3)—a member of the epidermal growth factor receptor (EGFR) family—contains 10 Asn-linked glycosylation sites, of which Asn^{414} corresponds to the wild-type amino acid of jtt1d_173 (N414H). Interestingly, Asn^{414}-linked-N-glycan in ErbB3 is known to play an essential role in regulating receptor hetero-dimerization with ErbB2 and also to have an effect on transforming activity [286]. In addition, it is reported that N414Q mutant of ErbB3 triggers auto-dimerization with ErbB2 without any ligand stimulation, which further accelerates phosphorylation of the receptor tyrosine. Eventually, the mutation promotes extracellular signal-regulated kinase (ERK) and Akt phosphorylation; sometimes overexpressed in a subset of human mammary tumors. Therefore, it is probable that the His variant at residue 414, induced by jtt1d_173, could also trigger spontaneous dimerization of the protein and further promote the signal transduction process and tumor development, but further molecular experiment is required to confirm this. Figure 5-9A illustrates the three-dimensional structure of ErbB3 (PDB: 1M6B), highlighting the wild-type amino acid (Asn^{395}) of jtt1d_173 interacting with a sugar molecule. In a distant homologue of the protein—type 1 insulin-like growth factor receptor extracellular domain (PDB: 1IGR)—the equivalent position (Asn^{72} of chain A) is responsible for interaction with a sulphate ion (SO4), but it seems that this may be an artefact promoting crystallization of a protein rather than physiologically relevant (see Figure 5-9B). I also investigated the three-dimensional structures of EGF receptor extracellular domains, homologues of ErbB3, (PDB: 1MOX, 1YY9 and 1NQL), but the equivalent position does not seem to inhibit EGF binding directly (see Figure 5-9C).

The ErbB3-binding protein 1, also known as a proliferation-associated protein 2G4 (PA2G4), interacts with ErbB3 (see above) and plays an important role in an ErbB3-regulated signal transduction pathway. Glu^{168} is the wild-type residue where the variant jtt1d_183 (E168G) is located. It is not clear whether Glu^{168} takes part in interactions with ErbB3, but if it does, the variant could possibly inhibit signal transduction. Indeed, Glu^{168} is located at the surface region based on the three-dimensional structure of PA2GA (PDB: 2Q8K), (see Figure 5-9D). Also, considering the physicochemical

properties of Glu, which is polar and negatively charged, it is likely that Glu^{168} is responsible for interaction, but further molecular studies are required to verify this. The amino acid substitution score from Glu to Gly is -2 according to both BLOSUM and PAM, and even lower (-4) based on ESST under the local structural environment of Glu; solvent accessible helical region without hydrogen-bond from side-chain.

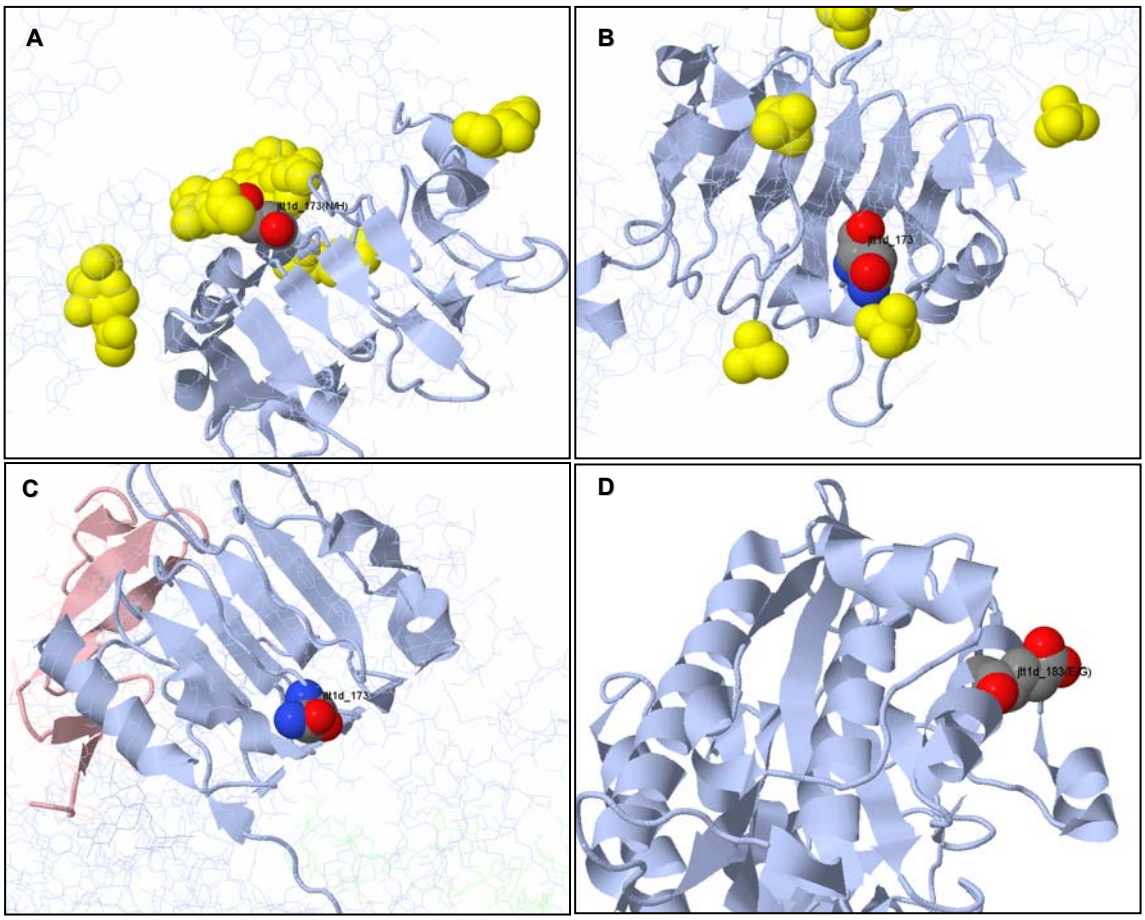


Figure 5-9 Three-dimensional structure of ErbB-3 and its binding protein

A. Three-dimensional structure of ErbB3 (PDB: 1M6B). Asn^{395} of chain A (jtt1d_173) and its interacting sugar molecule NAG (N-acetyl-D-glucosamine) are represented as space-filling models and coloured in CPK and yellow, respectively. Residues from 311 to 479, L domain (SCOP: d1m6ba2) are represented as a cartoon and wireframe elsewhere. **B**. Three-dimensional structure of a type 1 insulin-like growth factor receptor extracellular domain (PDB: 1IGR), a homologue of ErbB3. Asn^{72} of chain A, the equivalent position of jtt1d_173, and its interacting sulphate ligand are represented as space-filling models and coloured in CPK and yellow, respectively. Residues from 1 to 149, L domain (SCOP: d1igra1), are represented in a cartoon, and wireframe elsewhere. **C**. Three-dimensional structure of an EGF receptor

extracellular domain (PDB: 1MOX), a homologue of ErbB3. Asn^{79} of chain A, the equivalent position of jtt1d_173, is represented as a space filling model and coloured in CPK. Two SCOP domains, Epidermal Growth Factor (EGF) receptor (residue 1 to 162 of chain A) and EGF (residues from 2 to 50 of chain C), are represented as a cartoon and coloured in pale blue and pink, respectively. **D**. Three-dimensional structure of a proliferation-associated protein 2G4 (PDB: 2Q8K), the ErbB3-binding protein. Glu^{168}, the wild-type amino acid of jtt1d_183, is represented as a space filling model and coloured in CPK. Chain A is coloured in pale blue and represented as a cartoon.

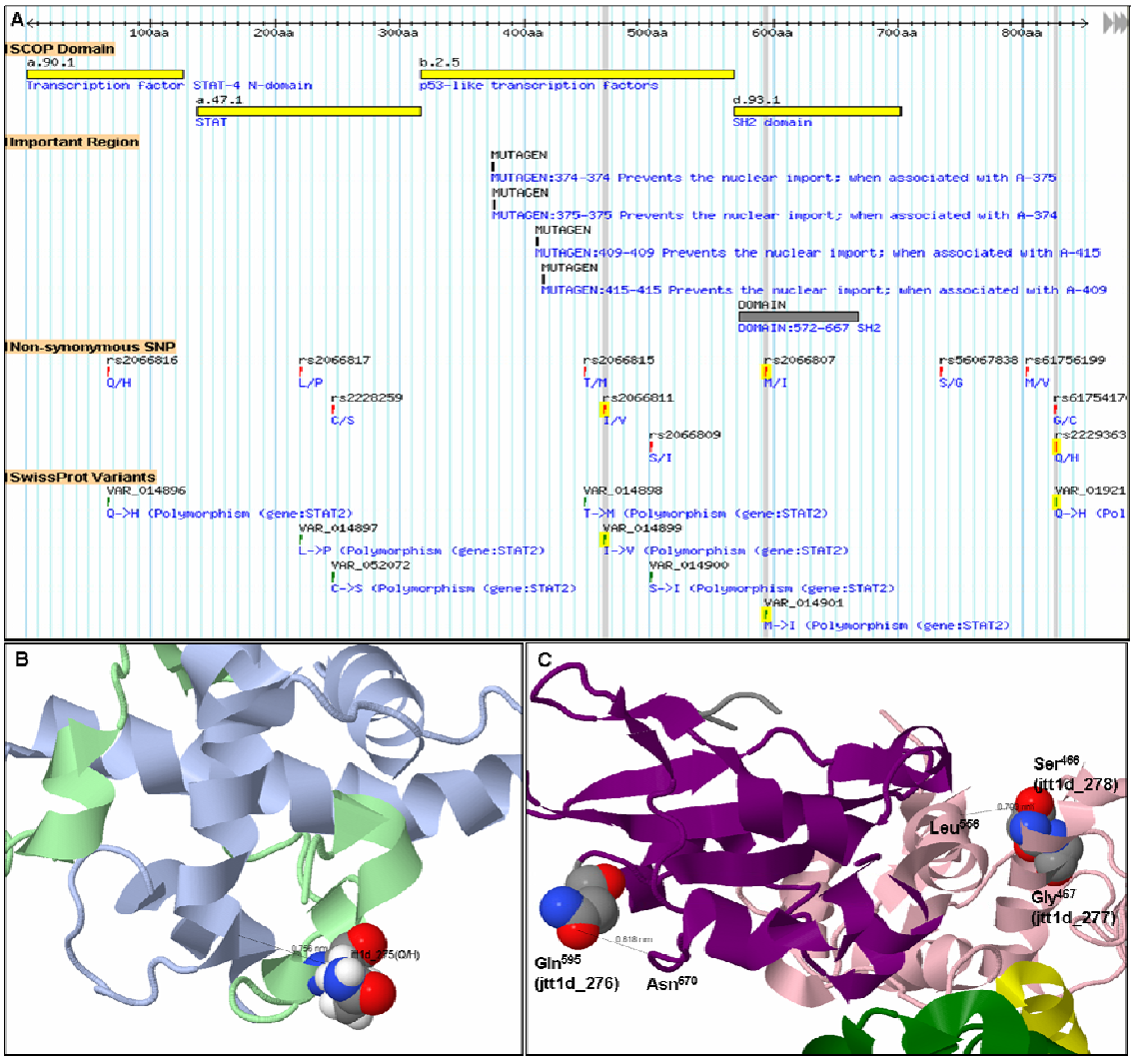
5.2.3.7 Variants in signal transducer activator of transcription 2 (STAT2)

10 genetic variants are found in DNA regions coding the signal transducer activator of transcription 2 (STAT2), of which four are non-synonymous SNPs: jtt1d_275 (Q826H), jtt1d_276 (M594I), jtt1d_277 (A465S), and jtt1d_278 (I464V). As shown in Figure 5-10A, there are 11 known amino acid variants of STAT2, of which three are at the positions where variant jtt1d_275, jtt1d_226 and jtt1d_278 are located; hence, only A465S is novel. STAT2 mediates signalling from type I interferons which trigger dimerization of phosphorylated of STAT1 and STAT2 via Jak kinases [287]. The phosphorylated STATs dimerise and interact with other molecules to form a transcription factor complex (ISGF3), which enters the nucleus. Four SCOP domains are assigned to residues 1 to 701, of which the p53-like transcription factor domain contains two variants (A465S and I464V) and the SH2 domain contains one (M594I).

There is a three-dimensional structure of STAT2 (PDB: 2KA4), but it contains only residues 783 to 838 (transactivation domain of STAT2) forming a complex with the TAZ1 domain of a CREB[44]-binding protein [288]. Based on the structure, Gln^{826}—wild-type amino acid of jtt1d_275—is responsible for interaction with a CREB-binding protein (chain A); C^{α}-distance is less than 7.5Å. Therefore the variant may interrupt the interaction (see Figure 5-10B). This could further inhibit dimerization of phosphorylated STAT2. The position of jtt1d_275 is same as that of rs222936345 of

[44] cAMP response element binding
[45] http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ss.cgi?subsnp_id=16361239

dbSNP [289] and VAR_01921346 of SwissVar [138]. In Figure 5-10C, equivalent positions of three remaining variants (jtt1d_276, jtt1d_277 and jtt1d_278) were investigated in the three-dimensional structure of STAT1 (PDB: 1YVL), a homologue of STAT2. Gln^{595}, the equivalent wild-type amino acid of jtt1d_276, is very close (<6.1Å) to Asn^{670} located in a loop region nearby; hence, the variant may incur local structural changes. Ser^{466} and Gly^{467}—equivalents of jtt1d_278 and jtt1d_277, respectively—are located within a helical region interfacing another helical segment in the p53-like transcription factor domain. In particular, Ser^{466} is fairly close (<7Å) to Leu^{556} of its nearby helical region, so it may interrupt helical packing. However, none of the substitution scores of these nsSNPs is negative.



[46] http://expasy.org/cgi-bin/variant_pages/get-sprot-variant.pl?VAR_019213

Figure 5-10 A schematic diagram and three-dimensional structure highlighting variants within STAT2 and its homologue

A. A schematic diagram of STAT2 illustrating the locations of four nsSNPs–jtt1d_275, jtt1d_276, jtt1d_277 and jtt1d_278–indicated by grey vertical lines. Their dbSNP and SwissVar equivalents are coloured in yellow boxes. Note that two consecutive variants (at residue 464 and 465) are coloured together. Other representations and colour schemes are the same as shown in Figure 5-2. Figure **B** shows the NMR structure of STAT2 (coloured in light green) and its interacting molecule CREB-binding protein (coloured in pale blue). The three-dimensional structure of STAT2 corresponds to the N-terminal region (residue 783 to838) shown in Figure **A**. Gln^{826} (jtt1d_275) is coloured in CPK and represented in a space filling model. The main-chain backbone is illustrated as a cartoon. The closest distance between jtt1d_173 and chain A is 7.5 Å. Figure **C** shows the crystal structure of STAT1, which is a homologue of STAT2 shown in Figure **A** (residue 1 to 678). Four SCOP domains are coloured in green, yellow, pink and purple in the same order as they appear in Figure **A**. The equivalent positions (Ser^{466}, Gly^{457} and Gln^{595}) of three nsSNPs are coloured in CPK with a space-filling model. The mainframe structure is represented as a cartoon. The two homologues share 44.3% sequence identity.

5.2.3.8 Variants in the myosin light chain (MLY6)

Two nsSNPs—jtt1d_195 (P112S) and jtt1d_197 (V145L) —are found within genetic regions coding myosin light chain 6 (MYL6). There are two UniProt proteins corresponding to the locus MYL6: myosin light polypeptide 6 (MYL6) and myosin light chain 6B (MYL6B), which share 81% sequence identity. Variant V145L—this is same as rs61938990 of dbSNP—is within the third EF-hand domain of MYL6, whereas P112S is between first and second EF-hand domains of MYL6B. The three-dimensional structure of MYL6 (PDB: 1BR1) reveals that Val^{145} is one of the residues interacting with chain B, a myosin heavy chain (see Figure 5-11A). In addition, Val^{145} is making a very close contact (<4.2Å) with its nearby helical segment, which constitutes an EF-hand motif. Therefore, replacement of Val with Leu could disturb native local structure by introducing a methyl group, even though the substitution scores are non-negative: 0 by PAM and 1 by BLOSUM and ESST. Figure 5-11B highlights the position of Pro^{112} (jtt1d_195), which is located in a coiled region linking two helices (PDB: 1OE9). Investigation of close homologues suggests that the equivalent position might disturb

interactions with the following ligands, presumably through a local conformational change: (i) calcium ion from troponin C (PDB: 1AVS, 1TNQ and 1YTZ; see Figure 5-11C), (ii) magnesium ion from troponin C (PDB: 1SBJ), and (iii) lead ion from calmodulin (PDB: 1N0Y Figure 5-11D). The substitution score from Pro to Ser is negative according to BLOSUM (-1) and ESST (-2).

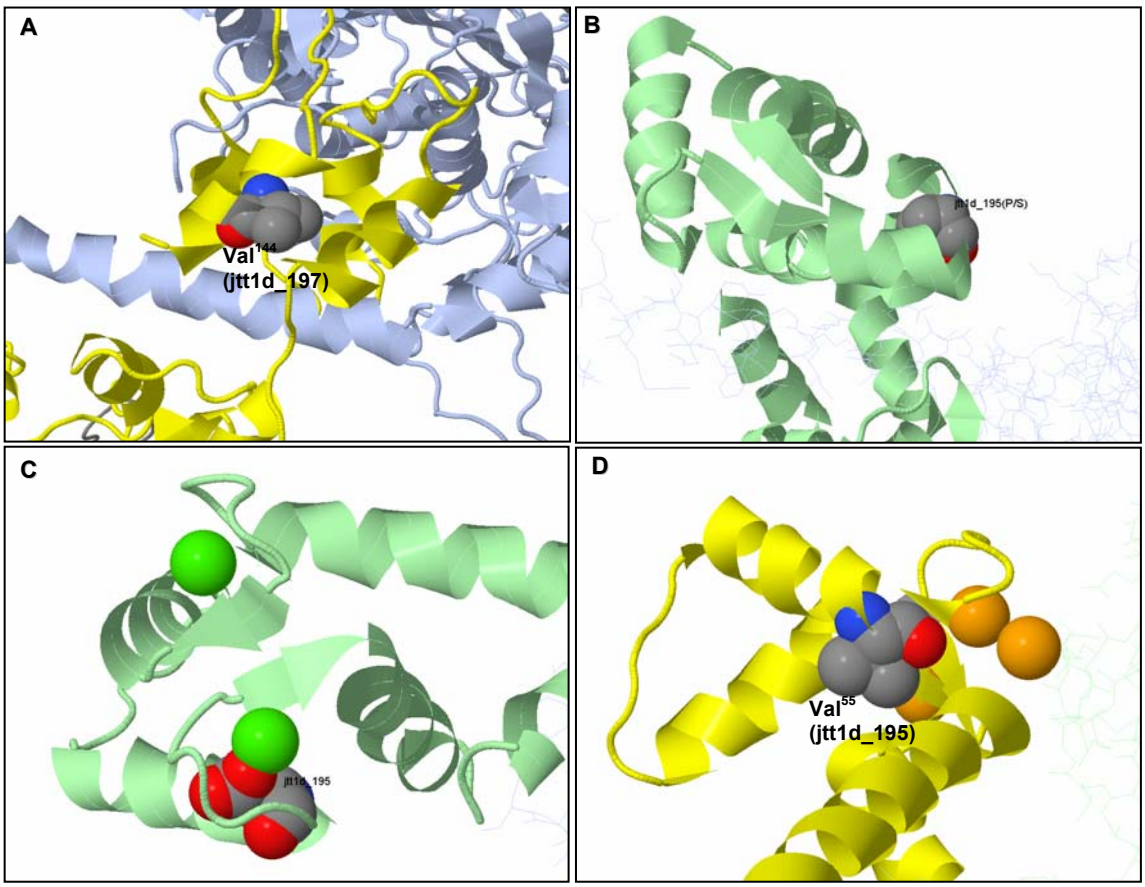


Figure 5-11 Three-dimensional structures highlighting the locations of two variants jtt1d_195 and jtt1d_197

A. A chicken homologue of MYL6 (coloured in yellow) is shown with a myosin heavy chain (MYH11) coloured in light blue. Val^{144} is the wild-type amino acid residue of jtt1d_197. Two homologues (MYL6 chicken and human) share 90% sequence identity. **B**. Three-dimensional structure of MYL6B (coloured in light green) and a myosin heavy chain (wireframe in light blue). The position of jtt1d_195 is represented as a space filling model and coloured in CPK. **C**. Three-dimensional structure of a calcium-saturated N-terminal domain of troponin, a homologue of MYL6B. Calcium ion is coloured in green and illustrated as a space filling model. **D**. Three-dimensional structure of a calmodulin protein, a homologue of MYL6B. Calcium and lead ion is coloured in green and orange in **C** and **D**, respectively, and illustrated as a space filling model.

5.2.3.9 Variants in ankyrin repeat domains

The serine/threonine-protein phosphatase 6 regulatory ankyrin repeat subunit C (ANR52) is a regulatory subunit of protein phospatase 6 that is involved in the recognition of phosphoprotein substrates. The protein is encoded by gene ANKRD52 onto which 22 genetic variants[47] were mapped; 10 are within the coding region, of which six are nsSNPs. There are 28 ankyrin (ANK) repeats, a 33-residue motif consisting of two alpha helices separated by loops, within the protein and the six nsSNPs are within the motifs except jtt1d_239 (S1061T), which is also known as rs59626664 (see Figure 5-12A). One more genetic variant (dbSNP: rs12305753) is already identified with this protein, which corresponds to jtt1d_241 (S499P); hence, only four nsSNPs are novel. Among them, substitution scores of variant C733W (jtt1d_240), located in ANK 21, and P492T (jtt1d_243), in ANK 15, are negative according to both BLOSUM and PAM. Figure 5-12B and Figure 5-12C show the three-dimensional structure of an ANK repeat motif (PDB: 1NOR), highlighting two equivalent positions of jtt1d_240 (His^{80}) and jtt1d_243 (Pro^{104}). His^{80} is within a solvent-accessible loop region linking two helices and makes hydrogen-bonds to amide and carbonyl groups of a main-chain in an adjacent helical region. Hence mutation of this residue could decrease local structural stability and further destabilize the ANK motif. Pro^{104}, however, is located in the solvent-inaccessible helical region of the motif without any hydrogen-bond from its side chain. Substitution with Thr would allow a hydrogen bond through the hydroxyl group of the sidechain; therefore it could incur local structural changes. In addition, considering the functional role of the ANK repeat (mediating protein-protein interactions), the two nsSNPs could be very deleterious.

[47] http://samul.org/T1D/353snps/gene/ANKRD52

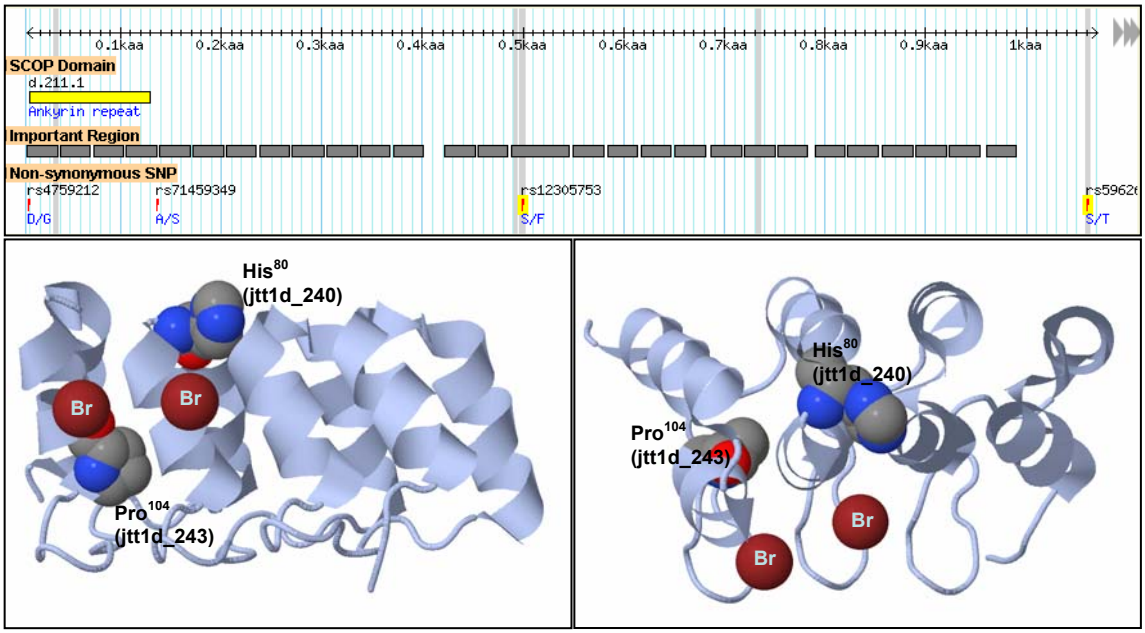


Figure 5-12 A schematic diagram highlighting the amino acid variants in ANK repeats and their equivalent positions within the three-dimensional structure

A. A schematic diagram of ANK52 illustrating the positions of six nsSNPs (indicated with grey vertical lines) within the protein: jtt1d_247 (N35K), jtt1d_243 (P492T), jtt1d_242 (A498P), jtt1d_241 (S499P), jtt1d_240 (C733W), and jtt1d_239 (S1061T). Two dbSNP equivalents (rs12305753 and rs59626664) are indicated in yellow. Note that two consecutive variants (residue 498 and 499) are coloured together. The ANK repeats are indicated as a grey box within the 'Important Region' track. Other representations and colour schemes are the same as shown in Figure 5-2. **B**. Three-dimensional structure of four ANK repeats (PDB: 1N0R). The structure corresponds to residues from 654 to 776 (or 375 to 497) shown in **A**. His^{80} and Pro^{104} are equivalent positions of the variant C733W and P492T, respectively. **C**. Same structure illustrated at different angle. Substitution scores of jtt1d_240 (C/W) are -2 and -11, according to BLOSUM and PAM, respectively. Substitution scores of jtt1d_243 (P/T) are -1 and -2, according to BLOSUM and PAM, respectively. The structure shares 43.3% sequence identity with the equivalent sequence region. (Br: Bromide ion)

5.2.4 Concluding Remarks

In this chapter, I have demonstrated a method for interrogating genetic variants responsible for disease aetiology using type 1 diabetes as an example. The main principle behind the approach explained in this chapter is simply applying lessons learnt from protein evolution to amino acid variants, in order to see whether they are acceptable or not. Therefore, I have mainly investigated structural and functional environments of amino acids variants; and interrogated them in terms of: i) their local structural environment to see whether native properties of wild-type amino acid have been impaired, and ii) the protein's functional niche to assess the impact of mutations. The claimed candidate variants underlying T1D aetiology still need further molecular studies to verify the significance of my approach. However one major gain of this approach is that it should be complementary to current genome-wide association studies by prioritizing genetic variants for further study. Considering one of the critiques of GWAS, which states that it does not provide any functional implication of genetic variation, the reductionist approach described here could be advantageous and indeed applicable to molecular diagnosis, especially if there is concensus between the two methods.

As shown in Table 5-2, amongst 353 SNPs, 192 (54.4%) are located within protein coding regions of which only 17.7% (34/192) are mapped onto their exact locations in protein three-dimensional structures. Hence, almost half (1 - 192/353) – those located within intronic regions and regulatory regions – could not be considered in this study, and even the majority (1 - 34/192) of protein-coding SNPs could not be interrogated in terms of their local structural environments. Here, I want to bring several points to the fore from this statistic. Firstly, modelling three-dimensional structures could help increase the coverage of nsSNPs to be interrogated within a structural context. Even though I tried to make the best of structural information by using sequence homology with proteins of known three-dimensional structure, this is limited by the quality of alignments especially for low sequence identity regions. Secondly, I focused only on protein-coding variants that replace amino acid types. However, complex diseases are not always influenced by the coding SNPs. Indeed more evidence is emerging for the

role of intronic SNPs that control splicing and expression (and timing of expression) of DNA and RNA products [290,291] and even synonymous SNPs are reported to control mRNA stability and for correct splicing [121]. In addition, I had to exclude many genetic variants responsible for insertions and deletions of DNA bases and larger copy number variants because they are more difficult to study with what we learnt from protein evolution. Lastly, the frequencies of 353 SNPs from the 80 samples (a mixture of cases and controls) were not accessible to me at this stage of analysis. Hence they need to be further analised to establish a causal relationship between genetic variations and disease phenotypes.

5.3 Materials and Methods

5.3.1 Locating SNPs in Genome

The SNP data from the work of John Todd's group has been considered on the basis of the Genome Reference Consortium[48] version 37 (GRCh37). The locations of 353 SNPs, within the Ensembl genebuild (database version: 57.37b), were identified by using Ensembl API [245] and transferred onto corresponding Ensembl human genes (ENSG), transcripts (ENST), and proteins (ENSP). If a coding sequence of a transcript does not start with a legacy translation initiation codon (AUG), no further mapping process could be proceeded, so an error flag has been raised as shown in Table 5-2.

5.3.2 Mapping Ensembl proteins onto three dimensional structures

Ensembl protein sequences were aligned with their corresponding UniProt sequences using BL2SEQ software, an implementation of the Smith-Waterman algorithm [292], of the NCBI Blast software package [64]. The aligned Ensembl-UniProt sequence was further mapped onto three-dimensional structures using Double-map method [193] explained in 2.3.2 and 4.3.3

[48] http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/index.shtml

5.3.3 Characterization of functional and structural environments

Table 5-4 shows the list of UniProt annotations used to characterize the functional features of amino acid residues where the SNPs are located. UniProt Knowledgebase XML files are downloaded from the FTP site of UniProt[49] [50] and their functional features are parsed using the Perl XML::Twig[51]. To identify the local structural environment of amino acid residue, JOY has been used [60]. The criteria applied to determine the local environment are explained in 4.3.4 in details.

As described in Chapter 1 the local structural environments of amino acid residues where SNPs occur are characterized on the basis of definitions suggested by Overington and colleagues [88,89]: 1) main-chain conformation and secondary structure, 2) solvent accessibility and 3) hydrogen bonding between side chains and main chains. In this framework, there could be 64 distinct environments for a residue from the combination of structural features: four from secondary structures (α-helix: H , β-strand: E, coil: C and residue with positive φ main-chain torsion angle: P), two from solvent accessibility (accessible: A and inaccessible: a), and eight (2^3) from hydrogen bonds to main-chain carbonyl (C and c) or amide (N and n) or to another side chain (S and s). In addition, three functional interaction types are sought from our in-house data sources: 1) protein-protein interaction from PICCOLO database [41], 2) protein-ligand interaction from CREDO [293], and 3) protein-nucleic acid interaction from BIPA [202].

[49] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.xml.gz
[50] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_trembl.xml.gz
[51] http://xmltwig.com/

Table 5-4 Lists of UniProt functional features used

Annotations	Descriptions
REGION	Extent of a region of interest in the sequence
VAR_SEQ	Description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting
VARIANT	Authors report that sequence variants exist
HUMSAVAR	Human polymorphisms and disease mutations
TRANSMEM	Extent of a transmembrane region
NP_BIND	Extent of a nucleotide phosphate-binding region
MUTAGEN	Site which has been experimentally altered by mutagenesis
DISULFID	Cysteine residues participating in disulfide bonds
METAL	Binding site for a metal ion
DNA_BIND	Denotes the position and type of a DNA-binding domain
MODRES	Modified residues excluding lipids, glycans and protein crosslinks
BINDING	Binding site for any chemical group (co-enzyme, prosthetic group, etc.)
ZN_FING	Denotes the position(s) and type(s) of zinc fingers within the protein
ACT_SITE	Amino acid(s) directly involved in the activity of an enzyme
PEPTIDE	Extent of an active peptide in the mature protein
MOTIF	Short (up to 20 amino acids) sequence motif of biological interest
COMPBIAS	Region of compositional bias in the protein
CARBOHYD	Covalently attached glycan group(s)
CA_BIND	Denotes the position(s) of calcium binding region(s) within the protein
PROPEP	Part of a protein that is cleaved during maturation or activation
SITE	Any interesting single amino acid site on the sequence
SIGNAL	Sequence targeting proteins to the secretory pathway or periplasmic space
TRANSIT	Extent of a transit peptide for organelle targeting
CROSSLNK	Residues participating in covalent linkage(s) between proteins
NON_TER	The sequence is incomplete. Indicate that a residue is not the terminal residue of the complete protein
LIPID	Covalently attached lipid group(s)

5.3.4 Building a web front-end

Web front-end: The web front-end (http://www-cryst.bioc.cam.ac.uk/t1d) has been built on the basis of the Perl Catalyst web application framework[52] as this employs the Model-View-Controller pattern, which simplifies application development and maintenance.

Database back-end: The MySQL[53] is used as a main relational database management system (RDBMS) and the Perl DBIx::Class[54] for mapping relation data to data objects.

Web server: The Apache HTTP server[55] (version 2.2.4) and mod_perl are used to deploy SAMUL on the web.

[52] http://www.catalystframework.org/
[53] http://www.mysql.com/
[54] http://search.cpan.org/dist/DBIx-Class/
[55] http://httpd.apache.org/

Chapter 6

SAMUL: A Web-based Database System for Visualizing Structural and Functional Features of Proteins

So far, I described structural and functional environments that shape and affect the occurrence of amino acid substitution from the perspective of protein evolution. Also I addressed what determines amino acid replacements and to what extent those environments contribute distinctive substitution patterns. Finally, I characterized structural and functional restraints of amino acid variations in human proteins and exemplified how the understanding of structural and functional restraints can help interrogating genetic variations identified from a genome-wide association study of type 1 diabetes. In this chapter, I describe development of a web-based database system which compiles data sources that I have used in previous chapters. Some of the material in this chapter has been published in Molecular BioSystems[56] which I co-authored with.

[56] *Lee S, Brown A, Pitt WR, Perez Higueruelo A, Gong S, et al. (2009) Structural interactomics: informatics approaches to aid the interpretation of genetic variation and the development of novel therapeutics. Mol Biosyst.*

6.1 Introduction

To understand the complex nature of molecular interactions within and between cells, it is desirable to employ an approach that can encompass the various kinds of genomic and proteomic data. Indeed, several centralised databases, such as Ensembl [245] and GenBank [295] harness the deluge of genome sequence information and automate functional annotations of genes and proteins needed for structural interactomics. In addition, recent technical advancements in X-ray crystallography and NMR experiments have enabled massive production of protein structure information. The Protein Data Bank (PDB) is the main repository of 3D structures of biological protein macromolecules [214]. As of 22 June 2009, more than 58,000 structures had been deposited in the PDB. These structures are made up of 75,574 polypeptide chains, 6,862 nucleotide chains, 13 polysaccharide chains, and 81,735 ligands. Thus it is essential that databases can handle massive quantities of structural data for large-scale analyses of protein structures and their interactions. With this motivation, multiple databases concerning the structure and interactions of protein–protein, protein–nucleic acid, protein–small molecule, and protein–carbohydrate complexes have been developed to provide the basis for the various analyses (see [294] for a review).

Whilst various individual databases enable interaction type-specific structural and functional restraints to be investigated, the interactome is the sum of individual interactions. This dictates the need for integration between the disparate databases and other informatics resources. There is a need to annotate the system fully, in which protein sequence and protein structure information are integrated. Also, despite the considerable structural information available for proteins and protein interactions, gaps still persist, such as the under-representation of transmembrane proteins in the PDB. In order to understand a system fully it becomes necessary to fill the gaps, a process that can be partially achieved through comparative modelling [213].

In this context, the Blundell group recently developed GLORIA, which is a structural information-centric meta-database, as an outcome of integrating comprehensive structural annotations with the results of automated modelling and nsSNP analysis

[115,294]. Through the mapping between sequence and structures (double-map), which has been described in Chapter 2, all the databases for protein–protein interactions (PICCOLO [41]), protein–nucleic acid interaction (BIPA [202]), protein–ligand interaction (CREDO [293]), protein–protein inhibitors (TIMBAL [296]), protein structure alignment (TOCCATA, [41]) and nsSNPs on protein structure and comparative models are interconnected (see Figure 6-1). This comprehensive relational scheme can be further extended by integrating genome-scale modelling pipeline, so functional residues and their mutations can be extended through homology at large-scale. Figure 6-1 shows a schematic diagram and workflow of GLORIA comprising major databases, categorised by interaction type, alongside our in-house databases.

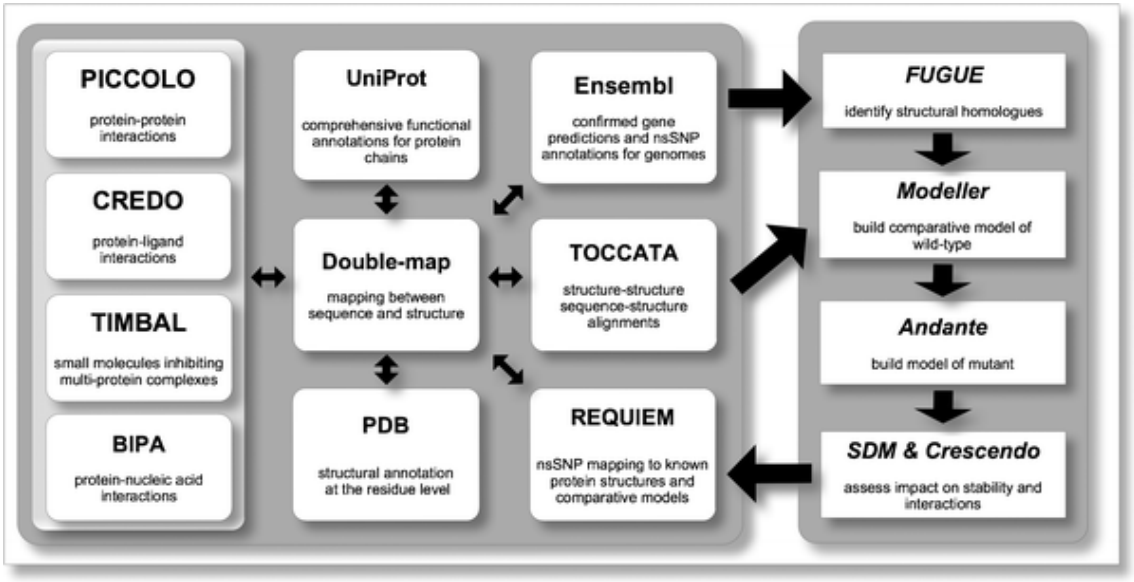


Figure 6-1 GLORIA and homology modelling-pipeline

GLORIA is a federation of interconnected databases integrating comprehensive biomolecular interactions and structural annotations with the results of the automated modelling at the genome-scale and analysis of impact of nsSNPs (this picture is taken from the reference [294] written by the Blundell group which I co-authored with).

In this chapter, I describe SAMUL[57] which is a web front-end of GLORIA. The main backbone of SAMUL is a sequence-to-structure mapping, as shown in Figure 6-1, which interconnects in-house databases and external data sources such as PDB, UniProt

[57] http://www-cryst.bioc.cam.ac.uk/samul (or http://samul.org/main, alternatively)

and Ensembl. SAMUL also provides structural and functional annotations of amino acid residues of proteins. The structural annotations are mainly from the local structural environments (by the scheme of 64 environments, described in section 1.2.2) of amino acid residues determined by JOY and presented and highlighted by Jmol − a molecular viewer [274]. For functional annotations, 26 UniProt feature descriptions are selected and the information is transferred onto their corresponding positions in 3D structures if available. In addition, SAMUL accommodates amino acid variations and mutations, which have been analyzed in Chapter 4, so that they can be browsed and interpreted in conjunction with the structural and functional environments of the wild type amino acid residues.

6.2 Results

6.2.1 Protein Sequence-to-Structure Mapping

Since the first identification of a protein sequence – that of insulin by Sanger and Tuppy in the 1950s [3,4], high-throughput sequencing techniques have enabled massive production of sequence information from different organisms. UniProt [216] is a central hub for protein sequences, providing rich annotation on function and cross-references. However, it does not explicitly provide any three-dimensional structure information of proteins at the amino acid residue level. Hence, in order to harness both UniProt and PDB information, sequences in UniProt have been mapped to their corresponding structures in the PDB [55,217,218,219,220,221,222].

In Chapter 2, I described a method, Double-map, to align a UniProt sequence to its corresponding PDB structure at residue level [193]. By using Double-map, UniProt annotations, especially feature (FT) records, can be harnessed and interpreted in the context of 3D structures of proteins. Further applications of Double-map are possible in combination with TOCCATA [41,115]. For example, the UniProt annotations can be extended across conserved positions within a TOCCATA alignment. In addition, nsSNPs that occur at protein coding regions can be mapped onto their corresponding amino acids in the context of their 3D structures if they are available in the PDB.

Alignment between chain G of 1cdl and P11799 (plain/text)

Index	PDB			UniProt		ENV[?]	Annotations
	SeqRes?	ResNum?	AtmRes?	ResNum?	Residue		
1	A	796	ala	1730	A	CAson	REGION, VAR_SEQ, PICCOLO
2	R	797	arg	1731	R	HAson	REGION, VAR_SEQ, PICCOLO
3	R	798	arg	1732	R	HAson	REGION, VAR_SEQ, PICCOLO
4	K	799	lys	1733	K	HAson	REGION, VAR_SEQ, PICCOLO
5	W	800	trp	1734	W	HAson	REGION, VAR_SEQ, PICCOLO
6	Q	801	gln	1735	Q	HAsON	REGION, VAR_SEQ, PICCOLO
7	K	802	lys	1736	K	HAson	REGION, VAR_SEQ, PICCOLO
8	T	803	thr	1737	T	HAsOn	REGION, VAR_SEQ, PICCOLO
9	G	804	gly	1738	G	HAson	REGION, VAR_SEQ, PICCOLO
10	H	805	his	1739	H	HASon	REGION, VAR_SEQ, PICCOLO
11	A	806	ala	1740	A	HAson	REGION, VAR_SEQ, PICCOLO
12	V	807	val	1741	V	HAson	REGION, VAR_SEQ, PICCOLO
13	R	808	arg	1742	R	HASon	REGION, VAR_SEQ, PICCOLO
14	A	809	ala	1743	A	HAson	REGION, VAR_SEQ, PICCOLO
15	I	810	ile	1744	I	HAson	REGION, VAR_SEQ, PICCOLO
16	G	811	gly	1745	G	HAson	REGION, VAR_SEQ, PICCOLO
17	R	812	arg	1746	R	HAson	REGION, VAR_SEQ, PICCOLO
18	L	813	leu	1747	L	CAson	REGION, VAR_SEQ, PICCOLO
19	S	814	ser	1748	S	CAsOn	REGION, MOD_RES, VAR_SEQ, PICCOLO
20	S	815	ser	1749	S	CAson	REGION, VAR_SEQ, PICCOLO

Figure 6-2 A screen shot[58] of SAMUL showing sequence-to-structure alignment between G chain of 1CDL and P11799

Two alignments (hence double-map) are shown here; 1) alignment between amino acid sequence defined in SEQRES record and that of ATOM record of a PDB file, 1CDL, 2) between amino acid sequence defined in SEQRES of the PDB file and the sequence from the corresponding UniProt entry (P11799). 'Index' is for the amino acid position of SEQRES and 'ResNum' is the residue number both in ATOM record of 1CDL and P11799. Amino acids, shown in 'AtmRes' column, are represented in JOY format (see Figure 1-1B). 'ENV' is for the local structural environment within the scheme of 64 (see Figure 1-1A). For the definitions of entries in Annotations column, see Table 6-1.

6.2.2 Rich Annotations

SAMUL provides 34 annotations at amino acid residue level from which 6 are for structural annotations of 3D structures and the rest 28 are for functional annotations mainly from UniProt features (FT) descriptions. Table 6-1 shows the full list of annotations available from SAMUL.

[58] http:// samul.org/main/pdb/1cdl/G/resmap

Table 6-1 Lists of structural and functional annotations provided from SAMUL (TLB for the in-house resource developed in the TLB group)

Source	Annotations	URL	Descriptions
TLB	PICCOLO	http://www-cryst.bioc.cam.ac.uk/piccolo	Protein-protein interaction database
	CREDO	http://www-cryst.bioc.cam.ac.uk/credo	A protein-ligand interaction database for drug discovery
	BIPA	http://www-cryst.bioc.cam.ac.uk/bipa	Biological Interaction database for Protein-nucleic Acid
UNIPROT	REGION	http://www.uniprot.org/manual/region	Extent of a region of interest in the sequence
	VAR_SEQ	http://www.uniprot.org/manual/var_seq	Description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting
	VARIANT	http://www.uniprot.org/manual/variant	Authors report that sequence variants exist
	HUMSAVAR	http://www.uniprot.org/docs/humsavar	Human polymorphisms and disease mutations
	TRANSMEM	http://www.uniprot.org/manual/transmem	Extent of a transmembrane region
	NP_BIND	http://www.uniprot.org/manual/np_bind	Extent of a nucleotide phosphate-binding region
	MUTAGEN	http://www.uniprot.org/manual/mutagen	Site which has been experimentally altered by mutagenesis
	DISULFID	http://www.uniprot.org/manual/disulfid	Cysteine residues participating in disulfide bonds
	METAL	http://www.uniprot.org/manual/metal	Binding site for a metal ion
	DNA_BIND	http://www.uniprot.org/manual/dna_bind	Denotes the position and type of a DNA-binding domain
	MODRES	http://www.uniprot.org/manual/mod_res	Modified residues excluding lipids, glycans and protein crosslinks
	BINDING	http://www.uniprot.org/manual/binding	Binding site for any chemical group (co-enzyme, prosthetic group, etc.)
	ZN_FING	http://www.uniprot.org/manual/zn_fing	Denotes the position(s) and type(s) of zinc fingers within the protein
	ACT_SITE	http://www.uniprot.org/manual/act_site	Amino acid(s) directly involved in the activity of an enzyme

	PEPTIDE	http://www.uniprot.org/manual/peptide	Extent of an active peptide in the mature protein
	MOTIF	http://www.uniprot.org/manual/motif	Short (up to 20 amino acids) sequence motif of biological interest
	COMPBIAS	http://www.uniprot.org/manual/compbias	Region of compositional bias in the protein
	CARBOHYD	http://www.uniprot.org/manual/carbohyd	Covalently attached glycan group(s)
	CA_BIND	http://www.uniprot.org/manual/ca_bind	Denotes the position(s) of calcium binding region(s) within the protein
	PROPEP	http://www.uniprot.org/manual/propep	Part of a protein that is cleaved during maturation or activation
	SITE	http://www.uniprot.org/manual/site	Any interesting single amino acid site on the sequence
	SIGNAL	http://www.uniprot.org/manual/signal	Sequence targeting proteins to the secretory pathway or periplasmic space
	TRANSIT	http://www.uniprot.org/manual/transit	Extent of a transit peptide for organelle targeting
	CROSSLNK	http://www.uniprot.org/manual/crosslnk	Residues participating in covalent linkage(s) between proteins
	NON_TER	http://www.uniprot.org/manual/non_ter	The sequence is incomplete. Indicate that a residue is not the terminal residue of the complete protein
	LIPID	http://www.uniprot.org/manual/lipid	Covalently attached lipid group(s)
CSA	CSA_PSI	http://www.ebi.ac.uk/thornton-srv/databases/CSA/	A database documenting enzyme active sites and catalytic residues in enzymes of 3D structure: homologous entries, found by PSI-BLAST alignment to one of the original entries
	CSA_LIT	http://www.ebi.ac.uk/thornton-srv/databases/CSA/	A database documenting enzyme active sites and catalytic residues in enzymes of 3D structure: original hand-annotated entries, derived from the primary literature
COSMIC	COSMIC	http://www.sanger.ac.uk/genetics/CGP/cosmic/	Catalogue Of Somatic Mutations In Cancer
ENSEMBL	ENVAR	http://www.ensembl.org/info/docs/variation/index.html	Ensembl Human variation database
PDB	MOD_RES	http://www.wwpdb.org/documentation/format32/sect3.html#MODRES	descriptions of modifications (e.g., chemical or post-translational) to protein and nucleic acid residues

6.2.3 Genetic Variation in Protein Structures and Disease

SAMUL houses amino acid sequence variants from *Homo sapiens* genome annotation provided by the following data sources; 1) Ensembl human variation database [245], 2) cancer somatic mutation from COSMIC [140], 3) UniProt human sequence variations [244]. They are integrated with various annotation information mentioned in the previous section. Table 6-2 shows the number of SNPs mapped onto UniProt, PDB, PICCOLO, CREDO, and BIPA at the time of writing. SNPs in Ensembl proteins were mapped onto their corresponding UniProt proteins and further to proteins in PDB *via* Double-map. SNPs in PICCOLO (4696), CREDO (3263), and BIPA (122) are subsets of SNPs in the PDB (18963). Among them, nsSNPs are of special interest especially if their allele types change corresponding amino acids which are presumably responsible for interactions in PICCOLO, CREDO and BIPA.

Table 6-2 Number of distinct SNPs categorized by annotations in SAMUL

Type	Database	NO of distinct SNPs
Sequence	Ensembl	203484
	UniProt	194053
Structure	PDB	18963
	PICCOLO	4696
	CREDO	3263
	TOPO_DOM	3068
	REGION	2412
	ZN_FING	183
	NP_BIND	140
	DNA_BIND	135
	BIPA	122
	PEPTIDE	115
	COSMIC	110
	DISULFID	100
	MOD_RES	92
	CSA_PSI	85
	CARBOHYD	81
	MUTAGEN	71
	SITE	63
	BINDING	62
	COMPBIAS	53
	MODRES	52
	TRANSMEM	47
	METAL	45
	PROPEP	42
	CA_BIND	37
	MOTIF	37
	ACT_SITE	23
	CROSSLNK	5
	CSA_LIT	4
	NON_TER	3
	TRANSIT	2
	SIGNAL	1

6.2.4 Visualization of Annotations

GBrowse: Structure and function annotations are graphically visualized and highlighted at the residue level of UniProt (or Ensembl) protein sequence using GBrowse (Generic Genome Browser) which is an open-source genome viewer widely used in the community [270]. Figure 6-3 shows a GBrowse generated image highlighting functional and structural annotations of a cell division protein kinase 2 (CDK2, UniProt accession: P24941). The image can be locally saved in various formats such as PNG, SVG and PDF through the web site. Annotations on the image are linked to the original sources of information so that users can investigate those features in depth.

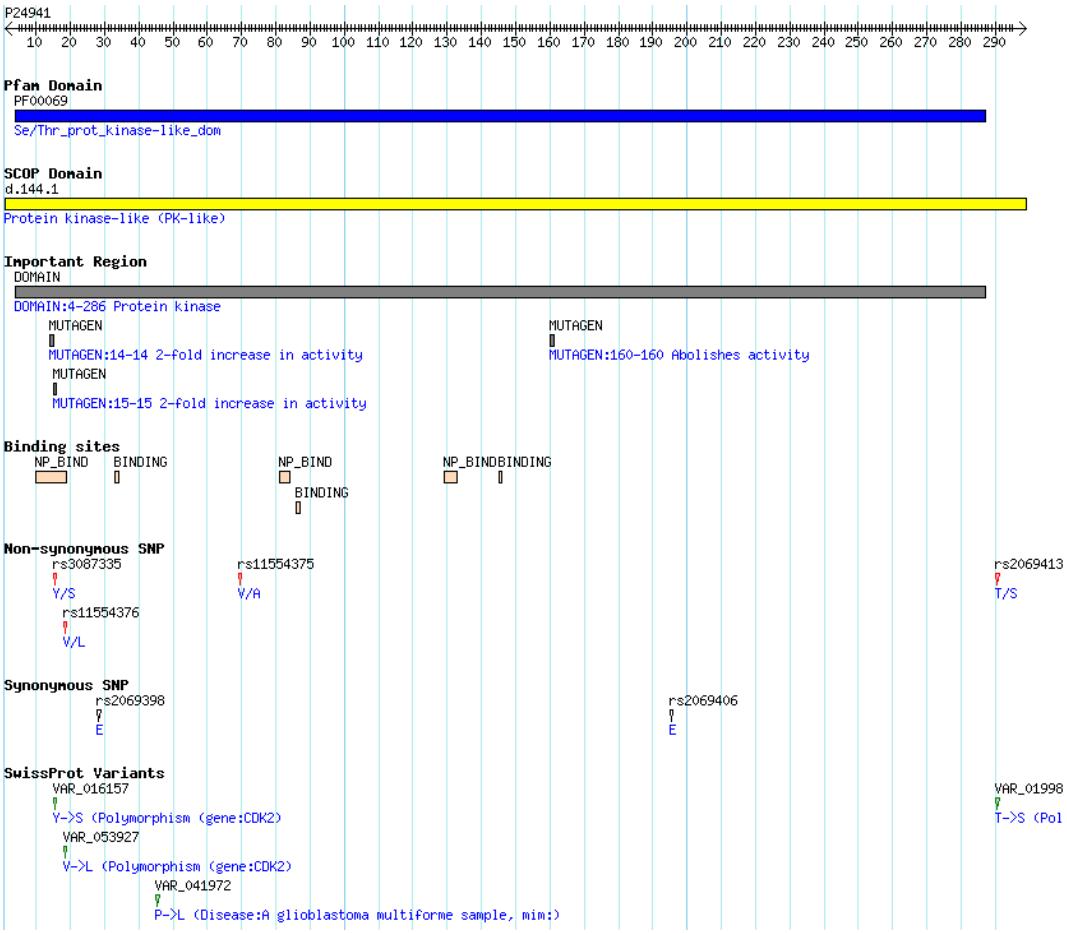


Figure 6-3 A screen shot[59] of GBrowse from SAMUL

Structural and functional annotations are provided by 9 tracks: 1) secondary structure, 2) Pfam and 3) SCOP for domain assignment information, 4) binding sites, 5) important regions, and 6) site for functional features, 7) synSNP, 8) nsSNP and 9) SwissVariants for amino acid variation information.

[59] http://samul.org/gb2/gbrowse/samul/?name=P24941

Jmol: Structural and functional annotations mapped onto 3D structure of PDB files could be selected and highlighted within the Jmol macromolecular view [274]. Figure 6-4 exemplifies a Jmol embedded SAMUL screen shot showing 3D structure of a cell division protein kinase 2 (CDK2, PDB code: 2VTI), featuring the location of various structural and functional features within the structure.

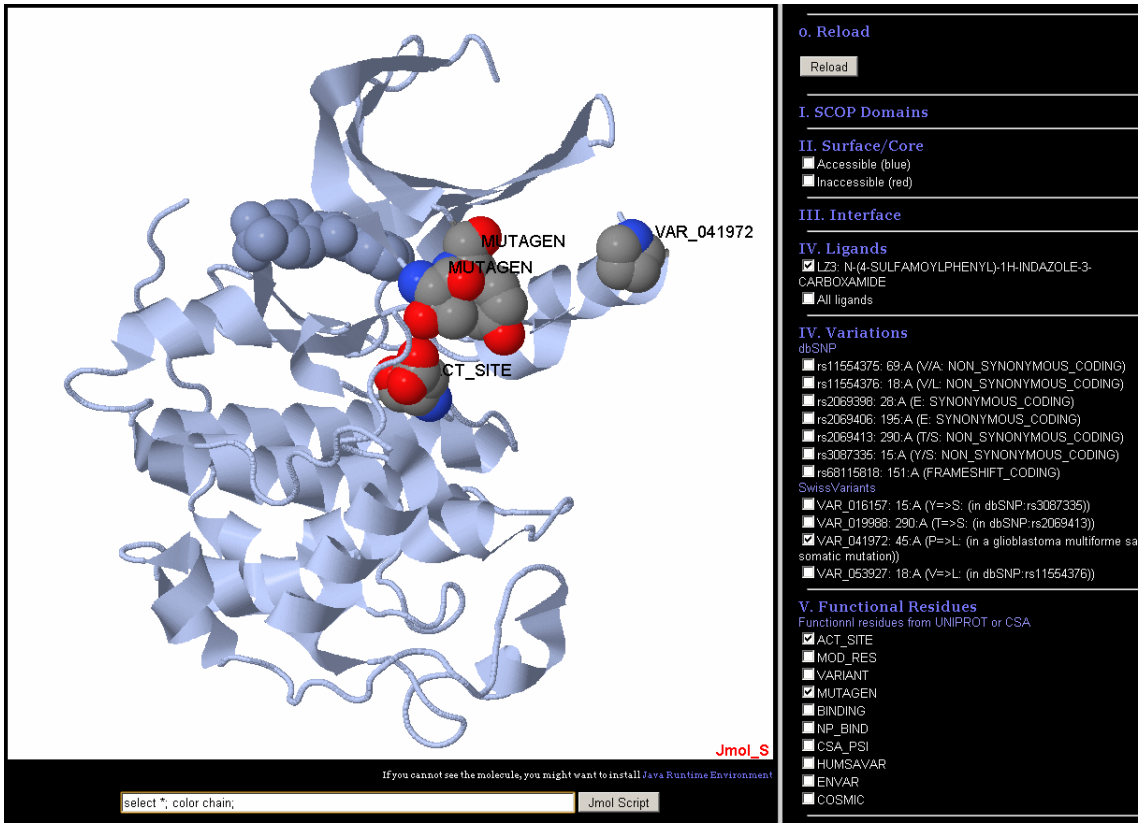


Figure 6-4 A screen shot[60] of Jmol from SAMUL

The navigation panel is on the right-hand side and the Jmol viewer is on the left. The pre-defined structural and functional annotations are presented as follows: 1) SCOP domain, 2) surface and core regions, 3) interface residues between two adjacent SCOP domain, 4) types of ligand, 5) amino acid variants, and 6) functional residues from the UniProt entry. There is also a form input field which accepts Jmol queries from advanced users who wish to manipulate visualisation options with their own flavours. The main chain of the protein molecule is presented as a cartoon with structural annotations in space-filled models of the individual amino acids.

[60] http://samul.org/main/pdb/2vti/jmol?hl=45:A&label=VARIANT&bgcolor=white

6.2.5 Distributed Annotation System (DAS)

SAMUL is a Distributed Annotation System (DAS) server, which provides XML-based web services to disseminate structural and functional annotations through the web. The DAS protocol is built on a client-server system which allows a single machine to communicate with a distant web server to gather different types of biological annotations, collate the information, and display it to the end user in a single view. Most of the major knowledge-based biological systems such as Ensembl, UCSC genome browser [297] and WormBase [298] provide DAS services. Numerous DAS resources are coordinated by the DAS registration server[61] [255]. Figure 6-5 shows an example of how the DAS service of SAMUL can be used in Jalview[62] which is a java-based multiple sequence alignment viewer and editor [252].

[61] http://www.dasregistry.org/
[62] http://www.jalview.org/

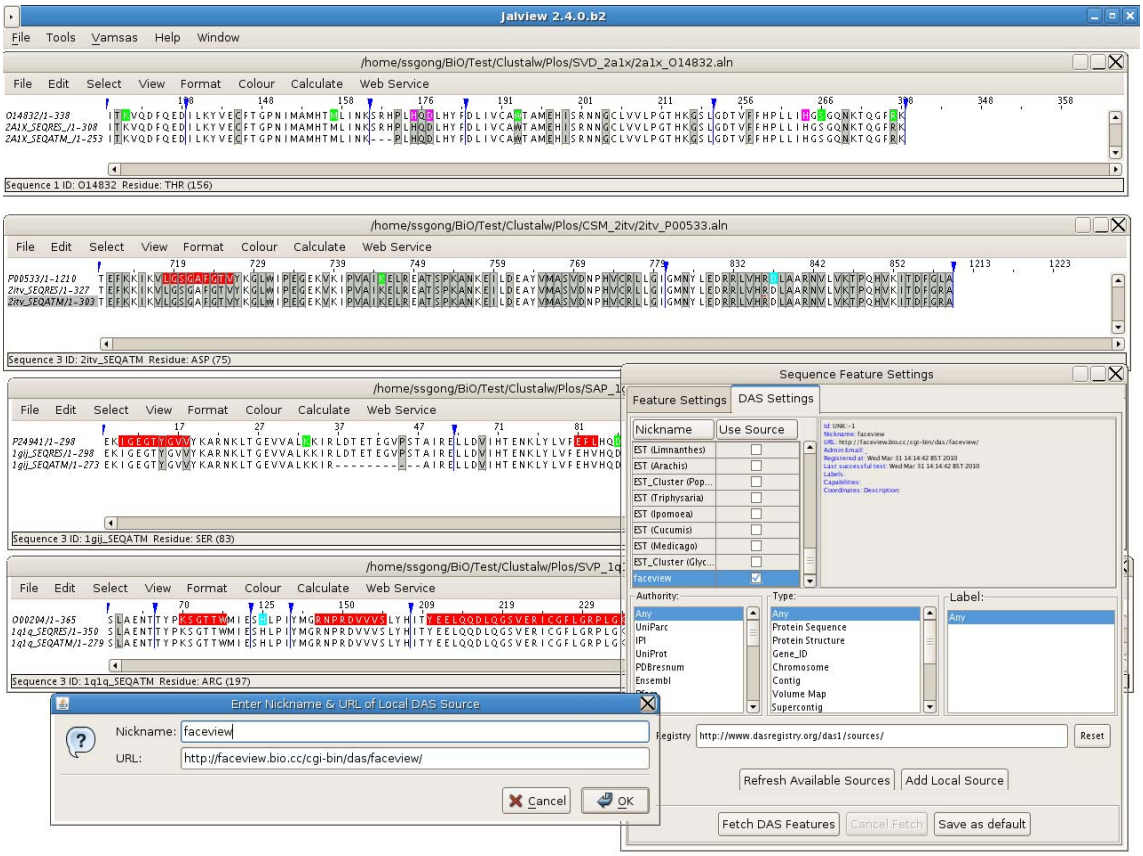


Figure 6-5 A screen dump showing the use of DAS service of SAMUL in Jalview

Four sequence alignment panels and two DAS configuration windows are shown. In the alignment panels, the following annotations ‘BINDING’, ‘ACT_SITE’, ‘METAL’ and ‘NP_BIND’ are coloured in green, cyan, magenta and red, respectively.

6.3 Materials and Methods

6.3.1 Data Source

Sequence-to-structure mapping: SAMUL employs the double-map method which aligns a sequence of UniProt to its corresponding PDB structure at residue level. See section 2.3.2 in details.

Sequence variations: The Ensembl human variation database is a major source of genetic variations. Also, COSMIC and UniProt are used as a source of cancer mutation data and disease-related amino acid variations, respectively. See section 4.3.1 for details.

Annotations: UniProt Knowledgebase XML files are downloaded from the FTP site of UniProt[63] [64] and their functional features are parsed using the Perl XML::Twig[65]. The TBL group's in-house databases − BIPA, CREDO, and PICCOLO − are used for the source of inter-molecular interaction types and CSA and PDB as the source of catalytic residue and modified residue information, respectively. For structural annotations Table 6-1 shows the full lists of annotations used in SAMUL.

6.3.2 Software

Calculation of local structural environments: JOY was used to identify the local structural environments of amino acids [60]. See section 3.3.2 for details.

Web front-end: SAMUL has been developed on the basis of the Perl Catalyst web application framework[66] as this employs the Model-View-Controller pattern, which simplifies application development and maintenance. The Jmol macromolecular viewer[67] is a default visualisation tool for a PDB file highlighting structural and functional features within the molecules. GBrowse[68] (version 2.0) is installed as a generic protein browser and a DAS server. SAMUL employs modern web 2.0 technology such as Google Ajax API, jQuery Javascript library and plugins such as the boxy, the jQuery tools and the coda-slider.

[63] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.xml.gz
[64] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_trembl.xml.gz
[65] http://xmltwig.com/
[66] http://www.catalystframework.org/
[67] http://www.jmol.org/
[68] http://gmod.org/wiki/Gbrowse

Database back-end: The MySQL[69] is used as a relational database management system (RDBMS) and the Perl DBIx::Class[70] for mapping relation data to data objects.

Web server: The Apache HTTP server[71] (version 2.2.4) and mod_perl are used to deploy SAMUL on the web.

[69] http://www.mysql.com/
[70] http://search.cpan.org/dist/DBIx-Class/
[71] http://httpd.apache.org/

Chapter 7
Concluding Remarks

In this thesis, I attempted to unravel the nature of amino acid replacements during protein evolution and tried to apply the principles to the understanding of the genetic variations or somatic mutations responsible for disease susceptibilities. However, I am deeply aware that assumptions underpinning this study are limited and reflect only some of the aspects out of many possible perspectives underlying how we understand protein evolution, genetic variations or mutations, genotype-phenotype causality and disease aetiology. Here, I enumerate limitations of methodologies in use and challenges raised during my study in the hope that this can give insights to those who wish to tackle the challenges in the future.

7.1 Restraints vs. constraints

When it comes to describing "evolution", the Blundell group has been using "restraints" for many years, although we accept that most evolutionary biologists use "constraints". However, protein structures are not "constrained" in evolution for the following reason. Firstly, based on the use of words written in a dictionary, the usual definition of "constrain" is "force, oblige, compel"; there is no option. The definition of "restrain" is to hold within bounds. In many areas of mathematical computation this is recognised; for example "restrained refinement" where we have target values and "constrained refinement" where the values are fixed. A "constraint" to us is a fixed aspect of a function, whereas a "restraint" is a target value which we seek to meet. Hence, in describing the evolution of proteins, we are mostly talking about "restraints" as we seek to retain secondary structure but know it changes even for orthologues. The packing of residues in the core is assumed to provide a local environment with which amino acid substitutions must be compatible, but once substitutions accumulate, so this becomes untrue. I therefore prefer the term "restraints", knowing that there will be variation in evolution. Secondly, I use "restraints" in the sense that these are structural, dynamic, systems or functional factors that influence the acceptance of amino acid substitutions that occur in divergent protein families. Given that selection occurs at the level of the organism and that individual proteins and the systems within which they evolve are plastic, these "constraints" tend not to "force", but rather to "restrain" the substitutions that occur in evolution.

7.2 Interaction types as functional restraints

In Chapter 2, I showed that discrimination of functional restraints from structural restraints could help describing the pattern of amino acid replacement and even enhance finding active site residues. However, considering the fact that all of the restraints to do with maintenance of tertiary structure are ultimately functional, it is not a trivial problem to make an explicit distinction between functional restraints and structural ones [299]. Indeed, many functions are mediated through quaternary interactions of proteins with other macromolecules in assemblies or with substrates, ligands or allosteric

regulators. The effects of these restraints are felt some distance away from the interaction site**,** but tend to have an increasing influence nearer to the site.

Integration of functional features, especially active sites of enzyme, into local structural environments is best exemplified by Chelliah *et al*. [100]. They measured the Euclidean distance between every amino acid and the known functional residues and compared the degree of conservation in terms of the proximity of functional residues. They showed that the degree of residue conservation is significantly higher in residues that are near to the active site compared with those that are far from it. Hence, geometrical distance from known active sites constitutes another restraint on amino acid substitutions in protein evolution and therefore can serve as an additional parameter to define the local structural environment in classifying amino acid substitution patterns (know as function-dependent ESST).

More recently, Richard Bickerton [41] considered the impact of protein-protein interactions on amino acid substitutions and made an interface-dependent ESST by taking four types of interacting accessibility interface residues: (i) interface core, (ii) interface periphery, (iii) core, and (iv) exposed. He showed that the strongest determinant is the interfacial accessibility environment followed by types of secondary structure. He also found that the interface environments are intermediate between the exposed surface and buried core; the interface core is more similar to the buried protein core and the interface periphery is more similar to the exposed surface.

Similarly, Semin Lee [294] considered residues involved in intermolecular interactions with nucleic acids and classified theses further into three types: (i) hydrogen bond; (ii) water-mediated hydrogen bond; and (iii) van der Waals contact. He found that residues interacting with nucleic acids have distinct substitution patterns when compared with the other sites and suggested the restraints of protein–nucleic acid interaction should also be considered.

The examples described above, which demonstrate restraints of amino acid substitution, also arise from interactions with other proteins; these are often components of

interaction networks, which are conserved throughout evolution [300] so that interacting proteins are under various restraints such as activity and life-time [301,302,303].

7.3 Toward integrated analysis of protein evolution

In Chapter 3, I focused on how amino acid substitutions, during divergent evolution of protein families, are constrained by the local structural environment of amino acid residues. I showed that strong restraints arise from the conservation of structure, not only from maintenance of a hydrophobic core and secondary structure, but also from buried, often charged hydrogen bonds. However, I have not attempted to discuss the origins of folds nor their evolution by additions and subtractions of elements of secondary structure, gene duplications and fusions; these have been widely reviewed elsewhere [304,305,306,307]. Neither did I consider restraints arising from the genomic position of the encoding genes, expression patterns, position in biological networks and robustness to translation [79]. (see section 1.1.4 for various non-structural restraints of protein evolution). Other factors can also be correlated with the rate of protein evolution. For example, expression level might be an important factor influencing evolutionary rate [72,308,309] as highly expressed proteins are constrained to have fewer mutations than relatively rare proteins to avoid the cost of misfolding effects. A proper understanding of the restraints on amino acid substitutions is an essential prerequisite to understanding protein evolution, but further insights will depend on integrated and multidisciplinary systems approaches [79,310].

7.4 Orthologues vs. paralogues

Chapter 3 defines each amino acid position, within a protein family or superfamily, in terms of its local structural environment and considers the impact of structural restraints on the amino acid substitutions that have been accepted during evolution. One major challenge here is to distinguish orthologues, which have the same functions in different organisms, and paralogues, in which gene duplication has occurred and new functions may have emerged [311]; in the latter case the restraints will have changed. Generally orthologs are defined on the basis of sequence similarity but this remains a source of uncertainty in comparative analyses [311]. Another argument that arises is the extent to

which the local environment is conserved in homologous families and therefore can provide restraints on amino acid substitutions. In other words, within a protein family or superfamily, how deep we should scan to extract important structural and functional features. For instance, if a set of sequences were only recently diverged they would not have accumulated enough substitutions to identify evolutionary restraint. Analyses of families and superfamilies of proteins show that the most critical packing arrangements of individual sidechains begin to differ when two proteins have less than 30% sequence identity due to relative movements of equivalent secondary structural elements [16,227], but some critical hydrogen-bonding interactions are retained at much greater levels of sequence divergence.

7.5 Obscure properties of cancer mutations

In Chapter 4, I showed that cancer somatic mutations and disease-related variants occur more frequently at amino acids making hydrogen bonds from side chains than neutral polymorphisms. In addition, based on substitution scores and amino acid property matrices, I showed that the severity of cancer somatic mutations lies between that of Mendelian disease-related variants and polymorphic variants; less deleterious than Mendelian disease causing variants but more severe than polymorphic variants. However, these properties of cancer mutations obscure the fact that cancers arise from mutations in a subset of genes that confer growth advantage to the tumour. Recently, Talavera *et al.* [246] investigated the pattern of cancer-related mutations and compared them with those from polymorphic variants. They showed that the distribution of cancerous amino acid substitutions is very similar to that of polymorphism, suggesting they are under similar selection pressures by neutral evolution, although polymorphic variants tend to occur at less conserved positions than cancer-related mutations. It is known that not all somatic mutations confer growth advantage to the cells. There are 'driver' somatic mutations which are the main contributors to the development of the cancers, whereas most somatic point mutations are likely to be 'passengers' that do not contribute to oncogenesis [158]. However, it is not a trivial problem discriminating between the two and our dataset almost certainly contains both types, obscuring the effect of 'driver' mutations. None the less, it is reported that driver mutations are more

clearly associated with key protein features than other somatic mutations (passengers) that have not been directly linked to tumour progression [312]. In addition, recent findings from the Stratton group could hopefully help identifying how the structural and functional properties of cancer mutations could contribute to cancer developments [120].

7.6 Other things to consider

At the time of this study, reported SNPs comprise 0.46% (0.13% for verified SNPs) of the total number of human DNA base pairs of which 53% of SNPs occur at intergenic regions and 36% occur at intronic region (See Table 7-1). Only 1.26% of human SNPs occur in protein coding regions of which more than half are non-synonymous SNPs (0.64%)—those that have been considered in Chapter 4—and the rest are synonymous SNPs (0.46%), frame shift (0.09%) and stop-gained mutations (0.02%). Throughout my analysis, I did not take the expression level into account; rather I assumed that proteins are expressed equally no matter whether they contain sequence variants or not. However, it is clear that proteins having deleterious mutations are selectively controlled by the protein degradation system to protect against misfolded or damaged proteins [77] and sometimes those mutations are compensated in other species [313].

Table 7-1 Total number of SNPs by different types of their consequences

Type	Occurrence	Ratio (%)
INTERGENIC	7,982,768	53.07
INTRONIC	5,481,863	36.45
UPSTREAM	663,985	4.41
DOWNSTREAM	556,742	3.70
3PRIME_UTR	137,639	0.92
NON_SYNONYMOUS_CODING	96,031	0.64
WITHIN_NON_CODING_GENE	86,955	0.58
SYNONYMOUS_CODING	69,035	0.46
5PRIME_UTR	28,343	0.19
FRAMESHIFT_CODING	14,002	0.09
REGULATORY_REGION,INTRONIC	13,365	0.09
SPLICE_SITE,INTRONIC	10,457	0.07
REGULATORY_REGION,UPSTREAM	4,951	0.03
REGULATORY_REGION,INTERGENIC	4,949	0.03
NON_SYNONYMOUS_CODING,SPLICE_SITE	2,845	0.02
STOP_GAINED	2,533	0.02

data from Ensemble human variations

Appendix I

Coordinates of 64 environments projected onto the principal component (PC) 1, 2 and 3

Structural Environment[1]	PC1	PC2	PC3
CASON	-11.39	2.94	-3.32
CASOn	-8.77	4.34	-2.31
CASoN	-11.32	2.41	-4.16
CASon	-10.54	3.84	-3.58
CAsON	-8.62	2.51	-2.60
CAsOn	-8.83	4.56	-1.26
CAsoN	-9.95	1.61	-4.27
CAson	-22.00	4.89	-6.02
CaSON	8.52	-3.65	-1.61
CaSOn	8.59	-3.68	0.06
CaSoN	5.44	-3.42	-1.57
CaSon	7.10	-2.77	-1.75
CasON	7.17	-3.41	-0.66
CasOn	8.49	-1.62	1.53
CasoN	4.48	-1.74	-2.20
Cason	13.77	-5.09	-0.09
EASON	-0.99	6.04	-7.85
EASOn	0.86	8.75	-7.83
EASoN	-2.87	4.78	-7.30
EASon	-1.31	10.34	-8.40
EAsON	0.80	5.43	-6.87
EAsOn	-0.07	7.20	-6.22
EAsoN	-3.01	5.35	-7.18
EAson	-9.10	18.97	-15.67
EaSON	17.56	-0.88	-6.35
EaSOn	16.33	0.37	-5.46
EaSoN	12.01	0.20	-5.58
EaSon	15.76	0.47	-4.34
EasON	14.01	-0.37	-4.89
EasOn	14.88	1.35	-4.04
EasoN	11.17	0.17	-3.52
Eason	27.48	-1.80	-10.81
HASON	-6.99	7.38	9.44
HASOn	-5.65	9.69	8.45
HASoN	-7.54	6.60	6.30

HASon	-7.79	10.49	6.77
HAsON	-5.30	6.62	7.24
HAsOn	-6.10	9.67	7.94
HAsoN	-6.30	7.12	5.69
HAson	-20.54	18.38	15.65
HaSON	11.31	-0.18	6.96
HaSOn	14.30	2.17	8.36
HaSoN	9.19	1.15	5.96
HaSon	12.79	4.74	6.82
HasON	10.46	0.13	6.37
HasOn	13.13	2.06	9.08
HasoN	8.90	1.34	5.50
Hason	22.85	1.85	17.37
PASON	-17.35	-7.81	0.77
PASOn	-13.31	-8.66	0.58
PASoN	-13.22	-6.88	-0.13
PASon	-14.36	-9.21	0.13
PAsON	-13.07	-6.54	0.44
PAsOn	-14.49	-7.92	-0.52
PAsoN	-12.29	-6.38	-0.11
PAson	-30.47	-20.53	0.55
PaSON	2.21	-11.83	2.33
PaSOn	1.24	-10.98	1.57
PaSoN	1.24	-8.70	1.01
PaSon	1.50	-8.07	0.98
PasON	1.34	-8.83	1.37
PasOn	0.93	-8.46	1.40
PasoN	1.35	-7.16	1.82
Pason	-3.62	-19.32	0.04
1: See Table 1-2 for details			

Appendix II

List of Single Nucleotide Polymorphisms from Type 1 Diabetes Genome-Wide Association Study

ID	Chr	Gene	ENSG	Strand	Position	Wt	Mut	prime_5	prime_3
jtt1d_1	2	GCG	ENSG00000115263	-1	163000583	C	T	ATTTTGGTCTGAATCAACCAGTTTATAAAGTCCCTGG	CGGCAAGATTATCAAGAATGGTGTTCATCTCATCAGA
jtt1d_2	2	GCG	ENSG00000115263	-1	163005674	G	A	TGCCTTGTACCAGCATTACAAATAATCCAGCCACAAA	GTAAATGCTTTTCATTTCTGCTGTCTGTCAGAACACA
jtt1d_3	2	AC007750.1	ENSG00000236841	1	163027570	G	A	GGGTGTATAAGTGGTTCGTGGACAGGCCGGATAAGCC	GTGGTTCTGGTCAGAGTACCACTGAAACACAAAGAAA
jtt1d_3	2	FAP	ENSG00000078098	-1	163027570	G	A	GGGTGTATAAGTGGTTCGTGGACAGGCCGGATAAGCC	GTGGTTCTGGTCAGAGTACCACTGAAACACAAAGAAA
jtt1d_4	2	FAP	ENSG00000078098	-1	163029378	T	C	GTGCATTAACCAGAGCTTTAGCAATCTGTGCTGAGTT	TTGAAAGTGCACATTATCTGCAACAAAGAGAGAGAGA
jtt1d_5	2	FAP	ENSG00000078098	-1	163055344	T	C	ACTTGCTGTGTAATATTGGCACCTTTCTTTCCTTAGA	TGGCAAGTAACACACTTCTTGCTTGGAGGATAGCTTC
jtt1d_6	2	FAP	ENSG00000078098	-1	163076380	G	A	TCAAAAATTTAAACACTTACTAATTTACTCCCAACAG	GCGACCAGCATAAATACTGAATTGGACGAGGAAGCTC
jtt1d_7	2	FAP	ENSG00000078098	-1	163080982	A	G	ATTGGATATAACCAAATTAAATAGTTGATACCTTTGA	ATAATCACTTTCTAGATATACAAATTGCCGATCAGGT
jtt1d_8	2	FAP	ENSG00000078098	-1	163081036	A	G	ATACAAATTGCCGATCAGGTGATAAGCCGTAATTTGA	AGCATTCACACTTTTCTGAAATTATGAAGAGGTTGAT
jtt1d_9	2	IFIH1	ENSG00000115267	-1	163123842	G	A	AATTATTTTTGAAAACCACTACAAAATTCCTTATTTT	GAGACAAGGCAAATCTAAGCCTTTGTGCACCATCATT
jtt1d_10	2	IFIH1	ENSG00000115267	-1	163124040	C	G	AGATGATTTCACCATTTATTTGATAGTCGGCACACTT	CTTTTGCAGTGCTTTGTTTTCTCTTACAATGTAAAGT
jtt1d_11	2	IFIH1	ENSG00000115267	-1	163124051	C	T	CCATTTATTTGATAGTCGGCACACTTCTTTTGCAGTG	CTTTGTTTTCTCTTACAATGTAAAGTTCCCTATAAGT
jtt1d_12	2	IFIH1	ENSG00000115267	-1	163124596	C	T	TTGTGGAAAAATGTAAAAATGGGTCTTTCTGGACTCA	CTTGAATTCTGGGGTCATATTGACGTGATGCATTTTC
jtt1d_13	2	IFIH1	ENSG00000115267	-1	163124637	T	C	AATTCTGGGGTCATATTGACGTGATGCATTTTCTCAA	TTACATGGATATCTTCCCCAGAACAGGCTAGCACACT
jtt1d_14	2	IFIH1	ENSG00000115267	-1	163128824	T	C	ATACATCATCTTCTCTCGGAAATCATTAACTGTCTCA	TGTTCGATAACTCCTGAACCACTGTGAGCAACCAGGA
jtt1d_15	2	IFIH1	ENSG00000115267	-1	163128828	C	T	ATCATCTTCTCTCGGAAATCATTAACTGTCTCATGTT	CGATAACTCCTGAACCACTGTGAGCAACCAGGACGTA
jtt1d_16	2	IFIH1	ENSG00000115267	-1	163128893	C	T	CAGGACGTAGGTGCTCTCATCAGCTCTGGCTCGACCA	CGGGCCTGAAAACACAAATAAATCAAGTAAATGAAAG
jtt1d_17	2	IFIH1	ENSG00000115267	-1	163133396	G	A	CCTAGTATATTGCTCCATTATGGTATTTCTTAATTTG	GTCAGCTTTTCATTTTCATATTCTGGGTTTTCAGCCA
jtt1d_18	2	IFIH1	ENSG00000115267	-1	163136505	C	G	TATTGTCAATCAATAGATATAAAACATTAAGCCCATA	CTTCTCTGGTTGCATCTGCAATGGCAAACTTCTTGCA
jtt1d_19	2	IFIH1	ENSG00000115267	-1	163136557	G	T	CTGCAATGGCAAACTTCTTGCATGGCTCCTGTATTTG	GTTTTTCAGTTGATCAAGGTTTTCTTTAACAGTTTTA
jtt1d_20	2	IFIH1	ENSG00000115267	-1	163137871	C	G	TTACTTTTAAAATGTGTTCTTCAGCTTTGGCTTGCTT	CGTGGCCCCTCCAACACCAGGTGAAGCTGTTAGTCCC
jtt1d_21	2	IFIH1	ENSG00000115267	-1	163137983	T	C	CTTGAGTCTATTGTTTTTCAACTTCTGCATCAAATAA	TGCCTCATGATGTTATTATACACTGCTTCTTTGTTGG
jtt1d_22	2	IFIH1	ENSG00000115267	-1	163144721	A	C	TTTTTTCTTCTTGTCTAAGTGATCCTTGGCAATGTAA	ACAGCCACTCTGGTTTTTCCACTCCCTGTAGGGAGGC
jtt1d_23	2	IFIH1	ENSG00000115267	-1	163167419	T	C	ACAATCCTTTTTAGTAGCTCTCTTACACCTGATTCAT	TTCCATTGTTTTCTGCAGCAGCAATCTGTTGTAAGAG
jtt1d_24	2	GCA	ENSG00000115271	1	163200678	T	C	TAGTGCGCCTTTCAGCCTCACCTGCAGCTGCGCCTCC	TTGCACCTGCGCCTGTGCTTTTTCTCCCAGCACTGCG
jtt1d_25	2	GCA	ENSG00000115271	1	163208893	T	G	AGTGGATGCTGAAGAACTTCAGAGATGTTTGACACAG	TCTGGAATTAATGGAACTTACTCTCGTGAGATCTTTT
jtt1d_26	2	GCA	ENSG00000115271	1	163217315	T	G	TGTCTGTGAAGAAGAAAATTATCTCCCTAGTTCAATC	TGTAGTGAAATAAGACTACAGAAGGCATTGTTTTTTC
jtt1d_27	2	KCNH7	ENSG00000184611	-1	163230011	T	C	ACTTACTTGTGAGGAAGGGCTGAAACTTCGGTCAGTT	TTGATGGATGCTTCCGGTTGACTGGTTCTCATCAGCT

jtt1d_28	2	KCNH7	ENSG00000184611	-1	163241287	C	G	TGTTTCTTCAAAATCGAGCCCAGATGCTTTCCCTATT	CCTGGAGAAGAGTCTACTATTCCTGAGAAGAGCGGCT
jtt1d_29	2	KCNH7	ENSG00000184611	-1	163250987	G	A	TGTCTCCTTCTGAATCATTCATGGATTGTGATCGTAG	GAGATCAGCCTTTGTTAATGGAACATGAAGATAAAAT
jtt1d_30	2	KCNH7	ENSG00000184611	-1	163279930	G	A	ATTCTTCAAGACGTTGCCTCAGAGGGTTGGGGATTTG	GTGAAAGCGAATGAACTCTTTTACTCGCAGCATCTGC
jtt1d_31	2	KCNH7	ENSG00000184611	-1	163292047	T	C	ACAGCAGCGCCATATTCTGAATATCGATCCAGTTTCC	TGGCCACGCGCACAAGACGGAGGAGTCGGGCAGTCTT
jtt1d_32	2	KCNH7	ENSG00000184611	-1	163302901	C	T	GAAAGGGCTGTAGTGCAATATCGTAAACTTGTTGATG	CGTGGTGTCTGCAGTTTGTATTCAGGTAGGACATCTG
jtt1d_33	2	KCNH7	ENSG00000184611	-1	163353469	T	C	TATTTTGTACTAATATCTACGGCCATACCACCCTGAG	TACATGTGATCTCATCTGATCTCAGAAATGCAACAGA
jtt1d_33	2	5S_rRNA	ENSG00000212312	1	163353469	T	C	TATTTTGTACTAATATCTACGGCCATACCACCCTGAG	TACATGTGATCTCATCTGATCTCAGAAATGCAACAGA
jtt1d_34	2	KCNH7	ENSG00000184611	-1	163360971	C	T	TAAGAAAATTGTGTCCTACCTGGGTCACTTTCTCAGT	CACATTGTGTGTTCGATCTTTAACCTTGGGTGCAATA
jtt1d_35	2	KCNH7	ENSG00000184611	-1	163361158	T	A	ATCTGATGTGGATCCCAGGAGGCTTGACTTGATATGA	TTAAAAGGCCCTAAAAAAATGGAAAGTATTTGTAAGA
jtt1d_36	2	CTLA4	ENSG00000163599	1	204732714	A	G	TGGATTTCAGCGGCACAAGGCTCAGCTGAACCTGGCT	ACCAGGACCTGGCCCTGCACTCTCCTGTTTTTTCTTC
jtt1d_37	2	CTLA4	ENSG00000163599	1	204738067	A	T	AAGGTTGTATTGCATATATACATATATATATATATAT	ATATATATATATATATATATATATATATATATATATA
jtt1d_38	2	CTLA4	ENSG00000163599	1	204738068	T	A	AGGTTGTATTGCATATATACATATATATATATATA	TATATATATATATATATATATATATATATATATATAT
jtt1d_39	2	CTLA4	ENSG00000163599	1	204738083	A	G	TATACATATATATATATATATATATATATATATAT	ATATATATATATATATATATATATTTTAATTTGATAG
jtt1d_40	2	CTLA4	ENSG00000163599	1	204738084	T	C	ATACATATATATATATATATATATATATATATATA	TATATATATATATATATATATATTTTAATTTGATAGT
jtt1d_41	2	CTLA4	ENSG00000163599	1	204738092	T	C	ATATATATATATATATATATATATATATATATATA	TATATATATATATATTTTAATTTGATAGTATTGTGCA
jtt1d_42	2	CTLA4	ENSG00000163599	1	204738094	T	C	ATATATATATATATATATATATATATATATATATA	TATATATATATATTTTAATTTGATAGTATTGTGCATA
jtt1d_43	2	ICOS	ENSG00000163600	1	204801577	C	T	GAAGTCAGGCCTCTGGTATTTCTTTCTCTTCTGCTTG	CGCATTAAAGTTTTAACAGGTAAGTGGTGTATTGAAT
jtt1d_44	2	ICOS	ENSG00000163600	1	204824324	A	C	TTATGCTGAATTTTTGTTACAGATGTGACCCTATAAT	ATGGAACTCTGGCACCCAGGCATGAAGCACGTTGGCC
jtt1d_45	2	ICOS	ENSG00000163600	1	204824355	T	C	TATAATATGGAACTCTGGCACCCAGGCATGAAGCACG	TTGGCCAGTTTTCCTCAACTTGAAGTGCAAGATTCTC
jtt1d_46	2	ICOS	ENSG00000163600	1	204824652	G	A	AGCAGTGCATCAGCCAGTAAAACAAACACATTTACAA	GAAAAATGTTTTAAAGATGCCAGGGGTACTGAATCTG
jtt1d_47	4	AC097533.2	ENSG00000237868	-1	122999081	G	A	GACAGGAACTGCTCGTCCACATACTGGGGTGTCCCAG	GGACAGCTGGAGGAGGCCGCCCATCATTTATATTAAA
jtt1d_48	4	AC097533.2	ENSG00000237868	-1	122999385	G	A	AGCTTCTCCTAGAATGTTCCTCACCCCTCTATTCTCC	GGCTCTGGCCTCTGTCCTTGAGGACGAGGAGGGGTCT
jtt1d_49	4	AC097533.2	ENSG00000237868	-1	122999452	T	C	GGGGTCTTCTCCCTGGGGTCCTGTTGCCCAGTGCTGC	TCCTTCCATAGAGGGGGATGATAAAAGGTATTTAAAA
jtt1d_50	4	AC097533.1	ENSG00000241037	-1	123008808	C	T	GGCCAGTCTGACCTGGGCCAGTTCACTCCTCCTTAGG	CAAACTGGCAGTCCCGGGAGGTCACCATATTGATGCT
jtt1d_51	4	KIAA1109	ENSG00000138688	1	123113428	A	G	TCCGGAAGAAACAGAAGAAAATATTGAAGGAGAAATG	AGCAGTGAGGATTGCAAATTACAAGACTTGCCTCCAT
jtt1d_52	4	KIAA1109	ENSG00000138688	1	123145751	T	A	TGAGAGCACACGCTATGTTCTCAGCAGAAGGTCTTCC	TTTGGGAAGCGATTCCTTAGAATACGCATGGTTAATT
jtt1d_53	4	KIAA1109	ENSG00000138688	1	123145825	C	T	GATGTGCAGGCTGGAAGTCTTACAGCTAAGGTCACAG	CACCACAGGTATGGTTTTCAGAGTACTATCTCCTGAC
jtt1d_54	4	KIAA1109	ENSG00000138688	1	123150286	T	C	TAGATGTACAATCCTTTTTTATTTTAGTTGAGTTGTA	TTCCTGGGCCTTGTCCAACTTCAGATGATTTGAAATA
jtt1d_55	4	KIAA1109	ENSG00000138688	1	123159256	A	G	TCACTTAACCCACCTTTTGTTTGTTTAGGTAATGTGA	ATGGCATGAAGAGGAAAGAATGGGAAAACAAATCAGT
jtt1d_56	4	KIAA1109	ENSG00000138688	1	123159262	T	A	AACCCACCTTTTGTTTGTTTAGGTAATGTGAATGGCA	TGAAGAGGAAAGAATGGGAAAACAAATCAGTGGGAAT

jtt1d_57	4	KIAA1109	ENSG00000138688	1	123159265	A	G	CCACCTTTTGTTTGTTTAGGTAATGTGAATGGCATGA	AGAGGAAAGAATGGGAAAACAAATCAGTGGGAATAGA
jtt1d_58	4	KIAA1109	ENSG00000138688	1	123159275	A	G	TTTGTTTAGGTAATGTGAATGGCATGAAGAGGAAAGA	ATGGGAAAACAAATCAGTGGGAATAGAAGTAGAGAGA
jtt1d_59	4	KIAA1109	ENSG00000138688	1	123159501	G	C	AGAGTTCTTGCATGGGACAAAAAGAGATGATGGCCAA	GCAAGGTCAGTATTCACTTAGATTTAGAAGCCTGATC
jtt1d_60	4	KIAA1109	ENSG00000138688	1	123160682	A	G	GGTAATAGCTTTGCTTTCTGTTTTAGTATCCCTACAG	AAATTTCAGGAAACAGCCCTGTGTCTCCTAATACTCA
jtt1d_61	4	KIAA1109	ENSG00000138688	1	123161331	C	T	AAACAGTGGAGAGTGAACAGATTACTCCGCAACAACC	CGTGATGAATTGTTATCAGACTTACCTTACTCAGTTC
jtt1d_62	4	KIAA1109	ENSG00000138688	1	123167847	T	C	CAGTTCTGAAATTTTGTGTGTGTGTTAGGTAAACTTA	TGTTTGTTACAAGCCTCAGTGGAAGAATCTCCAACTA
jtt1d_63	4	KIAA1109	ENSG00000138688	1	123168361	A	G	AAGAGACTTCAAACAATGCAGAACCTGGTAGAACATC	AAATTTTGATAGGTATGTTCATGCCACAAAGATGCAG
jtt1d_64	4	KIAA1109	ENSG00000138688	1	123171659	T	A	CTCCAACCGGCAGTGGCTATAATACTGATGTCTCTGA	TGATAATCTTCCATGTGACCGGACAAGCCCTTCCTCA
jtt1d_65	4	KIAA1109	ENSG00000138688	1	123175966	G	A	GTTTTTATAATTTTTTATTTATGTTAAAAAGGCAGCT	GAACCTTTTAAGCACTGCAACACCAGCTGTTGGTGCA
jtt1d_66	4	KIAA1109	ENSG00000138688	1	123176375	A	G	TTGCTCGCTTTCTCCAAGAAAATCCTTCATGTTTACT	ATGTAATATACTACACCACTATCTGCACCAGGCAAAT
jtt1d_67	4	KIAA1109	ENSG00000138688	1	123178574	T	C	TGAAACCTCAAATAGCTATGGACCATGAACATGAAGA	TGGACTTGGATTGGACAATGGGGGTGGTCTTCAAAGT
jtt1d_68	4	KIAA1109	ENSG00000138688	1	123178643	C	T	AAAGTGATACCAGTGCTGATGGAGCAGAATTTGAGTT	CGATGCAGGTAGTTTTGTAAGCCTCTATTGAGTACTT
jtt1d_69	4	KIAA1109	ENSG00000138688	1	123179900	C	T	CAGTGAACACACAATGCTATTAGAAGGAACAGCTAAC	CGGCCTCCACCTGGTAGCTCTGGACCTGTAACTGGAG
jtt1d_70	4	KIAA1109	ENSG00000138688	1	123184753	C	T	TCTCCTGTTACAAAATCAGGACACAATAGTCTTCCCA	CAGGTATTGAGTTATCACATTATTTTGCTTAACTGTC
jtt1d_71	4	KIAA1109	ENSG00000138688	1	123192240	A	C	GTTCAAACCAGCATTAATGTTGGGAACCTTTAGCATC	AGTGCTGTTGTAATGGAAAAGTCCGTGTGCACCCCTC
jtt1d_72	4	KIAA1109	ENSG00000138688	1	123192383	T	C	TTACTATTTCCTGTCAGTCAATAAGCCAGCATGTAGA	TATGGCTTTGGTTCGTCTTATTCATCAGTTTAGCACA
jtt1d_73	4	KIAA1109	ENSG00000138688	1	123201125	G	A	ACCAACTATCTAAACAAATCTCAGACCTAATCAGACA	GCCTTCTACAGCGTAAGTTATTTTATTTGTTCACATT
jtt1d_74	4	KIAA1109	ENSG00000138688	1	123207867	T	G	TATATAATGGAAGAACATGATAGTTATTCGGATCAGG	TGTGGAGTATAGATGAACTGCCTTCTAAACAAGGTTA
jtt1d_75	4	KIAA1109	ENSG00000138688	1	123229132	C	T	TATCAAAAGCTGTGCTGTTTTGGCTGAATTATAAGGC	CGCCTATGACAACTGGAATGAACAACGAATGGCTTTA
jtt1d_76	4	KIAA1109	ENSG00000138688	1	123245602	G	A	CTCTCCCATTTTGTTAGGCTGCTTCCCTAAAGGATAA	GTGGGGTTTGAGTTACAAACCAAGTTACAGCCGATCA
jtt1d_77	4	KIAA1109	ENSG00000138688	1	123249429	G	A	CTTTCCAAACTGAAGAGGGCCGACGGGATGACAGTTT	GTCTTCTACCAGTGAAGATTCCGAGAAGGATGAAAAA
jtt1d_78	4	KIAA1109	ENSG00000138688	1	123252539	C	T	AACAGGCTTTGCTGCTGTTCATCAGCTATTTACAGAA	CGCTGGCCAACAACACCAGTCAATAGAAGTCTTAGTG
jtt1d_79	4	KIAA1109	ENSG00000138688	1	123268859	A	G	CAATGAGCATATGACAAACAGCACCATGTCACCAGGG	ACAGTAGGACAGAGCCTAAAATCCCCAGCTTCCATAA
jtt1d_80	4	KIAA1109	ENSG00000138688	1	123271189	G	A	TGTCTACCTGGGGACCAGTTCCTTACCTTCCGCCAAA	GACAATGACTAGCAACCTAGAAAAAAGTTCACAAGAA
jtt1d_81	4	KIAA1109	ENSG00000138688	1	123274111	A	G	ATCATCGACACTGGCCTGGAGTATTGAAGGTGGTATC	AGGATGCCACATATCCTTATTTCAGATTCCATTACCA
jtt1d_82	4	KIAA1109	ENSG00000138688	1	123276961	G	A	ACTCTCTCTTTATCATTATGCATGTAGAGCTAAATCT	GCTTCGTAATGTTGATGCTAACAACACTGAGAATAGC
jtt1d_83	4	KIAA1109	ENSG00000138688	1	123277001	A	G	TCGTAATGTTGATGCTAACAACACTGAGAATAGCACT	ACTGTGAAGAATTCTAGTTTGTTGAGTGGATTCAGAG
jtt1d_84	4	KIAA1109	ENSG00000138688	1	123280860	T	C	CTGTGGACTGGAGAGATTTTATGTGCAATACATGGCA	TCTAGAACCTACTCTTAGGTAAGTAATGAGTATATAC
jtt1d_85	4	ADAD1	ENSG00000164113	1	123300267	G	A	TGGATGAAATGAGGGATTTTCTTGAAACAACCCTCAC	GCAGGTACCCCTTGGGCAGCCTTCAACCGCTCTGGGG
jtt1d_86	4	ADAD1	ENSG00000164113	1	123300599	C	T	GGCGCAAGCGCGGGGGCAAGAGCGCCGGCCTCCGAGA	CGGTTAGTGATTGGACGAAGCAGGGCGCGGGGGCGCA

jtt1d_87	4	ADAD1	ENSG00000164113	1	123300758	G	T	CGAGAGGTTGAGGCTGGGAGGTGGGAGCAACGGCGGC	GGCGGCCGCCTGCGAGCCCCCGGCCTGAGGCGCAGCA
jtt1d_88	4	ADAD1	ENSG00000164113	1	123302244	C	T	CTCCAAAAAAAATACCTAAGGAATTTATAATGAAATA	CAAACGTGGAGAGATAAATCCTGTGTCAGCCTTGCAC
jtt1d_89	4	ADAD1	ENSG00000164113	1	123302255	A	G	ATACCTAAGGAATTTATAATGAAATACAAACGTGGAG	AGATAAATCCTGTGTCAGCCTTGCACCAGTTTGCACA
jtt1d_90	4	IL2	ENSG00000109471	-1	123372753	C	A	TAAATAAGTGAAACCATTTTAGAGCCCCTAGGGCTTA	CAAAAAGAATCATAAAAGATCCATATTTATAGTTTTA
jtt1d_91	4	IL2	ENSG00000109471	-1	123377482	C	A	TTACATTAATTCCATTCAAAATCATCTGTAAATCCAG	CAGTAAATGCTCCAGTTGTAGCTGTGTTTTCTTTGTA
jtt1d_92	4	IL2	ENSG00000109471	-1	123377635	A	G	TGGCAGGAGTTGAGGTTACTGTGAGTAGTGATTAAAG	AGAGTGATAGGGAACTCTTGAACAAGAGATGCAATTT
jtt1d_93	4	IL21	ENSG00000138684	-1	123533820	C	T	CTCCTCCACTTGGAATACAAAGAAATGACTTTCACTA	CTATATTAGAGTATGTAACATAGTGTCCAACTGCAAG
jtt1d_94	4	IL21	ENSG00000138684	-1	123533834	G	A	ATACAAAGAAATGACTTTCACTACTATATTAGAGTAT	GTAACATAGTGTCCAACTGCAAGTTAGATCCTCAGGA
jtt1d_95	4	IL21	ENSG00000138684	-1	123536963	G	A	TTCCTGTATTTGCTGACTTTAGTTGGGCCTTCTGAAA	GCAGGAAAAAGCTGACCACTCACAGTTTGTCTGAAAG
jtt1d_96	10	IL2RA	ENSG00000134460	-1	6053568	C	T	GATCTTGCTCTGTTGCCCAGGCTGGAGTGCAGTGGTG	CCATCATGACTGACTGCAGCCTCGAACTCCTGGGCTC
jtt1d_97	10	IL2RA	ENSG00000134460	-1	6053809	C	T	GGGGTTATAGGCCTGAGCCACCGTGCCAGGCCTGATG	CGCTCTTTTCCGTGGTTACGTTCTACCAGTGTGACCT
jtt1d_98	10	IL2RA	ENSG00000134460	-1	6053866	C	T	TTCTACCAGTGTGACCTCCATCCCTTCTCCCTCTTCA	CTTCCTTCTTTCTTTCCTTCCTTGCATAAACATTGAA
jtt1d_99	10	IL2RA	ENSG00000134460	-1	6054083	T	C	CTTAAAGAGGCCAATTAGTAACGCACAGGTAAAACTT	TGCTAAGTATGATTCTCTGCCTGGGACCTCATTCATC
jtt1d_100	10	IL2RA	ENSG00000134460	-1	6054158	C	T	TTCTGTTCCTGACATTGCCTCATGGGTTTGGCTGCCC	CGTTTTGAAGTTACCCAAAGATTATTCTGCCATGGCC
jtt1d_101	10	IL2RA	ENSG00000134460	-1	6054765	C	T	GGATGTCTCCTGGGCGACCATTTAGCACCTTTGATTT	CACTTGGGCTTCATGACTTCTGTTGTCTGTTCCCGGC
jtt1d_102	10	IL2RA	ENSG00000134460	-1	6061401	G	T	TGTCCACAAAGCCAGTGCCCCACTCACCTGCTACCTG	GTACTCTGTTGTAAATATGGACGTCTCCATGGTTGCA
jtt1d_103	10	IL2RA	ENSG00000134460	-1	6063508	G	A	TGCATATGAGCTGGGGCTGGGTCCACCTTGTCTTCCC	GTGGGTCATTTTGCAGACGCTCTCAGCAGGACCTCTG
jtt1d_104	10	IL2RA	ENSG00000134460	-1	6066229	T	G	AGATTCATCTCTCACCTGGAAGGCTCGCTTGGTCCAC	TGGCTGCATTGGACTTTGCATTTCTGTGGTTTTCCTT
jtt1d_105	10	IL2RA	ENSG00000134460	-1	6066235	C	A	ATCTCTCACCTGGAAGGCTCGCTTGGTCCACTGGCTG	CATTGGACTTTGCATTTCTGTGGTTTTCCTTTCTTTC
jtt1d_106	10	IL2RA	ENSG00000134460	-1	6066236	A	C	TCTCTCACCTGGAAGGCTCGCTTGGTCCACTGGCTGC	ATTGGACTTTGCATTTCTGTGGTTTTCCTTTCTTTCT
jtt1d_107	10	IL2RA	ENSG00000134460	-1	6066302	G	A	TTCTTTCTGTTCTTCAGGTTGAGGTGTCACTTGTTTC	GTTGTGTTCCGAGTGGCTAGAAAATATAGATGGAATG
jtt1d_108	10	IL2RA	ENSG00000134460	-1	6067873	G	A	GGCTAGAGTTTCCTGTACAGAGCATATAGAGTGACCC	GCTTTTTATTCTGCGGAAACCTCTCTTGCATTCACAG
jtt1d_109	10	IL2RA	ENSG00000134460	-1	6067969	C	T	AGGCCATGGCTTTGAATGTGGCGTGTGGGATCTCTGG	CGGGTCATCGTCACAGAGCTCTGCAAAGCAAAAGAAG
jtt1d_109	10	AL137186.1	ENSG00000229664	1	6067969	C	T	AGGCCATGGCTTTGAATGTGGCGTGTGGGATCTCTGG	CGGGTCATCGTCACAGAGCTCTGCAAAGCAAAAGAAG
jtt1d_110	10	RP11-414H17.1	ENSG00000214015	-1	6113523	A	G	TTTGCTGCACAGCTTGGCACTGGGATTGGTGACTCCA	ATGGGCAGCTGGGCCACTGTTTCCAGGATGGCTTTGC
jtt1d_111	10	RP11-414H17.1	ENSG00000214015	-1	6113666	G	A	GCTTCTTGACGGGGTGGACCAGGAGCTTCTGGAAGCC	GCTGGGCAGCATGTACTCTGCTGCTCCCATAACCTGT
jtt1d_112	10	RP11-414H17.1	ENSG00000214015	-1	6113693	C	T	TCTGGAAGCCGCTGGGCAGCATGTACTCTGCTGCTCC	CATAACCTGTGCCAGGCATCAAAATTAGGCCCTTGAT
jtt1d_113	10	RP11-414H17.1	ENSG00000214015	-1	6113782	C	T	GTTGTCAAGAGGAGTTGTTGATGCCTCTGCATTTCTG	CCAGTTATGCTAATTTTGGGATATTGGCCTGACTGGT
jtt1d_114	10	RP11-414H17.1	ENSG00000214015	-1	6113805	T	A	CCTCTGCATTTCTGCCAGTTATGCTAATTTTGGGATA	TTGGCCTGACTGGTACTGGATGAACTTCTTGTCCTCT
jtt1d_115	10	RBM17	ENSG00000134453	1	6152058	C	T	CAGTGTACGAGGAACAAGACAGACCGAGATCTCCAAC	CGGACCTAGCAACTCCTTCCTCGCTAACATGGGGTAA

jtt1d_116	10	RBM17	ENSG00000134453	1	6154308	C	T	AGAAGACCAGCAAGCGTGGCGGCAAGATCATCGTGGG	CGACGCCACAGAGAAAGGTGTGTCCCCAGGGAAGCGT
jtt1d_117	10	RBM17	ENSG00000134453	1	6157654	G	A	TACCAAGACTCTTGGAAGGACTTCTAAGATATATGTT	GATTGATCCCTTTTTTATTTTGTGGTTTTTTAATATA
jtt1d_118	10	RBM17	ENSG00000134453	1	6158327	T	G	TTTTGGAAATGGCAGTTCCTTGGGGTCATGTTTCTAC	TGGCAAAATTTGCAATAGTGTTCTATTGTATGTAATT
jtt1d_119	10	RBM17	ENSG00000134453	1	6158386	C	T	CTATTGTATGTAATTTTAAAATTTATAAGATTATCCA	CGTTGGCCAAGTAAACTGTACTGCCAATAGAATTCTG
jtt1d_120	10	RBM17	ENSG00000134453	1	6158412	A	G	AAGATTATCCACGTTGGCCAAGTAAACTGTACTGCCA	ATAGAATTCTGGAATTGTGAGAAATTGTATCATTGAA
jtt1d_121	10	RBM17	ENSG00000134453	1	6158575	C	T	AGGTATTTCCAGAAAATACTCATGCCTGTGTTCTGTT	CCTTGCTTTCCCAAATACTGCATGTGACTTTCCTAAG
jtt1d_122	10	RBM17	ENSG00000134453	1	6158806	T	C	TAACATAAATAAAAGAATAACATTTTATCTTTTGTGG	TATTATTTTATTGAATAAAATTGAGTTTTATGATAAA
jtt1d_123	10	PFKFB3	ENSG00000170525	1	6191733	G	A	GTGCGTCCCTCCCAAAGCTGTGTGCTCGGTCCAAGAG	GATGACCATCCCCAATAGAGGAGGACTCATCTTCAGT
jtt1d_123	10	7SK	ENSG00000201581	1	6191733	G	A	GTGCGTCCCTCCCAAAGCTGTGTGCTCGGTCCAAGAG	GATGACCATCCCCAATAGAGGAGGACTCATCTTCAGT
jtt1d_124	10	PFKFB3	ENSG00000170525	1	6191735	T	C	GCGTCCCTCCCAAAGCTGTGTGCTCGGTCCAAGAGGA	TGACCATCCCCAATAGAGGAGGACTCATCTTCAGTCA
jtt1d_124	10	7SK	ENSG00000201581	1	6191735	T	C	GCGTCCCTCCCAAAGCTGTGTGCTCGGTCCAAGAGGA	TGACCATCCCCAATAGAGGAGGACTCATCTTCAGTCA
jtt1d_125	10	PFKFB3	ENSG00000170525	1	6191884	T	C	TAATATCCTCCAGTTCCATCCCAAGGATTTCACTCTT	TTTTATGGCTGAGTAGTATTCCATGTTGTATATGTAC
jtt1d_125	10	7SK	ENSG00000201581	1	6191884	T	C	TAATATCCTCCAGTTCCATCCCAAGGATTTCACTCTT	TTTTATGGCTGAGTAGTATTCCATGTTGTATATGTAC
jtt1d_126	10	PFKFB3	ENSG00000170525	1	6259115	G	A	CCCCACCTCACTTCCACCAGGCGTTTTTCATCGAGTC	GGTGTGCGACGACCCTACAGTTGTGGCCTCCAATATC
jtt1d_127	10	PFKFB3	ENSG00000170525	1	6262702	C	T	TGGTGAACCGGGTGCAGGACCACATCCAGAGCCGCAT	CGTGTACTACCTGATGAACATCCACGTGCAGCCGCGT
jtt1d_128	10	PFKFB3	ENSG00000170525	1	6264883	G	A	AGCCAGTGATCATGGAGCTGGAGCGGCAGGAGAATGT	GCTGGTCATCTGCCACCAGGCCGTCCTGCGCTGCCTG
jtt1d_129	10	PFKFB3	ENSG00000170525	1	6266128	C	T	CCCCCATCCCACGCCCTCCAGGCTGCCGTGTGGAATC	CATCTACCTGAACGTGGAGTCCGTCTGCACACACCGG
jtt1d_130	10	PFKFB3	ENSG00000170525	1	6268205	G	A	GACCTAACCCGCTCATGAGACGCAATAGTGTCACCCC	GCTAGCCAGCCCCGAACCCACCAAAAAGCCTCGCATC
jtt1d_131	12	DGKA	ENSG00000065357	1	56347577	C	T	CTTCTTGAGCTAAGGGGGACACCCTTGGCCTCCAAGC	CAGCCTTGAACCCACCTCCCTGTCCCTGGACTCTACT
jtt1d_132	12	SILV	ENSG00000185664	-1	56348028	G	A	AGAGTACTCAGACCTGCTGCCCACTGAGGAGGGGGCT	GTTCTCACCAATGGGACAAGAGCAGAAGATGCGGGGT
jtt1d_133	12	SILV	ENSG00000185664	-1	56351346	G	A	AGGTGTAGGAGAGGTCAGCTTCAGCCAGATAGCCACT	GGGGTCATGGAGCTGGAGGGCAAAGGTCAGAGGCTGA
jtt1d_134	12	CDK2	ENSG00000123374	1	56360876	G	A	GAGTTGTGTACAAAGCCAGAAACAAGTTGACGGGAGA	GGTGGTGGCGCTTAAGAAAATCCGCCTGGACACGTGA
jtt1d_135	12	CDK2	ENSG00000123374	1	56362711	T	G	TAGCAGACTTTGGACTAGCCAGAGCTTTTGGAGTCCC	TGTTCGTACTTACACCCATGAGGTGAGTCCCTTTATG
jtt1d_136	12	CDK2	ENSG00000123374	1	56365699	C	A	CTGAAGAGGGTTGGTATAAAAATAATTTTAAAAAAGC	CTTCCTACACGTTAGATTTGCCGTACCAATCTCTGAA
jtt1d_137	12	CDK2	ENSG00000123374	1	56365722	T	A	AATTTTAAAAAAGCCTTCCTACACGTTAGATTTGCCG	TACCAATCTCTGAATGCCCCATAATTATTATTTCCAG
jtt1d_138	12	CDK2	ENSG00000123374	1	56366031	A	C	AAAATGATTGGCCCCAGTCCCCTTGTTTGTCCCTTCT	ACAGGCATGAGGAATCTGGGAGGCCCTGAGACAGGGA
jtt1d_139	12	CDK2	ENSG00000123374	1	56366040	A	T	GGCCCCAGTCCCCTTGTTTGTCCCTTCTACAGGCATG	AGGAATCTGGGAGGCCCTGAGACAGGGATTGTGCTTC
jtt1d_140	12	CDK2	ENSG00000123374	1	56366160	G	C	TGTTTGAATTTTTCTCTTCCTTTTAGTATTCTTAGTT	GTTCAGTTGCCAAGGATCCCTGATCCCATTTTCCTCT
jtt1d_141	12	RAB5B	ENSG00000111540	1	56367901	C	G	GCTGCAGCTGTTTGTCTGTTCGACACAGGCTTGGGGC	CGACGGGGGAGACGGAGCCCCAGGTACCGAGCTGATG
jtt1d_142	12	RAB5B	ENSG00000111540	1	56374318	C	T	TCCTGATCTCCGGCCTCCTGACTTGAGCAAGATGTCC	CGGGCCAGGGAACTAAAAGCCTCATCCACATTCATAC

jtt1d_142	12	AC034102.1	ENSG00000237493	-1	56374318	C	T	TCCTGATCTCCGGCCTCCTGACTTGAGCAAGATGTCC	CGGGCCAGGGAACTAAAAGCCTCATCCACATTCATAC
jtt1d_143	12	RAB5B	ENSG00000111540	1	56374612	C	T	GGCTCCACGGTAGTAGGCAGTAGTTATTGTCTTGAAC	CGCTCTTGGCCAGCTGTGTCCCAGACTTGTAGTTTGA
jtt1d_143	12	AC034102.1	ENSG00000237493	-1	56374612	C	T	GGCTCCACGGTAGTAGGCAGTAGTTATTGTCTTGAAC	CGCTCTTGGCCAGCTGTGTCCCAGACTTGTAGTTTGA
jtt1d_144	12	RAB5B	ENSG00000111540	1	56374695	G	A	CCTCTATATCCACAGTGCGGATCTTGAAATCAATTCC	GATGGTGGAGATGTAAGTGTTGTTGAAGTTGTCCTCT
jtt1d_144	12	AC034102.1	ENSG00000237493	-1	56374695	G	A	CCTCTATATCCACAGTGCGGATCTTGAAATCAATTCC	GATGGTGGAGATGTAAGTGTTGTTGAAGTTGTCCTCT
jtt1d_145	12	RAB5B	ENSG00000111540	1	56374803	G	A	CCGAGTCCCCGATCAGCAGCAACTTGAAGAGGTGGTC	GTAGGCTTTGGCCATGGCGGACACCGGGGGAGCCGGG
jtt1d_146	12	RAB5B	ENSG00000111540	1	56386076	A	G	CTAAGAAATAACCTCCATCCCTACCCCTCAGCACACA	ACCCCTACGGTAACAGCACACTGAGCCCTGGCTCCCA
jtt1d_147	12	RAB5B	ENSG00000111540	1	56388136	G	T	CCCCCTTGGAGCAGGAGTGAAGATGTTTCATTATCTT	GGGCCTGGGAAACCACTTCCCCAGGCTTCTCCCTCCC
jtt1d_148	12	RAB5B	ENSG00000111540	1	56388137	G	T	CCCCTTGGAGCAGGAGTGAAGATGTTTCATTATCTTG	GGCCTGGGAAACCACTTCCCCAGGCTTCTCCCTCCCC
jtt1d_149	12	SUOX	ENSG00000139531	1	56391486	C	T	TGCGAGTCAGCCCTACCTGCACTGCTCTGGTCTAGTA	CAAACAGGCTGCTGGCATTGAGGTAGGTGGCAGAGAG
jtt1d_150	12	SUOX	ENSG00000139531	1	56395378	G	A	ATCCCCAGTGGATAAGGGGGTACTACTGTACTTGTTG	GCTCTTCATGTTAGCTCTGCTAGGCAGATGTCATTTC
jtt1d_151	12	SUOX	ENSG00000139531	1	56395420	T	A	TCATGTTAGCTCTGCTAGGCAGATGTCATTTCAGAGA	TGAGGAAGCAAGTTCAGAACGGCTTGGAATCTTGCTC
jtt1d_152	12	SUOX	ENSG00000139531	1	56395439	C	T	CAGATGTCATTTCAGAGATGAGGAAGCAAGTTCAGAA	CGGCTTGGAATCTTGCTCAGGAAATCGGGCTGGTTAA
jtt1d_153	12	SUOX	ENSG00000139531	1	56395577	T	C	ACCCATTTAGGCTGTCACTACTTTTTTTTCACTTTTT	TATCCCTGTTTAAGTCAGTCTGACCCACAGTTGTCCT
jtt1d_154	12	SUOX	ENSG00000139531	1	56395689	G	A	TTGGTTTCGGTCCTTTAGGCCCTTCGCCCCAGGCATC	GTTCTCTATGGTGGACAAAGTTCAGAATGGAAGATGG
jtt1d_155	12	SUOX	ENSG00000139531	1	56397807	C	T	CCCTGAGCTGCTGACAGAAAACTACATCACACCCAAC	CCTATCTTCTTCACCCGGAACCATCTGCCTGTACCTA
jtt1d_156	12	SUOX	ENSG00000139531	1	56398348	A	C	TGGCTGGGCAGAGTGAGTGTGCAGCCAGAGGAAAGTT	ACAGCCACTGGCAACGGCGGGATTACAAAGGCTTCTC
jtt1d_157	12	SUOX	ENSG00000139531	1	56398454	G	C	ACTCTGCTCCATCCATTCAGGAACTTCCTGTCCAGTC	GGCCATCACAGAGCCCCGGGATGGAGAGACTGTAGAA
jtt1d_158	12	SUOX	ENSG00000139531	1	56398531	G	A	GGGGAGGTGACCATCAAGGGCTATGCATGGAGTGGTG	GTGGCAGGGCTGTGATCCGGGTGGATGTGTCTCTGGA
jtt1d_159	12	IKZF4	ENSG00000123411	1	56415076	T	A	TCTCCCCTCTCCTTCTCTCCCTCTCTCTCTCTCTCTC	TCTCACACACACACACACACACACACTCAACACACAT
jtt1d_160	12	IKZF4	ENSG00000123411	1	56415078	T	A	TCCCCTCTCCTTCTCTCCCTCTCTCTCTCTCTCTCTC	TCACACACACACACACACACACACTCAACACACATAC
jtt1d_161	12	IKZF4	ENSG00000123411	1	56415317	G	A	GCATACACCACCCGCACTCCCTCGCCGTTTCCAAGGC	GGCGGCCGCGTTCGCACCCCAGGGTCTCACCGGCAAG
jtt1d_162	12	IKZF4	ENSG00000123411	1	56415348	G	A	CAAGGCGGCGGCCGCGTTCGCACCCCAGGGTCTCACC	GGCAAGGGAAGGATAATGTAAGTTCAGGCAGAAGGCG
jtt1d_163	12	IKZF4	ENSG00000123411	1	56429575	A	G	AGCTTCTTGCTTTAAGTCCTCACCCTTTACATTATCT	AATTCTTCAGTTTTGATGCTGATACCTGCCCCCGGCC
jtt1d_164	12	IKZF4	ENSG00000123411	1	56429707	A	G	TACCTCTTGTGCCCTCTCACTTTAGGCAGCTTGCACT	ATTCTTGAATGAATGAAGAATTATTTCCTCATTTGGA
jtt1d_165	12	IKZF4	ENSG00000123411	1	56430367	C	T	TTCTCTCTTCTAATTTTCAGTATAACCAAAAATTATC	CCAGCATGAGCACGGGCACGTGCCCTTCACCCCATTC
jtt1d_166	12	IKZF4	ENSG00000123411	1	56430764	T	C	CAAGTTGTAACTCTTGGTCCTTCTCTCTCTCCTTTTC	TCTTCCCTTCCTTCCCCTTCCATCTTTCTTTCCACAT
jtt1d_167	12	IKZF4	ENSG00000123411	1	56431851	C	T	GCAGCTTCTTTCCTTGTGTACATAATATATATATATA	CATATATATATATATTTTTAATCAGAAGTTATGAAGA
jtt1d_168	12	RPS26P20	ENSG00000197728	1	56435929	C	G	ATGCTATATAGGAGGGCCCTGCCAGGCACCGTCTCCT	CTCTCCGGTCCGTGCCTCCAAGATGGTGAGTCTTCTT
jtt1d_169	12	RPS26P20	ENSG00000197728	1	56437235	A	G	CAATTCACAGCAAAGTAGTCAGGAATCGATCTCGTGA	AGCCCGCAAGGACCGAACACCCCCACCCCGATTTAGA

jtt1d_170	12	ERBB3	ENSG00000065361	1	56473892	A	T	CCCTCTGCGTTCCTCCCTCCCTCTCTCTCTCTCTC	ACACACACACACCCCTCCCCTGCCATCCCTCCCCGGA
jtt1d_171	12	ERBB3	ENSG00000065361	1	56478809	G	A	GTCACAGTGGATTCGAGAAGTGACAGGCTATGTCCTC	GTGGCCATGAATGAATTCTCTACTCTACCATTGCCCA
jtt1d_172	12	ERBB3	ENSG00000065361	1	56481661	C	T	GGCCCAACCCCAACCAGTGCTGCCATGATGAGTGTGC	CGGGGGCTGCTCAGGCCCTCAGGACACAGACTGCTTT
jtt1d_173	12	ERBB3	ENSG00000065361	1	56486826	A	C	CTGGCCGCCCCACATGCACAACTTCAGTGTTTTTTCC	AATTTGACAACCATTGGAGGCAGAAGCCTCTACAAGT
jtt1d_174	12	ERBB3	ENSG00000065361	1	56487201	T	C	ATGTCACATCTCTGGGCTTCCGATCCCTGAAGGAAAT	TAGTGCTGGGCGTATCTATATAAGTGCCAATAGGCAG
jtt1d_175	12	ERBB3	ENSG00000065361	1	56490379	C	T	TCAAAGAGACAGAGCTAAGGAAGCTTAAAGTGCTTGG	CTCGGGTGTCTTTGGAACTGTGCACAAAGTGAGTGAC
jtt1d_176	12	ERBB3	ENSG00000065361	1	56494932	T	C	ACACCCAATGCCACGGGGATGCCTGGCATCAGAGTCA	TCAGAGGGGCATGTAACAGGCTCTGAGGCTGAGCTCC
jtt1d_177	12	ERBB3	ENSG00000065361	1	56494991	G	A	CTGAGGCTGAGCTCCAGGAGAAAGTGTCAATGTGTAG	GAGCCGGAGCAGGAGCCGGAGCCCACGGCCACGCGGA
jtt1d_178	12	ERBB3	ENSG00000065361	1	56494998	A	T	TGAGCTCCAGGAGAAAGTGTCAATGTGTAGGAGCCGG	AGCAGGAGCCGGAGCCCACGGCCACGCGGAGATAGCG
jtt1d_179	12	ERBB3	ENSG00000065361	1	56495049	C	A	CCCACGGCCACGCGGAGATAGCGCCTACCATTCCCAG	CGCCACAGTCTGCTGACTCCTGTTACCCCACTCTCCC
jtt1d_180	12	ERBB3	ENSG00000065361	1	56495339	C	A	TCAACCCCCAGGTACTCCCTCCTCCCGGGAAGGCACC	CTTTCTTCAGTGGGTCTCAGTTCTGTCCTGGGTACTG
jtt1d_181	12	ERBB3	ENSG00000065361	1	56496809	A	G	AGCCTTAAAGAGATGAAATAAATTAAGCAGTAGATCC	AGGATGCAAAATCCTCCCAATTCCTGTGCATGTGCTC
jtt1d_182	12	ERBB3	ENSG00000065361	1	56496940	A	C	TGTTTCTTGTTTTTGCACTGAATCAAGTCTAACCCCA	ACAGCCACATCCTCCTATACCTAGACATCTCATCTCA
jtt1d_183	12	PA2G4	ENSG00000170515	1	56503030	A	G	CATTTGATGTTGTACTTCTAGAACACACAAGTGACAG	AAGCCTGGAACAAAGTTGCCCACTCATTTAACTGCAC
jtt1d_184	12	PA2G4	ENSG00000170515	1	56504991	T	C	TTTTTGCCATAGGTGAATTTGTTGCCCAGTTTAAATT	TACAGTTCTGCTCATGCCCAATGGCCCCATGCGGATA
jtt1d_185	12	ZC3H10	ENSG00000135482	1	56514749	G	C	CCTTGGCCTTTCACCGGCTGACCTACCAAATGGCAAG	GAGGAGGTCCCTATCTGCCGTGACTTTCTCAAGGGTG
jtt1d_186	12	ZC3H10	ENSG00000135482	1	56516156	A	G	GTTGGACAATACAGGAATTGCTTCTGGGCCCTGGGAA	AGCTGGGACCATAGTGCTCCAGCCCAAAGACTAGGGG
jtt1d_187	12	ESYT1	ENSG00000139641	1	56527373	G	T	GGTATCTGATCTCTACTACATCTCAATTTCTTCTAGT	GGTTCCCTCTACAAGGTGGGCAAGGCCAAGTTCACTT
jtt1d_188	12	ESYT1	ENSG00000139641	1	56528164	G	A	CTGTCTACAGTACCAACTGCCCAGTGTGGGAGGAAGC	GTTCCGGTTCTTCCTACAAGACCCTCAAAGCCAGGAG
jtt1d_189	12	ESYT1	ENSG00000139641	1	56531154	T	C	GACCCTGTCACACGACTCCTGATAGCCAGTTTGGGAC	TGAGGTGAGTCTATATCTGGAAAGGACTAGGGTCTGT
jtt1d_190	12	ESYT1	ENSG00000139641	1	56531660	C	T	TCACATCAGTTCCAGGCCAAGAGCTAGAGGTTGAAGT	CTTTGACAAGGACTTGGACAAGGATGATTTTCTGGGC
jtt1d_191	12	ESYT1	ENSG00000139641	1	56532009	C	T	GACCCTGGAGGATGTCCCATCTGGCCGCCTGCACTTG	CGCCTGGAGCGTCTCACCCCCCGTCCCACTGCTGCTG
jtt1d_192	12	ESYT1	ENSG00000139641	1	56536711	G	T	TACAGTGAAGAACGAAAGCTGGTCAGCATTGTTCATG	GTTGCCGGTGAGACCCCATCCCTCCTGTCCTCCAGAT
jtt1d_193	12	ESYT1	ENSG00000139641	1	56537387	A	C	ATCTTCCAACACAGGTGCAGCTGGACCTAGCTGAGAC	AGACCTTTCCCAGGGTGTAGCCCGGTGGTGAGTGTCT
jtt1d_194	12	ESYT1	ENSG00000139641	1	56538340	T	G	GAGAGGGCTTTGGAGGACTTGGGACAGCAGGGCCAAT	TTTTTTGCCCAAGTGCCTAGGCTGCTAACTCACTGAC
jtt1d_195	12	MYL6	ENSG00000196465	1	56548970	C	T	CCCCACCAACGCCGAGGTGCTCAAGGTCCTGGGGAAC	CCCAAGAGTGATGGTGAGGGGACCCTTGGGAACAATT
jtt1d_196	12	MYL6	ENSG00000092841	1	56554411	G	A	CTTGTCTTCACCATGAATGTCTCTTCCTTCCTGCAGC	GTTTGTGAGGCATATCCTGTCGGGGTGACGGGCCCAT
jtt1d_197	12	MYL6	ENSG00000092841	1	56554415	G	C	TCTTCACCATGAATGTCTCTTCCTTCCTGCAGCGTTT	GTGAGGCATATCCTGTCGGGGTGACGGGCCCATGGGG
jtt1d_198	12	SMARCC2	ENSG00000139613	-1	56556817	T	C	TTAAAAACAAAACTGACATTCAGAGGGAAAGGAATCA	TTGGCTGAGCTGGGGTGGCCTAAAACAGCAACAATGA
jtt1d_199	12	SMARCC2	ENSG00000139613	-1	56556911	G	C	TTAGTACGAATGAACTCGAATAAGCTCAGCGTAGGGT	GGGGGAGGGGAGTTGGGGCCTTGACTTAGTCACTAAA

jtt1d_200	12	SMARCC2	ENSG00000139613	-1	56556957	G	A	GAGTTGGGGCCTTGACTTAGTCACTAAAAAGGGGCTT	GGGGAGAGATGGAATCTGCGCCCTGTTCTATCCCCAG
jtt1d_201	12	SMARCC2	ENSG00000139613	-1	56557345	C	T	CATGCCTCTGTTAGGCATGGTGAGGGCTTTGGAGGGG	CGGAAGGAGCTTTCCTTACGTAGTGATGAACTCTCCA
jtt1d_202	12	SMARCC2	ENSG00000139613	-1	56558351	G	C	GGAGGAGCGGGGAGGTTAATACTGATGGAGTCAGCTA	GACTACCAAATGGGATGATGGATGGAGCAGGAGGAGG
jtt1d_203	12	SMARCC2	ENSG00000139613	-1	56558397	C	T	ATGGGATGATGGATGGAGCAGGAGGAGGAGGAGGAGG	CGGCATGCCAAAAGGCAAACCCAAAGGAGCATTACCC
jtt1d_204	12	SMARCC2	ENSG00000139613	-1	56559142	A	G	GGGGTCCAGGGGGGGGAACCCCTGGTGGGACTGCCCC	AGGCTGGGGGGCTCCAGCTGGTTGCTGCTGCTGTGGC
jtt1d_205	12	RNF41	ENSG00000181852	-1	56599106	C	A	GGTTACAGAGGTTTGGGGCAGATATCGTAAGTTCAGC	CGAGGACTCCCTTGGCTCTTCCTATATTAAAGCCAAA
jtt1d_206	12	RNF41	ENSG00000181852	-1	56599270	A	T	GTTTCAAGTTTGATTTTTTTTTCTTTTTTCCTTTCTT	AAAAAAAAAAAAAGGAAGTAAATAAATTAAATTGCCA
jtt1d_207	12	RNF41	ENSG00000181852	-1	56599366	G	A	GATCAGGCCATTCTTAAAAAAAAGAGGGGGGGGGGCA	GTAGGTGGAGTTTGTGAAATATAAACAAACAATGGCC
jtt1d_208	12	RNF41	ENSG00000181852	-1	56599769	T	C	GCCATCAGTGATGGCCAATTAACGGCCTCACTACTTA	TTTCTAGAGATTTGGCTCCACCCTTACCATTTCTTCA
jtt1d_209	12	RNF41	ENSG00000181852	-1	56600095	T	C	CCTGAAAGGCTGAGCAAACTACCCCAAGGCCCTTCAG	TGCCAGAAGGGCAGGGAGATGTGTGGCTCAGGTATAA
jtt1d_210	12	RNF41	ENSG00000181852	-1	56600528	C	T	AGCGCTTGATTACAGCCTGGAGCACAGCATCAGGAGT	CGAGATCATCCCTCCCCAGCGGGTCACTCTTGCTGGC
jtt1d_211	12	OBFC2B	ENSG00000139579	1	56618292	G	A	GATGGACCGAGTCCCGGCTTGTCGGGATGAGGGTTCC	GGAAGATCTGGCCAGTAAGATTCTACTCCCTGGCTGT
jtt1d_212	12	OBFC2B	ENSG00000139579	1	56618305	A	C	CCGGCTTGTCGGGATGAGGGTTCCGGAAGATCTGGCC	AGTAAGATTCTACTCCCTGGCTGTGCACCGGGTTCCC
jtt1d_213	12	OBFC2B	ENSG00000139579	1	56618412	C	A	GCTTGAGACTAAAAGAGCATCCCGGCAGGGGGCCTTC	CAGCCCCAAAGCAGCCTGTCCAGAGACCCCCAAATTC
jtt1d_214	12	OBFC2B	ENSG00000139579	1	56623347	C	T	GGAGTTCAAGACCAGCCTGACCAACATGGAGAAACCC	CGTGTCTACTAAAAATACAGAATTAGCCAGGCATGGT
jtt1d_215	12	OBFC2B	ENSG00000139579	1	56623525	A	G	AGCCTGGGCAATAAGAGCGAAACTCCATCTCAAAAAA	AAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAATGGCA
jtt1d_216	12	OBFC2B	ENSG00000139579	1	56623529	G	A	TGGGCAATAAGAGCGAAACTCCATCTCAAAAAAAAAA	GAAAGAAAGAAAGAAAGAAAGAAAGAAATGGCAGTTA
jtt1d_217	12	OBFC2B	ENSG00000139579	1	56623533	G	A	CAATAAGAGCGAAACTCCATCTCAAAAAAAAAAGAAA	GAAAGAAAGAAAGAAAGAAAGAAATGGCAGTTACCAT
jtt1d_218	12	SLC39A5	ENSG00000139540	1	56625045	T	C	TCGGGGAGAATAGGAGCCAGAACCTGAGCCCCTAAGC	TATTCCCCTCACCAATGATGGGGTCCCCAGTGAGTCA
jtt1d_219	12	SLC39A5	ENSG00000139540	1	56626535	G	T	GATGTCTGGGCAGGGATGCCTCTGGGTCCCTCAGGGT	GGGGTGACCTGGAAGAGTCAAAGGCCCCTCACCTACC
jtt1d_221	12	SLC39A5	ENSG00000139540	1	56628700	G	A	TGACCCCTCGTCAGTTTGCTCTGCTGTGCCCAGCCCT	GCTTTATCAGATCGACAGCCGCGTCTGCATCGGCGCT
jtt1d_222	12	SLC39A5	ENSG00000139540	1	56628706	T	C	CTCGTCAGTTTGCTCTGCTGTGCCCAGCCCTGCTTTA	TCAGATCGACAGCCGCGTCTGCATCGGCGCTCCGGCC
jtt1d_223	12	SLC39A5	ENSG00000139540	1	56630444	G	C	CCACCAGCTCTGGCCCCTCCTGGGCACCAAGGCCACA	GTCATGGGCACCAGGGTGGCACTGATATCACGTGGAT
jtt1d_224	12	SLC39A5	ENSG00000139540	1	56630764	G	A	CTGATGGCTTCTCCAGCGGCCTCAGTACCACCTTAGC	GGTCTTCTGCCATGAGCTGCCCCACGAACTGGGTAGG
jtt1d_225	12	SLC39A5	ENSG00000139540	1	56630985	T	C	CTGCTCCAGTCAGGGCTGTCCTTTCGGCGGCTGCTGC	TGCTGAGCCTCGTGTCTGGAGCCCTGGGATTGGGGGG
jtt1d_226	12	ANKRD52	ENSG00000139645	-1	56632048	G	A	TTCTCTTTACTTCCTACAACCGAGTACATGGGTCACA	GGGTGGAGGGTGCAACAGGACATGGAACATGCCCCTC
jtt1d_227	12	ANKRD52	ENSG00000139645	-1	56632575	A	C	CACACATACAAAGCTGAGCTATCCAGGAACACAAGGG	AAACAAGGAGATTGTCCAGGGTGGGAGCGGAGGCAGC
jtt1d_228	12	ANKRD52	ENSG00000139645	-1	56633003	G	A	ATTTGGTCGCTTCTCTAGGGGTTGGGTTGGGAGGAGG	GAGCCCCCAAGGCAGACCCTTCCCTCTCTACCTCCCG
jtt1d_229	12	ANKRD52	ENSG00000139645	-1	56633209	C	A	CTTAAAAGGTAGGGTTCAAACTAGGCGGGATGGGGGC	CCATACTGGTTTGCCCCAGGAGTAGGGTTTCTGGGCT
jtt1d_230	12	ANKRD52	ENSG00000139645	-1	56633732	C	T	GGGCCATGGCTGGGGTTGGAGAGGGAGGTAGGCCCTC	CTCAGCCCCTCCACCCCAAGAAACACATCTACGTGGG

jtt1d_231	12	ANKRD52	ENSG00000139645	-1	56634583	C	G	GAGGACTCCCCACCACCTCCCAGCACTGCTGACAGTG	CGCTGAGGGCAGCAGGGCGCAGACAGCCCCCAGAAAT
jtt1d_232	12	ANKRD52	ENSG00000139645	-1	56634639	G	C	CAGACAGCCCCCAGAAATCTCATCTAGCAAGACAACG	GGGCTCTGACACTCAGACGCTTCCCGCCATCACCAAA
jtt1d_233	12	ANKRD52	ENSG00000139645	-1	56634891	T	G	TTGGGGATGGGGGCTGCAGCCCTTAAGAGAGGTCACA	TTGATTGTCATATAGGGGAGGACACGGTGTGGAGGGA
jtt1d_234	12	ANKRD52	ENSG00000139645	-1	56635080	G	C	GAACTGAGGGTTGGGAGAAGCAAAACACAGAGAGACA	GGGGATCAAAAGGGACCATGAAAAGATAAGGACTTGG
jtt1d_235	12	ANKRD52	ENSG00000139645	-1	56635640	G	A	CCCTCTTCACTCCAGCCCAGTTTGGCTTTTGGGGTGC	GACTTTAGAAATCTTATCAGTGCAGCCCCCAGCTCAA
jtt1d_236	12	ANKRD52	ENSG00000139645	-1	56636170	G	A	CTCCCAAAAGAAGATAAAAGAAAGAAGCAAGGTTAAA	GTGCGTGGTTAGGGGCCAGGCTAGGAGTGGGAGGGAA
jtt1d_237	12	ANKRD52	ENSG00000139645	-1	56636530	A	C	GACCGGCCGAGCAGGGAGGCAGTGATGGGTATGGAAG	AAGAGGGGATCTGCCTGGCAGTAGGGGCAGGGGAGAA
jtt1d_238	12	ANKRD52	ENSG00000139645	-1	56636962	G	A	GCTACTCAGAGTAGCAGCCATCTAACCCAATGGCGCC	GGGCCGCTCCTGGCTGTAGGGGCAGGAGGCCCCATGG
jtt1d_239	12	ANKRD52	ENSG00000139645	-1	56636975	C	G	GCAGCCATCTAACCCAATGGCGCCGGGCCGCTCCTGG	CTGTAGGGGCAGGAGGCCCCATGGGGCAGGGCGCCGC
jtt1d_240	12	ANKRD52	ENSG00000139645	-1	56639366	A	C	ATGCGTCGTGGTCCAGCAGGGCAGCCAGGCAGTCCTC	ACAGCCAGTCACTGCCTGTGAGTAACATGGGGGTGTG
jtt1d_241	12	ANKRD52	ENSG00000139645	-1	56645975	A	G	GGACCAGACTGGTAGCCTCACCTCCTGTAAGTGTCAG	AAGCGGCAGCGTAGTGGAGGGGAGAGCAGCCTTTACA
jtt1d_242	12	ANKRD52	ENSG00000139645	-1	56645978	C	G	CCAGACTGGTAGCCTCACCTCCTGTAAGTGTCAGAAG	CGGCAGCGTAGTGGAGGGGAGAGCAGCCTTTACAGTC
jtt1d_243	12	ANKRD52	ENSG00000139645	-1	56645996	G	T	CTCCTGTAAGTGTCAGAAGCGGCAGCGTAGTGGAGGG	GAGAGCAGCCTTTACAGTCGGCCTCGTTGACACCTGC
jtt1d_244	12	ANKRD52	ENSG00000139645	-1	56645997	A	G	TCCTGTAAGTGTCAGAAGCGGCAGCGTAGTGGAGGGG	AGAGCAGCCTTTACAGTCGGCCTCGTTGACACCTGCC
jtt1d_245	12	ANKRD52	ENSG00000139645	-1	56647911	G	C	TATTAACCAGTAGCTCCAAGCAGAGAGCGCCATTGGT	GGAGACTGCAGCCACATGCAGTGGCGTGAAGCCCTTG
jtt1d_246	12	ANKRD52	ENSG00000139645	-1	56649601	A	G	TCACCTCAAGATGCCCACTATGCACTGCATGGTGCAG	AGCACTGCGCCCGCTCCTGTCAGCCACGTTGAGGCTG
jtt1d_247	12	ANKRD52	ENSG00000139645	-1	56651618	A	T	GGTCTCGCCATCCTCCCTGCTCCCAACTCACCAGCAC	ATTGATGTTCTCCTTCTGCGAGAGTAGGGAACGCACT
jtt1d_248	12	COQ10A	ENSG00000135469	1	56662842	G	A	GCTGGCAGCTCCTTGGCCTGTGATTCTTCTTCTCCTA	GGTACTCAATGCAGGAGATGTATGAGGTGGTGTCCAA
jtt1d_249	12	COQ10A	ENSG00000135469	1	56664041	G	A	AGAATGTTGCTGCCTTTGAGCGTCGGGCAGCCACCAA	GTTTGGTCCAGAAACAGCCATCCCCCGTGAACTGATG
jtt1d_250	12	COQ10A	ENSG00000135469	1	56664084	G	T	TCCAGAAACAGCCATCCCCCGTGAACTGATGTTCCAT	GAGGTGCACCAGACTTGAGGCAAGGGATTGCTCCCTG
jtt1d_251	12	COQ10A	ENSG00000135469	1	56664231	C	T	AGTCTGTGTTCATAATACTGTTTCTCCTCTCAATTTC	CCAGAAATTGGGTTCTATGCTGGCTGGAAATGTTGGG
jtt1d_252	12	COQ10A	ENSG00000135469	1	56664433	T	C	CCTTATCAAGACACCTTAGTGTCTGACCAGGGGACGA	TAGTAACTTTTCTAAGGATTGAATAAATTGAGCTTTT
jtt1d_253	12	COQ10A	ENSG00000135469	1	56664743	A	G	TGGCCTGAGTTTTTATAAAATTTCAATAAATTGTGAC	AGTGTGAATTTGGCTTTATTATATTGTTTCTTGGGGC
jtt1d_254	12	CS	ENSG00000062485	-1	56666514	G	A	CTTATTGAGGGCTTGGCAGAGAAGCTAAAGCTCCAAA	GTGACTACAGATTCTCTGCAACCGGCTTTGACCCATG
jtt1d_255	12	CS	ENSG00000062485	-1	56666524	A	T	GCTTGGCAGAGAAGCTAAAGCTCCAAAGTGACTACAG	ATTCTCTGCAACCGGCTTTGACCCATGGAAACAGGAG
jtt1d_256	12	CS	ENSG00000062485	-1	56667528	G	A	ATTAGGCAGGTGTTTCAGAGCAAACTCTCGCTGACAG	GTATATCGCGGATCAGTCTTCCTTAGTACTGCATGGC
jtt1d_257	12	CS	ENSG00000062485	-1	56669799	G	A	TTGCTGGGTCTAGCTTACCTGTGGATGGTGAGGTACA	GGCGCGTGAGCTCAGTGAACTGATGATCAGTATAGCC
jtt1d_258	12	CS	ENSG00000062485	-1	56679751	A	G	TTTGGCCCACCACCGTCTTGCCATGTTGCTGCCTGAA	AGTCTTAATTCTGGCCTGCTCCTTAGGTATCAGGTCA
jtt1d_259	12	CNPY2	ENSG00000144785	-1	56705028	G	A	TCACCGCAAACTTGAGGGTGCCGCTAATATCTGAGTC	GATTCGGATGCCTTGTAGGTCCAGTTCACTGGATTCT
jtt1d_260	12	CNPY2	ENSG00000144785	-1	56708706	C	T	CGGAAAGATCCCATCTGAATGGTCTTCTTGGGGTCCA	CCTGGGCAATTTCCCATTCTAGTTCATCCACCAGAGC

jtt1d_261	12	CNPY2	ENSG00000144785	-1	56709652	G	A	AGGGCCGCCGCCGTGGCCCAAGCGCTGGAAGACCGCT	GGACTCTCACTTTGGCCCCAGTGGCTCCCGAAACTAC
jtt1d_262	12	PAN2	ENSG00000135473	-1	56711066	C	G	TGTCCAGGACTTCAATCCCTAGCCAGCTAGGAACTTA	CAGTTATGGTTCCAGGAGCTTCTCTGCCTGGATATTA
jtt1d_263	12	PAN2	ENSG00000135473	-1	56711235	G	A	TGAATCTGGCTCCTAGAACCTTTGCAACAATTCTGCT	GTGTTCCAGTTCAATTAATAGCACCATCTGTGACCCT
jtt1d_264	12	PAN2	ENSG00000135473	-1	56711371	C	T	CCAGTTCTGGGGCTATAGAACAGTAAAGGGAGAGGGC	CGTGGTTCTTTGGGAAGGGTAGTCAGAGCGCCAGCAC
jtt1d_265	12	PAN2	ENSG00000135473	-1	56716948	A	C	GAATGGAGAAAGGAAGCCAGACGTTCTTCAACTCCTC	AATGGAGGGACACACCAGCACACCCTCTGGACTCCCT
jtt1d_266	12	PAN2	ENSG00000135473	-1	56718422	T	C	AGAGGTCCAACATGTGAAACAGGAAGCCCAGCTCACA	TGCCAGACAGAACTCCTTCTGGCAAAGGTGGTTTTGA
jtt1d_267	12	PAN2	ENSG00000135473	-1	56718817	G	A	ACCCAACCCTTACCTGGATCATGCAGTTACAGTAGGC	GTTGGGAATGTGGGGCTCTAATCCAGCAAACAAGGTC
jtt1d_268	12	PAN2	ENSG00000135473	-1	56722060	T	G	TTCTGAGTCTCCTGGACAGTGTTAAGATCAATCTCTA	TTATGTGATTCTGCAGCCCACCAACGAGTAGAGTGCT
jtt1d_269	12	PAN2	ENSG00000135473	-1	56727705	G	A	TTTTGGAGCGCGTGGAATTAGAACGAGTAGGGGGAGC	GCAAGCGCTGTCAGCTCCGCGGGAAATTCCAGTTTCC
jtt1d_270	12	IL23A	ENSG00000110944	1	56733531	G	A	ACCAGGGTCTGATTTTTTATGAGAAGCTGCTAGGATC	GGATATTTTCACAGGGGAGCCTTCTCTGCTCCCTGAT
jtt1d_271	12	STAT2	ENSG00000170581	-1	56735599	A	G	GGAATAGCTAAGGTGTGAGATTGTCCAGAGTCCTATG	ACAGACCTTCAAGGTTTTAAGTTCCACAGACTTGGAC
jtt1d_272	12	STAT2	ENSG00000170581	-1	56735990	G	T	ATCTGTAATCCCAGCTACTGGGGAGGCTGAGGCAGGA	GAGTCACTTGAACCCGGAAGGCGGAGGTTGCAGTGAG
jtt1d_274	12	STAT2	ENSG00000170581	-1	56737126	G	T	TGATTCCCATCCTTGGAGAACAATATCATGCTATGAG	GAGTAGGAAGGGCAAGAGATATGAAAAGAACAGAGGA
jtt1d_275	12	STAT2	ENSG00000170581	-1	56737251	C	A	GGCTGGGGCGGGAGACGTAAACCTCATCCACGGTGTT	CTGGCCAGCCAACAGTGGGTCACCATTCGGCATGATT
jtt1d_276	12	STAT2	ENSG00000170581	-1	56740682	C	G	CTGACGATTCACTGAAGCGCAGTAGAAAGGTGCCAGA	CATGGTCTTCTTCAGCAGCCGGCGCTCCTGGCTCCGA
jtt1d_277	12	STAT2	ENSG00000170581	-1	56742994	C	A	GGGCTGAGCAAATTGAACCAGAGAACTGAAGCCCAGG	CAATTGAGAGCTGGTTCATGTTGGAAATAATCACCAC
jtt1d_278	12	STAT2	ENSG00000170581	-1	56742997	T	C	CTGAGCAAATTGAACCAGAGAACTGAAGCCCAGGCAA	TTGAGAGCTGGTTCATGTTGGAAATAATCACCACAGG
jtt1d_279	12	STAT2	ENSG00000170581	-1	56744612	C	G	GCCCGTGTCCTGGCCACCCTTCACCTGCTCCAATTTA	CCTGTCAATGGAGACTTCCACAGTCAGTGACTCATTG
jtt1d_280	12	STAT2	ENSG00000170581	-1	56745195	C	T	TAACCAGGCAACTCAGTCCCTTCAGCTCCTTCAGCAG	CTGCCTCAGGTGAAACAACAGCTTTGCTCCAGCTGTG
jtt1d_281	12	STAT2	ENSG00000170581	-1	56753870	C	T	GGCCCGTACCTGATTAGGGTTGCAGTCCCCGCGCCCT	CCAATGGCTCTGGTCGCGACTTCCCGTCCCTAGTATG
jtt1d_282	12	APOF	ENSG00000175336	-1	56754466	C	T	GCGATCTCGGCTCGCTGCAGCCTCGACCTCCCAGGCT	CAGGTGATTCTCCCGCCTCAGCCTCCCAGGTAGTTGG
jtt1d_283	12	APOF	ENSG00000175336	-1	56755058	A	G	ACTCCCAGCCCCAGGATCTAAGTCATAGCTCTTGATT	ATGGCCCACCCCCAGTAGGGAGCTGAACTTACTACTT
jtt1d_284	12	APOF	ENSG00000175336	-1	56755120	T	C	AACTTACTACTTCTGATATGAAAGAAGCCAGAGTAGT	TGTTTCTTCCAAGTCACTCACATCTGAGATGGCCCTC
jtt1d_285	12	APOF	ENSG00000175336	-1	56755474	C	G	GCTGGACTACATTGTGCACAGCTTGCTCCTTCTCATT	CTCACAGTCCTCTGTCGGGAGGGAGCGCCCGACCCTT
jtt1d_286	12	APOF	ENSG00000175336	-1	56755793	T	G	GTGGCTGAAACCAGGCAGTGACTTTGGGTGCAGAAAT	TGGCAGGACAAGGGGTCTGAGGAGGGTGTCTGGGATT
jtt1d_287	16	CIITA	ENSG00000179583	1	10992793	C	A	TTGGTCTCTGGTTTTTCTCAAAGTAGAGCACATAGGA	CCAGATGAAGTGATCGGTGAGAGTATGGAGATGCCAG
jtt1d_288	16	CIITA	ENSG00000179583	1	10995933	A	G	TGAGCCCCCCACTGTGGTGACTGGCAGTCTCCTAGTG	AGACCAGTGAGCGACTGCTCCACCCTGCCCTGCCTGC
jtt1d_289	16	CIITA	ENSG00000179583	1	10998628	C	G	CCATCTCCAGAGCACAAGACGTCCCCCACCCAATGCC	CGGCAGCTGGAGAGGTCTCCAACAAGCTTCCAAAATG
jtt1d_290	16	CIITA	ENSG00000179583	1	11000848	G	C	TTGAAGAGACCTGACCGCGTTCTGCTCATCCTAGACG	GCTTCGAGGAGCTGGAAGCGCAAGATGGCTTCCTGCA
jtt1d_291	16	CIITA	ENSG00000179583	1	11001032	C	T	CCCGGGGCCGCCTGGTCCAGAGCCTGAGCAAGGCCGA	CGCCCTATTTGAGCTGTCCGGCTTCTCCATGGAGCAG

jtt1d_292	16	CIITA	ENSG00000179583	1	11001421	C	A	CAGTTCCCATCCGCAGACGTGAGGACCTGGGCGATGG	CCAAAGGCTTAGTCCAACACCCACCGCGGGCCGCAGA
jtt1d_293	16	CIITA	ENSG00000179583	1	11001671	C	T	TCTTCCAGCCTCCCGCCCGCTGCCTGGGAGCCCTACT	CGGGCCATCGGCGGCTGCCTCGGTGGACAGGAAGCAG
jtt1d_294	16	CIITA	ENSG00000179583	1	11001680	G	T	CTCCCGCCCGCTGCCTGGGAGCCCTACTCGGGCCATC	GGCGGCTGCCTCGGTGGACAGGAAGCAGAAGGTGCTT
jtt1d_295	16	CIITA	ENSG00000179583	1	11001691	C	T	TGCCTGGGAGCCCTACTCGGGCCATCGGCGGCTGCCT	CGGTGGACAGGAAGCAGAAGGTGCTTGCGAGGTACCT
jtt1d_296	16	CIITA	ENSG00000179583	1	11001694	T	C	CTGGGAGCCCTACTCGGGCCATCGGCGGCTGCCTCGG	TGGACAGGAAGCAGAAGGTGCTTGCGAGGTACCTGAA
jtt1d_297	16	CIITA	ENSG00000179583	1	11001743	G	A	AGAAGGTGCTTGCGAGGTACCTGAAGCGGCTGCAGCC	GGGGACACTGCGGGCGCGGCAGCTGCTGGAGCTGCTG
jtt1d_298	16	CIITA	ENSG00000179583	1	11001770	G	T	GGCTGCAGCCGGGGACACTGCGGGCGCGGCAGCTGCT	GGAGCTGCTGCACTGCGCCCACGAGGCCGAGGAGGCT
jtt1d_299	16	CIITA	ENSG00000179583	1	11001821	C	T	GCGCCCACGAGGCCGAGGAGGCTGGAATTTGGCAGCA	CGTGGTACAGGAGCTCCCCGGCCGCCTCTCTTTTCTG
jtt1d_300	16	CIITA	ENSG00000179583	1	11001914	G	A	CTGATGCACATGTACTGGGCAAGGCCTTGGAGGCGGC	GGGCCAAGACTTCTCCCTGGACCTCCGCAGCACTGGC
jtt1d_301	16	CIITA	ENSG00000179583	1	11002904	G	A	TGAGGCCCTCCCTCCACAGGGCTGCCTTGAGCGACAC	GGTGGCGCTGTGGGAGTCCCTGCAGCAGCATGGGGAG
jtt1d_302	16	CIITA	ENSG00000179583	1	11002927	A	G	GCCTTGAGCGACACGGTGGCGCTGTGGGAGTCCCTGC	AGCAGCATGGGGAGACCAAGCTACTTCAGGCAGCAGA
jtt1d_303	16	CIITA	ENSG00000179583	1	11016045	C	T	CCTCTGTTTCCGACAGCTTGTACAATAACTGCATCTG	CGACGTGGGAGCCGAGAGCTTGGCTCGTGTGCTTCCG
jtt1d_304	16	CIITA	ENSG00000179583	1	11016265	G	C	CAAGGGCCAGGCCCCAAGGTGAGTTTCTCTTGCCAGC	GTCCAGTACAACAAGTTCACGGCTGCCGGGGCCCAGC
jtt1d_305	16	CIITA	ENSG00000179583	1	11017815	T	C	TACTTGTGGACACAGCTCTTCTCCAGGCTGTATCCCA	TGAGCCTCAGCATCCTGGCACCCGGCCCCTGCTGGTT
jtt1d_306	16	CIITA	ENSG00000179583	1	11017869	C	T	GCACCCGGCCCCTGCTGGTTCAGGGTTGGCCCCTGCC	CGGCTGCGGAATGAACCACATCTTGCTCTGCTGACAG
jtt1d_307	16	CIITA	ENSG00000179583	1	11017870	G	A	CACCCGGCCCCTGCTGGTTCAGGGTTGGCCCCTGCCC	GGCTGCGGAATGAACCACATCTTGCTCTGCTGACAGA
jtt1d_308	16	CIITA	ENSG00000179583	1	11017973	C	T	CCCAGTTGGGTGGATGCCTGGTGGCAGCTGCGGTCCA	CCCAGGAGCCCCGAGGCCTTCTCTGAAGGACATTGCG
jtt1d_309	16	CIITA	ENSG00000179583	1	11018402	C	T	CAAGCGTGAGCCACTGCACCGGGCCACAGAGAAAGTA	CTTCTCCACCCTGCTCTCCGACCAGACACCTTGACAG
jtt1d_310	16	CIITA	ENSG00000179583	1	11018447	G	A	CCCTGCTCTCCGACCAGACACCTTGACAGGGCACACC	GGGCACTCAGAAGACACTGATGGGCAACCCCCAGCCT
jtt1d_311	16	CIITA	ENSG00000179583	1	11018622	T	C	GGCCAGATGCACCAGCCCTTAGCAGGGAAACAGCTAA	TGGGACACTAATGGGGCGGTGAGAGGGGAACAGACTG
jtt1d_312	16	CIITA	ENSG00000179583	1	11018623	G	A	GCCAGATGCACCAGCCCTTAGCAGGGAAACAGCTAAT	GGGACACTAATGGGGCGGTGAGAGGGGAACAGACTGG
jtt1d_313	16	CIITA	ENSG00000179583	1	11023208	T	C	TGCAGGGAGGCAAACTCTGGCTGGGTTCCTGTAAACA	TCCATCGCAGCTGCAAATAATCAGAAGCCAAGGCCAG
jtt1d_313	16	DEXI	ENSG00000182108	-1	11023208	T	C	TGCAGGGAGGCAAACTCTGGCTGGGTTCCTGTAAACA	TCCATCGCAGCTGCAAATAATCAGAAGCCAAGGCCAG
jtt1d_314	16	CIITA	ENSG00000179583	1	11023406	G	T	TCAAACAGGAACCTCTCTGTTGGCACGAAGCTTTTGA	GGGGAGCAGGTCTAACAAGAAGGAAAAAGGGGGGTTA
jtt1d_314	16	DEXI	ENSG00000182108	-1	11023406	G	T	TCAAACAGGAACCTCTCTGTTGGCACGAAGCTTTTGA	GGGGAGCAGGTCTAACAAGAAGGAAAAAGGGGGGTTA
jtt1d_315	16	CLEC16A	ENSG00000038532	1	11038360	C	T	GCCGCCGAAGGCCACGCGGGTGAACTGCATTTCCCAG	CGCCCCACGCGGCGGCGGCCGTAAAGCGCGGCGGTCG
jtt1d_316	16	CLEC16A	ENSG00000038532	1	11038464	T	C	GGGCTGTGGGCCGGGGAGGAAGGCGGCTCGCGGTTCC	TCCACCGCCTCCGCCGCCGCATCCTCCGCTTGTGCTA
jtt1d_317	16	CLEC16A	ENSG00000038532	1	11038467	A	C	CTGTGGGCCGGGGAGGAAGGCGGCTCGCGGTTCCTCC	ACCGCCTCCGCCGCCGCATCCTCCGCTTGTGCTACCG
jtt1d_318	16	CLEC16A	ENSG00000038532	1	11038558	G	T	CTCTGCTGGTCCGGCATGAGACCGTGAGACGAGAGAC	GGGTCGGGGCCGCCGACATGTTTGGCCGCTCGCGGAG
jtt1d_319	16	CLEC16A	ENSG00000038532	1	11056378	C	T	TCTTCTTGAACATCTTGCGGCAAAAGTCGGGCCGTTA	CGTGTGCGTTCAGCTGCTGCAGACCTTGAACATCCTC

jtt1d_320	16	CLEC16A	ENSG00000038532	1	11056426	C	A	AGCTGCTGCAGACCTTGAACATCCTCTTTGAGAACAT	CAGTCACGAGACCTCACTTTGTAAGGACATTCCTTGG
jtt1d_321	16	CLEC16A	ENSG00000038532	1	11073195	C	T	CTCTCTCTCTCCTGCCACCCTGCACTAGGGAGGAGAA	CGGCCGAAAATTAGCCTGCCGGTGTCTCTTTATCTTC
jtt1d_322	16	CLEC16A	ENSG00000038532	1	11076776	A	G	TCTTAATTATACATCATGCACCGCTGGTGAACTCGTT	AGCTGAAGTCATTCTGAATGGTGATCTGTCTGAGATG
jtt1d_323	16	CLEC16A	ENSG00000038532	1	11154770	G	A	CTCTGCTCTCTGAACTGTTGGTCCAGGCCATCCGGGT	GTTCTTCATGCTGCGTTCCCTGTCACTGCAATTGCGA
jtt1d_324	16	CLEC16A	ENSG00000038532	1	11220123	G	A	GCTGCCCTTCCTCTCTCAGAAGCCCCGTCGGCTGGCA	GCACCAGCTTCTTAGAATTTGTCAAAGCACAGCGCAA
jtt1d_325	16	CLEC16A	ENSG00000038532	1	11260274	C	T	TCTTGCAGGCTTCGCCGTGGCCCAGTGCATAAACCAG	CACAGCTCCCCGTCCCTGTCCTCACAGTCGCCACCCT
jtt1d_326	16	CLEC16A	ENSG00000038532	1	11260278	G	A	GCAGGCTTCGCCGTGGCCCAGTGCATAAACCAGCACA	GCTCCCCGTCCCTGTCCTCACAGTCGCCACCCTCCGC
jtt1d_327	16	CLEC16A	ENSG00000038532	1	11272287	G	A	CAACGAAACGGAAGCAGACTCTAAGCCCAGCAAGAAC	GTGGCCAGGAGCGCAGCCGTGGAGACAGCCAGCCTGT
jtt1d_328	16	CLEC16A	ENSG00000038532	1	11272330	G	A	AGGAGCGCAGCCGTGGAGACAGCCAGCCTGTCCCCCA	GCCTCGTCCCTGCCCGGCAGCCCACCATTTCCCTGCT
jtt1d_329	16	CLEC16A	ENSG00000038532	1	11272572	G	A	CGCTGAGGACTGAGTCAGTGCCGGGGCCTCCCTTTGT	GTGTGTGGCCCCGCTGGTAGGGACCCCAGTGCCGCTG
jtt1d_330	16	CLEC16A	ENSG00000038532	1	11272740	C	T	CCCCACGTTGTCCTTGAATTCCTTTTTCACTTTGCAT	CTCTTCACGTGCAGGCTGGGACCAGCGGAGACACCGC
jtt1d_331	16	CLEC16A	ENSG00000038532	1	11273405	C	A	TTTCTCCAGGAAAAGGAGGAATGTAGCCAGCTCCCCA	CTCAGGACGCTTCCTCATTTCTCTTCACCAAAACCAA
jtt1d_332	16	CLEC16A	ENSG00000038532	1	11273459	C	T	TTTCTCTTCACCAAAACCAAACAGAGACAGCTTCCAG	CACCTTCTTCAGTGTTACCATCTCTAAGAAGGAACCA
jtt1d_333	16	CLEC16A	ENSG00000038532	1	11274064	C	G	AGAACATGGTCTCTGTCTCCTCGGCCCAGCCAGCTGT	CCCGGCAAGGCCTGCCGAGGGCAGTTTTCAACCTCAT
jtt1d_334	16	CLEC16A	ENSG00000038532	1	11274079	C	T	TCTCCTCGGCCCAGCCAGCTGTCCCGGCAAGGCCTGC	CGAGGGCAGTTTTCAACCTCATGAAGGAAACACAGTC
jtt1d_335	16	CLEC16A	ENSG00000038532	1	11274456	A	G	CCTGTGTGTTGCTTAATTTTTTAAGAGCAAGAGGGGT	AGAGAGGATCAAGCTGGCCCTGGCTGGAGATGGCTAG
jtt1d_336	16	CLEC16A	ENSG00000038532	1	11274485	A	C	AGAGGGGTAGAGAGGATCAAGCTGGCCCTGGCTGGAG	ATGGCTAGCCCCTGAGACATGCACTTCTGGTTTTGAA
jtt1d_337	16	CLEC16A	ENSG00000038532	1	11274748	T	C	GGGCTGGACAGCATGCCCGGGAGGACCAGCAGAGGAT	TAAAGGTGACTGGGAGGACCAGCGGAGGATAAAAGAC
jtt1d_338	16	CLEC16A	ENSG00000038532	1	11275128	C	T	CTCATAGCTGGGGCGCTCCCAGACAGGCCAGTCCAGA	CAGGACACGCTGGGCCCCTGGCATCCAGAGGAAGAGC
jtt1d_339	16	CLEC16A	ENSG00000038532	1	11275672	G	A	CCTAAGGGGCAGGTGAAGAAGCGCAGCCCTGCCAGAC	GCGCTAGATTCCTCTAAGGTCTCTGAGATGCACCGTT
jtt1d_340	16	CLEC16A	ENSG00000038532	1	11275720	C	T	CTCTAAGGTCTCTGAGATGCACCGTTTTTTAAAAAGG	CGTGGGGTGAACTGATTTTGATCTTCTTGTCTAGATG
jtt1d_341	16	CLEC16A	ENSG00000038532	1	11275881	T	G	TATGTAAATAATTTTTGTCCCAGTGAGAACCGAGGGT	TAGAAAACCTCGATGCCTCTGAGCCTCGGGACCGCTC
jtt1d_342	16	CLEC16A	ENSG00000038532	1	11275913	C	T	AGGGTTAGAAAACCTCGATGCCTCTGAGCCTCGGGAC	CGCTCTAGGGAAGTACCTGCTTTCGCCAGCATGACTC
jtt1d_343	18	DOK6	ENSG00000206052	1	67508495	C	T	GCTCTCCTTTCTTCCTCTCTAGGTCATGGGTTTGGTT	CGTCAAAGATGTCTCGTGCACAGACATTTCCCAGCTA
jtt1d_344	18	DOK6	ENSG00000206052	1	67508929	T	C	GGTCTGACAGTAACAGAAACCTCAGCACTGGGAAAAG	TTGCCCACACTGGGGTATGCCTGGGTGATGGGCACCT
jtt1d_345	18	DOK6	ENSG00000206052	1	67509073	A	G	CACAAGCTCTGTGGCTTTTAAAGTTCTGACAGGGATA	AATACAGTAAGCTCTGCAAATGACATCGTAGCTGCAT
jtt1d_346	18	DOK6	ENSG00000206052	1	67509199	C	T	AAGCAGGGCCCTGTAGCTCTACTCGTGTGTGTGTGTG	CGTGTGTGTGTGTGTGTGTAGACAAATGGATATTGCT
jtt1d_347	18	DOK6	ENSG00000206052	1	67509341	T	C	TTGTTGAAAACCATAATTTGGTGCATTAAAGTTAAGA	TTGTTATGTTGAATAGCTATTTTTAAAATAGTGCTGT
jtt1d_348	18	CD226	ENSG00000150637	-1	67530195	C	T	GAATGATCACTATATTTAACAATAACCATTGTCCTTT	CAAGGTAACCATGACAGTTTTATTCAGTTTGGCAAAT
jtt1d_349	18	CD226	ENSG00000150637	-1	67530439	G	A	CCAATTTCTTTTTTCCCTCCCAAATTTCTACCCTACC	GTCCTATGCCACCACTTTCAACTGTCACTTTTAGTGG

jtt1d_350	18	CD226	ENSG00000150637	-1	67530796	G	A	GGTATTCTTCTGGACTATGCTGATAGAGTGGATTCTA	GAAGTATGGACAGAGATACCCCATTTCTAGAATCCAT
jtt1d_351	18	CD226	ENSG00000150637	-1	67531026	C	A	AGTGAACCCTTAGTTTACTTAAGCCATTCCATCCTTT	CTGGTCATGTTGCCTATTTAACAATGAAAAAATAATC
jtt1d_352	18	CD226	ENSG00000150637	-1	67531642	T	C	TCTCTTGTATCATCCATGGATTGATTGGTAGGTTGAC	TGGTAGAGATGGGACTTCTATAGTTATTGGGTGCCTA
jtt1d_353	18	CD226	ENSG00000150637	-1	67534632	C	T	TCTGTGTATCCCAGGACTCTGTAAATAGATCTCTTCT	CTCTCTCCTTCTCCTTCTGGAATGCATATTCAATAAA
jtt1d_354	18	CD226	ENSG00000150637	-1	67534642	C	T	CCAGGACTCTGTAAATAGATCTCTTCTCTCTCTCCTT	CTCCTTCTGGAATGCATATTCAATAAAGGATATAAAG
jtt1d_355	18	CD226	ENSG00000150637	-1	67563156	G	T	ACCAAGTTGCAGTAAGTTAAGAGGTCGATCTGACGGG	GCTGGATCTTTTCCCACCTCACTGCCTGCACAGGCCA

Appendix III

Substitution scores of the 100 nsSNPs

ID	BLOSUM	PAM	ESST
jtt1d_1	0	1	N/A
jtt1d_5	0	0	1
jtt1d_6	-1	0	1
jtt1d_10	0	0	1
jtt1d_11	0	1	1
jtt1d_13	3	3	4
jtt1d_14	0	0	N/A
jtt1d_15	1	-2	N/A
jtt1d_16	0	0	N/A
jtt1d_17	-1	-1	N/A
jtt1d_19	0	0	N/A
jtt1d_21	0	0	2
jtt1d_22	-3	-3	-5
jtt1d_23	1	3	N/A
jtt1d_25	1	1	2.8571
jtt1d_27	2	2	N/A
jtt1d_28	0	0	N/A
jtt1d_31	-2	-6	N/A
jtt1d_32	0	0	N/A
jtt1d_35	-3	-3	N/A
jtt1d_36	0	1	N/A
jtt1d_43	-3	-5	N/A
jtt1d_51	0	0	N/A
jtt1d_53	0	-1	N/A
jtt1d_54	-1	-1	N/A
jtt1d_55	1	1	N/A
jtt1d_56	-1	0	N/A
jtt1d_57	2	2	N/A
jtt1d_59	-1	0	N/A
jtt1d_60	-2	-2	N/A
jtt1d_62	-3	-5	N/A
jtt1d_64	2	3	N/A
jtt1d_69	-3	0	N/A
jtt1d_70	-1	-1	N/A
jtt1d_71	-1	-1	N/A
jtt1d_74	-3	-3	N/A
jtt1d_78	-3	-5	N/A
jtt1d_79	0	1	N/A
jtt1d_83	0	1	N/A

jtt1d_89	-2	-2	N/A
jtt1d_105	1	1	N/A
jtt1d_106	-1	-2	N/A
jtt1d_107	-1	-2	N/A
jtt1d_155	-1	0	N/A
jtt1d_156	-2	-5	N/A
jtt1d_158	-1	-1	N/A
jtt1d_161	0	0	N/A
jtt1d_162	1	0	N/A
jtt1d_171	1	0	0
jtt1d_173	1	1	2.25
jtt1d_176	-1	0	N/A
jtt1d_178	-1	-1	N/A
jtt1d_179	-1	-1	N/A
jtt1d_180	2	1	N/A
jtt1d_183	-2	-2	-4
jtt1d_185	2	2	N/A
jtt1d_187	-2	-4	N/A
jtt1d_191	-3	-5	N/A
jtt1d_192	-3	-3	N/A
jtt1d_195	-1	0	-1
jtt1d_197	1	0	N/A
jtt1d_202	1	0	N/A
jtt1d_219	-2	-4	N/A
jtt1d_223	1	2	N/A
jtt1d_225	-3	-5	N/A
jtt1d_239	1	2	N/A
jtt1d_240	-2	-11	N/A
jtt1d_241	-1	0	N/A
jtt1d_242	-1	0	N/A
jtt1d_243	-1	-2	N/A
jtt1d_247	0	0	N/A
jtt1d_256	-1	-1	N/A
jtt1d_260	1	0	N/A
jtt1d_265	1	1	N/A
jtt1d_268	2	1	N/A
jtt1d_275	0	2	0
jtt1d_276	1	1	N/A
jtt1d_277	1	1	N/A
jtt1d_278	3	3	N/A
jtt1d_283	-1	-1	N/A
jtt1d_285	2	3	N/A
jtt1d_286	-1	-1	N/A
jtt1d_287	-1	-2	N/A
jtt1d_288	-2	-6	N/A
jtt1d_289	-2	-2	N/A
jtt1d_290	0	0	N/A

jtt1d_292	-2	-1	N/A
jtt1d_295	-2	-6	N/A
jtt1d_296	0	-1	N/A
jtt1d_302	1	0	N/A
jtt1d_304	1	0	N/A
jtt1d_321	-3	0	N/A
jtt1d_325	2	-1	N/A
jtt1d_326	1	1	N/A
jtt1d_327	1	0	N/A
jtt1d_328	1	1	N/A
jtt1d_343	-2	-6	N/A
jtt1d_352	0	0	N/A
jtt1d_354	2	2	N/A
jtt1d_355	-1	-2	N/A

References

1. Bajaj M, Blundell T (1984) Evolution and the tertiary structure of proteins. Annual Review of Biophysics and Bioengineering 13: 453.
2. Orengo CA, Thornton JM (2005) Protein families and their evolution-a structural perspective. Annual Review of Biochemistry 74: 867.
3. Sanger F, Tuppy H (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem J 49: 481-490.
4. Sanger F, Tuppy H (1951) The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem J 49: 463-481.
5. Sanger F (1988) Sequences, Sequences, and Sequences. Annual Review of Biochemistry 57: 1-29.
6. Adams MJ, Blundell TL, Dodson EJ, Dodson GG, Vijayan M, et al. (1969) Structure of Rhombohedral 2 Zinc Insulin Crystals. Nature 224: 491-495.
7. Blundell TL, Cutfield JF, Cutfield SM, Dodson EJ, Dodson GG, et al. (1971) Atomic positions in rhombohedral 2-zinc insulin crystals. Nature 231: 506-511.
8. Blundell TL, Cutfield JF, Cutfield SM, Dodson EJ, Dodson GG, et al. (1972) Three-dimensional atomic structure of insulin and its relationship to activity. Diabetes 21: 492-505.
9. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624-626.
10. Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246: 96-98.
11. Zuckerkandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. J Mol Evol 7: 167-183.
12. Zuckerkandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. II. A selectionist model for random fixations in proteins. J Mol Evol 7: 269-311.
13. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296: 750-752.

14. Bloom JD, Adami C (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol Biol 3: 21.
15. Jordan IK, Wolf YI, Koonin EV (2003) No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3: 1.
16. Hubbard TJ, Blundell TL (1987) Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. Protein Eng 1: 159-171.
17. Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Journal of Molecular Biology 120: 97.
18. Gibrat JF, Garnier J, Robson B (1987) Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. Journal of Molecular Biology 198: 425.
19. Levin JM, Robson B, Garnier J (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS Letters 205: 303.
20. Pauling L, Corey RB (1951) Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. Proc Natl Acad Sci U S A 37: 729-740.
21. Pauling L, Corey RB, Branson HR (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 37: 205-211.
22. Hutchinson EG, Thornton JM (1994) A revised set of potentials for beta-turn formation in proteins. Protein Sci 3: 2207-2216.
23. Sibanda BL, Blundell TL, Thornton JM (1989) Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. J Mol Biol 206: 759-777.
24. Wilmot CM, Thornton JM (1988) Analysis and prediction of the different types of beta-turn in proteins. J Mol Biol 203: 221-232.

25. Baker EN, Hubbard RE (1984) Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 44: 97-179.
26. Presta LG, Rose GD (1988) Helix signals in proteins. Science 240: 1632-1641.
27. Richardson JS, Richardson DC (1988) Amino acid preferences for specific locations at the ends of alpha helices. Science 240: 1648-1652.
28. Jansen GA, Ferdinandusse S, Hogenhout EM, Verhoeven NM, Jakobs C, et al. (1999) Phytanoyl-CoA hydroxylase deficiency. Enzymological and molecular basis of classical Refsum disease. Adv Exp Med Biol 466: 371-376.
29. Chan AWE, Hutchinson EG, Thornton JM (1993) Identification, classification, and analysis of beta-bulges in proteins. Protein Sci 2: 1574-1590.
30. Richardson JS, Getzoff ED, Richardson DC (1978) The beta bulge: a common small unit of nonrepetitive protein structure. Proc Natl Acad Sci U S A 75: 2574-2578.
31. Barlow DJ, Thornton JM (1988) Helix geometry in proteins. J Mol Biol 201: 601-619.
32. Eswar N, Ramakrishnan C (1999) Secondary structures without backbone: an analysis of backbone mimicry by polar side chains in protein structures. Protein Eng 12: 447-455.
33. Cubellis MV, Caillez F, Blundell TL, Lovell SC (2005) Properties of polyproline II, a secondary structure element implicated in protein-protein interactions. Proteins 58: 880-892.
34. Stapley BJ, Creamer TP (1999) A survey of left-handed polyproline II helices. Protein Sci 8: 587-595.
35. Milner-White E, Ross BM, Ismail R, Belhadj-Mostefa K, Poet R (1988) One type of gamma-turn, rather than the other gives rise to chain-reversal in proteins. J Mol Biol 204: 777-782.
36. Milner-White EJ (1987) Beta-bulges within loops as recurring features of protein structure. Biochim Biophys Acta 911: 261-265.
37. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.
38. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. (1997) CATH--a hierarchic classification of protein domain structures. Structure 5: 1093-1108.

39. Mizuguchi K, Deane CM, Blundell TL, Overington JP (1998) HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 7: 2469-2471.
40. Bhaduri A, Pugalenthi G, Sowdhamini R (2004) PASS2: an automated database of protein alignments organised as structural superfamilies. BMC Bioinformatics 5: 35.
41. Bickerton GR (2009) Molecular Characterization and Evolutionary Plasticity of Protein-Protein Interfaces. Cambridge: Emmanuel College, University of Cambridge. 264 p.
42. Holm L, Sander C (1996) The FSSP database: fold classification based on structure-structure alignment of proteins. Nucleic Acids Research 24: 206.
43. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11: 739.
44. Marchler-Bauer A, Addess KJ, Chappey C, Geer L, Madej T, et al. (1999) MMDB: Entrez's 3D structure database. Nucleic Acids Res 27: 240-243.
45. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucl Acids Res 36: D281-288.
46. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2009) InterPro: the integrative protein signature database. Nucl Acids Res 37: D211-215.
47. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, et al. (2006) The PROSITE database. Nucleic Acids Res 34: D227-230.
48. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, et al. (2003) PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Research 31: 400.
49. Servant F, Bru C, Carrère S, Courcelle E, Gouzy J, et al. (2002) ProDom: automated clustering of homologous domains. Briefings in Bioinformatics 3: 246.
50. Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. Proceedings of the National Academy of Sciences of the United States of America 95: 5857.
51. Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Research 31: 371.

52. Buchan DW, Rison SC, Bray JE, Lee D, Pearl F, et al. (2003) Gene3D: structural assignments for the biologist and bioinformaticist alike. Nucleic Acids Res 31: 469-473.

53. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, et al. (2009) SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Research 37: D380-386.

54. Krishnamurthy N, Brown D, Kirshner D, Sjolander K (2006) PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. Genome Biology 7: R83.

55. Wang Y, Addess KJ, Chen J, Geer LY, He J, et al. (2007) MMDB: annotating protein sequences with Entrez's 3D-structure database. Nucleic Acids Res 35: D298-300.

56. Heger A, Korpelainen E, Hupponen T, Mattila K, Ollikainen V, et al. (2008) PairsDB atlas of protein sequence space. Nucl Acids Res 36: D276-280.

57. Orengo CA, Stilltoe I, Reeves G, Pearl FMG (2001) What can structural classifications reveal about protein evolution? J Struc Biol 134: 145-165.

58. Orengo CA, Taylor WR (1993) A local alignment method for protein structure motifs. Journal of Molecular Biology 233: 488.

59. Sali A, Blundell TL (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J Mol Biol 212: 403-428.

60. Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP (1998) JOY: protein sequence-structure representation and analysis. Bioinformatics 14: 617-623.

61. Madej T, Gibrat JF, Bryant SH (1995) Threading a database of protein cores. Proteins 23: 356.

62. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872-876.

63. Eddy SR, Mitchison G, Durbin R (1995) Maximum discrimination hidden Markov models of sequence consensus. J Comput Biol 2: 9-23.

64. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
65. (2008) The universal protein resource (UniProt). Nucleic Acids Res 36: D190-195.
66. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313: 903-919.
67. Krishnamurthy N, Brown D, Sjolander K (2007) FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. BMC Evol Biol 7 Suppl 1: S12.
68. Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Research 32: W327-331.
69. Blundell TL, Wood SP (1975) Is the evolution of insulin Darwinian or due to selectively neutral mutation? Nature 257: 197.
70. Akashi H (2001) Gene expression and molecular evolution. Curr Opin Genet Dev 11: 660-666.
71. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA 102: 14338-14343.
72. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23: 327-337.
73. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341-352.
74. Hamill SJ, Cota E, Chothia C, Clarke J (2000) Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. J Mol Biol 295: 641-649.
75. Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 75: 333-366.
76. Hamada D, Tanaka T, Tartaglia GG, Pawar A, Vendruscolo M, et al. (2009) Competition between folding, native-state dimerisation and amyloid aggregation in beta-lactoglobulin. J Mol Biol 386: 878-890.
77. Goldberg AL (2003) Protein degradation and protection against misfolded or damaged proteins. Nature 426: 895-899.

78. Wolffe AP, Matzke MA (1999) Epigenetics: regulation through repression. Science 286: 481-486.

79. Pal C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nat Rev Genet 7: 337-348.

80. Dayhoff MO, Eck RV (1968) Atlas of Protein Sequence and Structure. Natl Biomed Res Found 3: 33.

81. Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185: 862-864.

82. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89: 10915-10919.

83. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275-282.

84. Gonnet GH, Cohen MA, Benner SA (1992) Exhaustive matching of the entire protein sequence database. Science 256: 1443-1445.

85. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691-699.

86. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25: 1307-1320.

87. Luthy R, McLachlan AD, Eisenberg D (1991) Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins 10: 229-239.

88. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci 1: 216-226.

89. Overington J, Johnson MS, Sali A, Blundell TL (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc Biol Sci 241: 132-145.

90. Koshi JM, Goldstein RA (1995) Context-dependent optimal substitution matrices. Protein Eng 8: 641-645.

91. Koehl P, Levitt M (2002) Protein topology and stability define the space of allowed sequences. Proc Natl Acad Sci USA 99: 1280-1285.

92. Rice DW, Eisenberg D (1997) A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 267: 1026-1038.
93. Lee S, Blundell TL (2009) Ulla: a program for calculating environment-specific amino acid substitution tables. Bioinformatics 25: 1976-1977.
94. DeLano WL (2002) The PyMOL Molecular Graphics System. Palo Alto, CA, USA: DeLano Scientific.
95. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7: 95-99.
96. Lovell SC, Davis IW, Arendall WB, 3rd, de Bakker PI, Word JM, et al. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 50: 437-450.
97. Worth CL, Blundell TL (2009) Satisfaction of hydrogen-bonding potential influences the conservation of polar sidechains. Proteins 75: 413-429.
98. Wako H, Blundell TL (1994) Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. J Mol Biol 238: 693-708.
99. Johnson MS, Overington JP, Blundell TL (1993) Alignment and searching for common protein folds using a data bank of structural templates. J Mol Biol 231: 735-752.
100. Chelliah V, Chen L, Blundell TL, Lovell SC (2004) Distinguishing structural and functional restraints in evolution in order to identify interaction sites. J Mol Biol 342: 1487-1504.
101. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231-238.
102. Sunyaev S, Hanke J, Aydin A, Wirkner U, Zastrow I, et al. (1999) Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes. J Mol Med 77: 754-760.

103. Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314-331.
104. Solomon E, Bodmer WF (1979) Evolution of sickle variant gene. Lancet 1: 923.
105. Kan YW, Dozy AM (1978) Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. Proc Natl Acad Sci U S A 75: 5631-5635.
106. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, et al. (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 13: 399-408.
107. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, et al. (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30: 233-237.
108. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl: 228-237.
109. Kruglyak L (2008) The road to genome-wide association studies. Nat Rev Genet 9: 314-318.
110. Sunyaev S, Ramensky V, Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16: 198-200.
111. Wang Z, Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17: 263-270.
112. 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.
113. 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. Journal of Molecular Biology 315: 771-786.
114. Steward RE, MacArthur MW, Laskowski RA, Thornton JM (2003) Molecular basis of inherited diseases: a structural perspective. Trends Genet 19: 505-513.
115. Worth CL, Bickerton GR, Schreyer A, Forman JR, Cheng TM, et al. (2007) A structural bioinformatics approach to the analysis of nonsynonymous single

nucleotide polymorphisms (nsSNPs) and their relation to disease. J Bioinform Comput Biol 5: 1297-1318.

116. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11: 863-874.

117. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306: 234-238.

118. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, et al. (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245: 1073-1080.

119. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245: 1066-1073.

120. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, et al. (2010) Signatures of mutation and selection in the cancer genome. Nature 463: 893-898.

121. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Rev Genet 7: 98-108.

122. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463: 191-196.

123. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9: 387-402.

124. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.

125. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241-251.

126. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356-369.

127. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, et al. (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21: 577-581.

128. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.

129. Brookes AJ, Lehvaslaiho H, Siegfried M, Boehm JG, Yuan YP, et al. (2000) HGBASE: a database of SNPs and other variations in and around human genes. Nucleic Acids Res 28: 356-360.

130. Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, et al. (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucleic Acids Res 30: 387-391.

131. Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, et al. (1999) ProTherm: Thermodynamic Database for Proteins and Mutants. Nucleic Acids Res 27: 286-288.

132. Thorn KS, Bogan AA (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 17: 284-285.

133. Martin AC, Facchiano AM, Cuff AL, Hernandez-Boussard T, Olivier M, et al. (2002) Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. Hum Mutat 19: 149-164.

134. Kwok CJ, Martin AC, Au SW, Lam VM (2002) G6PDdb, an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations. Hum Mutat 19: 217-224.

135. Mooney SD, Altman RB (2003) MutDB: annotating human variation with functionally relevant data. Bioinformatics 19: 1858-1860.

136. Riva A, Kohane IS (2002) SNPper: retrieval and analysis of human SNPs. Bioinformatics 18: 1681-1685.

137. Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, et al. (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum Mutat 23: 464-470.

138. Mottaz A, David FP, Veuthey AL, Yip YL (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics 26: 851-852.

139. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, et al. (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 91: 355-358.

140. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, et al. (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet Chapter 10: Unit 10 11.

141. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 32: D520-522.

142. Ryan M, Diekhans M, Lien S, Liu Y, Karchin R (2009) LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. Bioinformatics 25: 1431-1432.

143. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, et al. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21: 2814-2820.

144. Hurst JM, McMillan LE, Porter CT, Allen J, Fakorede A, et al. (2009) The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Hum Mutat 30: 616-624.

145. Reumers J, Conde L, Medina I, Maurer-Stroh S, Van Durme J, et al. (2008) Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases. Nucleic Acids Res 36: D825-829.

146. Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F (2006) SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. Bioinformatics 22: 2183-2185.

147. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, et al. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res 33: D527-532.

148. Han A, Kang HJ, Cho Y, Lee S, Kim YJ, et al. (2006) SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences. Nucleic Acids Res 34: W642-644.

149. Hulbert EM, Smink LJ, Adlem EC, Allen JE, Burdick DB, et al. (2007) T1DBase: integration and presentation of complex data for type 1 diabetes research. Nucleic Acids Res 35: D742-746.

150. Jegga AG, Gowrisankar S, Chen J, Aronow BJ (2007) PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. Nucleic Acids Res 35: D700-706.
151. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.
152. Church DM, Lappalainen I, Sneddon TP, Hinton J, Maguire M, et al. (2010) Public data archives for genomic structural variation. Nat Genet 42: 813-814.
153. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
154. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, et al. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464: 713-720.
155. Bardelli A, Parsons DW, Silliman N, Ptak J, Szabo S, et al. (2003) Mutational analysis of the tyrosine kinome in colorectal cancers. Science 300: 949.
156. Davies H, Hunter C, Smith R, Stephens P, Greenman C, et al. (2005) Somatic mutations of the protein kinase gene family in human lung cancer. Cancer Res 65: 7591-7595.
157. Awan A, Bari H, Yan F, Moksong S, Yang S, et al. (2007) Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network. IET Syst Biol 1: 292-297.
158. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446: 153-158.
159. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, et al. (2007) The genomic landscapes of human breast and colorectal cancers. Science 318: 1108-1113.
160. Burke DF, Worth CL, Priego EM, Cheng T, Smink LJ, et al. (2007) Genome bioinformatic analysis of nonsynonymous SNPs. BMC Bioinformatics 8: 301.
161. Gilis D, Rooman M (2000) PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. Protein Eng 13: 849-856.
162. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714-2726.

163. Brooks BR, Brooks CL, 3rd, Mackerell AD, Jr., Nilsson L, Petrella RJ, et al. (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30: 1545-1614.

164. Christen M, Hunenberger PH, Bakowies D, Baron R, Burgi R, et al. (2005) The GROMOS software for biomolecular simulation: GROMOS05. J Comput Chem 26: 1719-1751.

165. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320: 369-387.

166. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, et al. (2005) The FoldX web server: an online force field. Nucleic Acids Res 33: W382-388.

167. Cheng J, Randall A, Baldi P (2006) Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62: 1125-1132.

168. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33: W306-310.

169. Capriotti E, Fariselli P, Casadio R (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 20 Suppl 1: i63-68.

170. Masso M, Vaisman, II (2008) Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics 24: 2002-2009.

171. Dell'Orco D (2009) Fast predictions of thermodynamics and kinetics of protein-protein recognition from structures: from molecular design to systems biology. Mol Biosyst 5: 323-334.

172. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 31: 19-20.

173. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894-3900.

174. Topham CM, Srinivasan N, Blundell TL (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng 10: 7-21.

175. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, et al. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics.

176. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812-3814.

177. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249.

178. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129-2141.

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

180. Bao L, Zhou M, Cui Y (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33: W480-482.

181. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, et al. (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21: 3176-3178.

182. Parthiban V, Gromiha MM, Schomburg D (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 34: W239-242.

183. Yuan HY, Chiou JJ, Tseng WH, Liu CH, Liu CK, et al. (2006) FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. Nucleic Acids Res 34: W635-641.

184. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7: 166.

185. Yin S, Ding F, Dokholyan NV (2007) Eris: an automated estimator of protein stability. Nat Methods 4: 466-467.

186. Ye ZQ, Zhao SQ, Gao G, Liu XQ, Langlois RE, et al. (2007) Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics 23: 1444-1450.

187. Uzun A, Leslin CM, Abyzov A, Ilyin V (2007) Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways. Nucleic Acids Res 35: W384-392.

188. Li S, Ma L, Li H, Vang S, Hu Y, et al. (2007) Snap: an integrated SNP annotation platform. Nucleic Acids Res 35: D707-710.

189. Masso M, Vaisman, II (2010) AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng Des Sel 23: 683-687.

190. Cheng TM, Lu YE, Vendruscolo M, Lio P, Blundell TL (2008) Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. PLoS Comput Biol 4: e1000135.

191. Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, et al. (2008) Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. Hum Mutat 29: 198-204.

192. Lee PH, Shatkay H (2008) F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res 36: D820-824.

193. Gong S, Blundell TL (2008) Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. PLoS Comput Biol 4: e1000179.

194. Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8: 357-366.

195. Zuckerkandl E, Pauling LB (1962) Molecular disease, evolution, and genetic heterogeneity; Kasha M, Pullman B, editors: Academic Press. 189-225 p.

196. Blundell TL, Cooper J, Donnelly D, Driessen H, Edwards Y, et al. (1991) Patterns of sequence variation in families of homologous proteins. In: Jornvall/Hoog/Gustavsson, editor. Methods in Proteins Sequence Analysis. Basel: Birkhauser Verlag AG. pp. 373-385.

197. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310: 243-257.

198. Chelliah V, Blundell T, Mizuguchi K (2005) Functional restraints on the patterns of amino acid substitutions: application to sequence-structure homology recognition. Proteins 61: 722-731.

199. Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32: D129-133.

200. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32: D115-119.

201. Gong S, Park C, Choi H, Ko J, Jang I, et al. (2005) A protein domain interaction interface database: InterPare. BMC Bioinformatics 6: 207.

202. Lee S, Blundell TL (2009) BIPA: a database for protein-nucleic acid interaction in 3D structures. Bioinformatics 25: 1559-1560.

203. McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238: 777-793.

204. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. J Mol Biol 280: 1-9.

205. Guharoy M, Chakrabarti P (2005) Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Acad Sci U S A 102: 15447-15452.

206. Fox BA, Yee VC, Pedersen LC, Le Trong I, Bishop PD, et al. (1999) Identification of the calcium binding site and a novel ytterbium site in blood coagulation factor XIII by x-ray crystallography. J Biol Chem 274: 4917-4923.

207. Lin Y, Hwang WC, Basavappa R (2002) Structural and functional analysis of the human mitotic-specific ubiquitin-conjugating enzyme, UbcH10. J Biol Chem 277: 21913-21921.

208. Hoang C, Ferre-D'Amare AR (2001) Cocrystal structure of a tRNA Psi55 pseudouridine synthase: nucleotide flipping by an RNA-modifying enzyme. Cell 107: 929-939.

209. Stec B, Holtz KM, Kantrowitz ER (2000) A revised mechanism for the alkaline phosphatase reaction involving three metal ions. J Mol Biol 299: 1303-1311.
210. Koike A, Takagi T (2004) Prediction of protein-protein interaction sites using support vector machines. Protein Eng Des Sel 17: 165-173.
211. Sikic M, Tomic S, Vlahovicek K (2009) Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. PLoS Comput Biol 5: e1000278.
212. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257: 342-358.
213. Sali A, Overington JP, Johnson MS, Blundell TL (1990) From comparisons of protein sequences and structures to protein modelling and design. Trends Biochem Sci 15: 235-240.
214. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.
215. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.
216. (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res 37: D169-174.
217. David FP, Yip YL (2008) SSMap: a new UniProt-PDB mapping resource for the curation of structural-related information in the UniProt/Swiss-Prot Knowledgebase. BMC Bioinformatics 9: 391.
218. Laskowski RA (2009) PDBsum new things. Nucleic Acids Res 37: D355-359.
219. Martin AC (2005) Mapping PDB chains to UniProtKB entries. Bioinformatics 21: 4297-4301.
220. Reichert J, Suhnel J (2002) The IMB Jena Image Library of Biological Macromolecules: 2002 update. Nucleic Acids Res 30: 253-254.
221. Velankar S, McNeil P, Mittard-Runte V, Suarez A, Barrell D, et al. (2005) E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res 33: D262-265.
222. Via A, Zanzoni A, Helmer-Citterich M (2005) Seq2Struct: a resource for establishing sequence-structure links. Bioinformatics 21: 551-553.

223. Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6: 31.
224. Word MJ (2000) All-atom small-probe contact surface analysis: An information-rich description of molecular goodness-of-fit. Durham: Duke University.
225. Worth CL, Gong S, Blundell TL (2009) Structural and functional constraints in the evolution of protein families. Nat Rev Mol Cell Biol 10: 709-720.
226. Gong S, Worth CL, Bickerton GR, Lee S, Tanramluk D, et al. (2009) Structural and functional restraints in the evolution of protein families and superfamilies. Biochem Soc Trans 37: 727-733.
227. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. Embo J 5: 823-826.
228. Kisters-Woike B, Vangierdegom C, Mueller-Hill B (2000) On the conservation of protein sequences in evolution. Trends in Biochemical Sciences 25: 419-421.
229. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160-174.
230. Hotelling H (1933) Analysis of Complex Statistical Variables into Principal Components. J Educ Psychol 24: 417-441.
231. Michener CD, Sokal RR (1957) A Quantitative Approach to a Problem in Classification. Evolution 11: 130.
232. Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. Proc Natl Acad Sci U S A 103: 5869-5874.
233. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, et al. (2005) Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci U S A 102: 606-611.
234. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127-128.
235. Deane CM, Allen FH, Taylor R, Blundell TL (1999) Carbonyl-carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid. Protein Eng 12: 1025-1028.
236. Gallivan JP, Dougherty DA (1999) Cation-pi interactions in structural biology. Proc Natl Acad Sci U S A 96: 9459-9464.

237. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.

238. Gong S, Blundell TL (2010) Structural and functional restraints on the occurrence of single amino Acid variations in human proteins. PLoS One 5: e9186.

239. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, et al. (2009) The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. Genome Res 19: 1622-1629.

240. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The diploid genome sequence of an individual human. PLoS Biol 5: e254.

241. Wang J, Wang W, Li R, Li Y, Tian G, et al. (2008) The diploid genome sequence of an Asian individual. Nature 456: 60-65.

242. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, et al. (2008) Genetic variation in an individual human exome. PLoS Genet 4: e1000160.

243. Bao L, Cui Y (2006) Functional impacts of non-synonymous single nucleotide polymorphisms: selective constraint and structural environments. FEBS Lett 580: 1231-1234.

244. Yip YL, Famiglietti M, Gos A, Duek PD, David FP, et al. (2008) Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat 29: 361-366.

245. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. Nucleic Acids Res 37: D690-697.

246. Talavera D, Taylor MS, Thornton JM (2009) The (non)malignancy of cancerous amino acidic substitutions. Proteins.

247. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327-335.

248. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

249. Jansen GA, Hogenhout EM, Ferdinandusse S, Waterham HR, Ofman R, et al. (2000) Human phytanoyl-CoA hydroxylase: resolution of the gene structure and the molecular basis of Refsum's disease. Hum Mol Genet 9: 1195-1200.

250. Jansen GA, Ofman R, Ferdinandusse S, Ijlst L, Muijsers AO, et al. (1997) Refsum disease is caused by mutations in the phytanoyl-CoA hydroxylase gene. Nat Genet 17: 190-193.

251. Mihalik SJ, Morrell JC, Kim D, Sacksteder KA, Watkins PA, et al. (1997) Identification of PAHX, a Refsum disease gene. Nat Genet 17: 185-189.

252. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

253. Valdar WS (2002) Scoring residue conservation. Proteins 48: 227-241.

254. Prlic A, Down TA, Hubbard TJ (2005) Adding some SPICE to DAS. Bioinformatics 21 Suppl 2: ii40-41.

255. Prlic A, Down TA, Kulesha E, Finn RD, Kahari A, et al. (2007) Integrating sequence and structural biology with DAS. BMC Bioinformatics 8: 333.

256. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577-2637.

257. Ihaka R, Gentleman R (1996) R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5: 299-314.

258. Shenkin PS, Erman B, Mastrandrea LD (1991) Information-theoretical entropy as a measure of sequence variability. Proteins 11: 297-313.

259. Schuler LD, Walde P, Luisi PL, van Gunsteren WF (2001) Molecular dynamics simulation of n-dodecyl phosphate aggregate structures. Eur Biophys J 30: 330-343.

260. Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet 10: 135-151.

261. Weir BS (2008) Linkage disequilibrium and association mapping. Annu Rev Genomics Hum Genet 9: 129-142.

262. Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, et al. (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature 448: 591-594.

263. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, et al. (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39: 857-864.

264. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881-885.

265. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, et al. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316: 1336-1341.

266. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 39: 1329-1337.

267. Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, et al. (2010) Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. Nature 463: 360-363.

268. Lowe CE, Cooper JD, Brusko T, Walker NM, Smyth DJ, et al. (2007) Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. Nat Genet 39: 1074-1082.

269. Reed P, Cucca F, Jenkins S, Merriman M, Wilson A, et al. (1997) Evidence for a type 1 diabetes susceptibility locus (IDDM10) on human chromosome 10p11-q11. Hum Mol Genet 6: 1011-1016.

270. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, et al. (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12: 1599-1610.

271. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, et al. (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res 14: 2121-2127.

272. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, et al. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36: 40-45.

273. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, et al. (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 38: 617-619.

274. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

275. Chistyakov DA, Savost'anov KV, Turakulov RI, Petunina NA, Trukhina LV, et al. (2000) Complex association analysis of graves disease using a set of polymorphic markers. Mol Genet Metab 70: 214-218.

276. Marron MP, Raffel LJ, Garchon HJ, Jacob CO, Serrano-Rios M, et al. (1997) Insulin-dependent diabetes mellitus (IDDM) is associated with CTLA4 polymorphisms in multiple ethnic groups. Hum Mol Genet 6: 1275-1282.

277. Vaidya B, Imrie H, Perros P, Dickinson J, McCarthy MI, et al. (1999) Cytotoxic T lymphocyte antigen-4 (CTLA-4) gene polymorphism confers susceptibility to thyroid associated orbitopathy. Lancet 354: 743-744.

278. Deng Z, Morse JH, Slager SL, Cuervo N, Moore KJ, et al. (2000) Familial primary pulmonary hypertension (gene PPH1) is caused by mutations in the bone morphogenetic protein receptor-II gene. Am J Hum Genet 67: 737-744.

279. Thio CL, Mosbruger TL, Kaslow RA, Karp CL, Strathdee SA, et al. (2004) Cytotoxic T-lymphocyte antigen 4 gene and recovery from hepatitis B virus infection. J Virol 78: 11258-11262.

280. Auld DS (2001) Zinc coordination sphere in biochemical zinc sites. Biometals 14: 271-313.

281. Auld DS (2009) The ins and outs of biological zinc sites. Biometals 22: 141-148.

282. Yamashita MM, Wesson L, Eisenman G, Eisenberg D (1990) Where metal ions bind in proteins. Proc Natl Acad Sci U S A 87: 5648-5652.

283. Johnson JL, Coyne KE, Garrett RM, Zabot MT, Dorche C, et al. (2002) Isolated sulfite oxidase deficiency: identification of 12 novel SUOX mutations in 10 patients. Hum Mutat 20: 74.

284. Kisker C, Schindelin H, Pacheco A, Wehbi WA, Garrett RM, et al. (1997) Molecular basis of sulfite oxidase deficiency from the structure of sulfite oxidase. Cell 91: 973-983.

285. Wilson HL, Wilkinson SR, Rajagopalan KV (2006) The G473D mutation impairs dimerization and catalysis in human sulfite oxidase. Biochemistry 45: 2149-2160.

286. Yokoe S, Takahashi M, Asahi M, Lee SH, Li W, et al. (2007) The Asn418-linked N-glycan of ErbB3 plays a crucial role in preventing spontaneous heterodimerization and tumor promotion. Cancer Res 67: 1935-1942.

287. Bluyssen HA, Levy DE (1997) Stat2 is a transcriptional activator that requires sequence-specific contacts provided by stat1 and p48 for stable interaction with DNA. J Biol Chem 272: 4600-4605.

288. Wojciak JM, Martinez-Yamout MA, Dyson HJ, Wright PE (2009) Structural basis for recruitment of CBP/p300 coactivators by STAT1 and STAT2 transactivation domains. Embo J 28: 948-958.

289. Nickerson DA, Rieder MJ, Crawford DC, Carlson CS, Livingston RJ (2005) An Overview of the Environmental Genome Project. Essays on the Future of Envrionmental Health Research: 42-53.

290. Alimonti A, Carracedo A, Clohessy JG, Trotman LC, Nardella C, et al. (2010) Subtle variations in Pten dose determine cancer susceptibility. Nat Genet 42: 454-458.

291. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.

292. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195-197.

293. Schreyer A, Blundell T (2009) CREDO: a protein-ligand interaction database for drug discovery. Chem Biol Drug Des 73: 157-167.

294. Lee S, Brown A, Pitt WR, Perez Higueruelo A, Gong S, et al. (2009) Structural interactomics: informatics approaches to aid the interpretation of genetic variation and the development of novel therapeutics. Mol Biosyst.

295. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, et al. (2000) GenBank. Nucleic Acids Res 28: 15-18.

296. Higueruelo AP, Schreyer A, Bickerton GR, Pitt WR, Groom CR, et al. (2009) Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. Chem Biol Drug Des 74: 457-467.

297. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

298. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, et al. (2010) WormBase: a comprehensive resource for nematode research. Nucleic Acids Res 38: D463-467.

299. Haliloglu T, Keskin O, Ma B, Nussinov R (2005) How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hot spots and conserved residues. Biophys J 88: 1552-1559.

300. Park J, Bolser D (2001) Conservation of protein interaction network in evolution. Genome Inform 12: 135-140.

301. Batada NN, Hurst LD, Tyers M (2006) Evolutionary and physiological importance of hub proteins. PLoS Comput Biol 2: e88.

302. Pal C, Papp B, Hurst LD (2003) Genomic function: Rate of evolution and gene dispensability. Nature 421: 496-497; discussion 497-498.

303. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, et al. (2005) Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A 102: 5483-5488.

304. Andreeva A, Murzin AG (2006) Evolution of protein fold in the presence of functional constraints. Current Opinion in Structural Biology 16: 399.

305. Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE (2009) The origin, evolution and structure of the protein world. The Biochemical Journal 417: 621.

306. Copley RR, Letunic I, Bork P (2002) Genome and protein evolution in eukaryotes. Current Opinion in Chemical Biology 6: 39-45.

307. Kinch LN, Grishin NV (2002) Evolution of protein structures and functions. Current Opinion in Structural Biology 12: 400.

308. Choi JK, Kim SC, Seo J, Kim S, Bhak J (2007) Impact of transcriptional properties on essentiality and evolutionary rate. Genetics 175: 199-206.

309. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102: 14338-14343.

310. Zeldovich KB, Shakhnovich EI (2008) Understanding Protein Evolution: From Protein Physics to Darwinian Selection. Annual Review of Physical Chemistry 59: 105-127.
311. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39: 309-338.
312. Izarzugaza JM, Redfern OC, Orengo CA, Valencia A (2009) Cancer-associated mutations are preferentially distributed in protein kinase functional sites. Proteins 77: 892-903.
313. Ferrer-Costa C, Orozco M, de la Cruz X (2007) Characterization of compensated mutations in terms of structural and physico-chemical properties. J Mol Biol 365: 249-256.