Prediction of Factors Determining Changes in Stability in Protein Mutants

I n a u g u r a l - D i s s e r t a t i o n

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Parthiban Vijayarangakannan

aus Dindigul (Indien)

KÖLN 2005

Berichterstatter: Prof. Dr. Dietmar Schomburg

Prof. Dr. Heinz Saedler

Tag der letzten mündlichen Prüfung: 13 Jan 2006

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Preview

The relationship between the conformational stability and chemical integrity of a protein is of particular importance to understanding the mechanisms of protein folding and inactivation. On exposure to changes in environmental conditions (elevated temperatures, acidic/basic conditions, or the presence of structure perturbing solutes), protein molecules may undergo either conformational changes (local changes in secondary and tertiary structure), reversible unfolding (cooperative loss of higher ordered structure), or inactivation (irreversible changes in structural or chemical integrity of the molecule). Perturbation of protein structure often leads to the exposure of previously buried amino acid residues, facilitating their chemical reactivity. In many cases, partial unfolding of a protein is often observed prior to the onset of irreversible chemical or conformational processes. Moreover, protein conformation generally may control the rate and extent of deleterious chemical reactions. Conversely, chemical changes to the polypeptide backbone or amino acid side chains of a protein may lead to loss of conformational stability. For instance, the reduction of disulphides or the oxidation of cysteine residues can induce protein unfolding and aggregation. The interplay between these reactions and protein conformation is crucial to emphasise the understanding of protein stability.

1.2 Non-covalent Forces in Protein Stability

1.2.1 Electrostatic Interactions

(i) Van der Waals Interactions and Electronic Shell Repulsion

Van der Waals interactions, also known as London dispersion forces, result from attractive transient oscillating diploes that non-bonded atoms induce in each other. This transient dipole is generated by electrons moving in relation to the nucleus. In a pair of atoms each dipole polarizes the opposing atom. The attraction energy is proportional to r^{-6}, the distance between the nuclei, and to

the polarisability of the atoms. Such interactions, also ubiquitous, are fairly weak and short range. Because of the strong distance dependence of van der Waals interactions, the packing of atoms in the protein core, relative to their interaction with solvent, is important in determining whether they will stabilise or destabilise the native state.

The electronic shell repulsion is due to sterical hindrance when neighbouring atoms start to have overlap of the electron clouds. The repulsion of the electronic shells is proportional to r^{-12}. The attractive (distant) and the repulsive (close) components are usually taken together and described by the Lennard-Jones potential. Electrostatic repulsion may be more important, not only in destabilising the native state but also in terms of its effect on the degree of extension of the unfolded state.

A series of simplifications have been made in calculating van der Waals interactions. Mainly, only atoms in contact are taken as the close neighbours that must be considered. Also, electrostatic interactions are simplified with regard to geometry of interactions. These simplifications result in a van der Waals potential which is isotropic (equal in all directions) and is a function solely of contact distance. The values of the partial charges are subject to large inaccuracies. The complete expression for dispersion forces and electron repulsion is given below:

$$E_{vdw,ij} = -A/r_{ij}^{6} + B/r_{ij}^{12} \qquad (1)$$

Here A and B are constants depending on the atoms. The parameter, B, is taken from the sum of van der Waals radii; the parameter, A, is from the polarisability of the atoms. The resulting minimum corresponds to the most favourable atomic distance.

(ii) Hydrogen Bonding

Whenever two heavy (non-hydrogen) atoms with opposite partial charges [donor (D)-acceptor(A) pairs] were found to be within a distance (d) of 3.5Å, a hydrogen bond has been inferred. The geometrical goodness of the hydrogen bond was assessed by computing the values of the following angles:

(1) Angle θ_D between vectors BD-D and D-A, BD is the atom covalently bonded to the donor (D) atom.

(2) Angle θ_A between vectors D-A and A-BA, BA is the atom covalently bonded to the acceptor (A) atom.

A hydrogen bond was taken to have good geometry if both of these angles lie in the range of 90-150°. The distance *d* also slightly varies according D-A pairs. Hydrogen bonding is quite sensitive to distance constraints. Hydrogen bonds between NH-O, OH-N and OH-O need an approximate distance range of 2.55-3.04Å, 2.62-2.93Å and 2.65-2.93Å respectively. The amount of energy one hydrogen bond contributes towards the stabilisation of a protein is calculated to be around 1-3 $kcalmol^{-1}$.

(iii) Salt Bridges (Ion Pairs)

Salt bridges or ion-pairs are a special form of particularly strong hydrogen bonds made up of the interaction between two charged residues. On the other hand, there are also non-hydrogen bonded salt bridges and this discrimination is solely based on geometric considerations. In folded proteins, pairs of neighbouring, oppositely charged residues often interact to form salt bridges.

Salt bridges play important roles in protein structure and function such as in oligomerisation, molecular recognition, allosteric regulation, domain motions, and α-helix capping (Kumar and Nussinov 1999). An early calculation (Honig and Hubbell 1984) estimated that the cost of transferring a salt bridge from water to the protein environment is approximately 10-16 kcal/mol. Using continuum electrostatic calculations, it has been shown that the desolvation penalty due to the burial of polar and charged groups in the protein interior (a low dielectric environment) during protein folding, may not be fully recovered by favourable electrostatic interactions in the folded (Hendsch and Tidor 1994) state. Salt bridges can be stabilising or destabilising to the protein structure depending on their geometry, location in the protein, electrostatic interaction between salt-bridging side-chains with each other, and that between the salt bridge and its surroundings. But, most of the salt bridges are stabilising

irrespective of whether they are buried or exposed, isolated or networked, hydrogen bonded or not.

Salt bridge formation is inferred for a pair of oppositely charged residues (Asp or Glu with Arg, Lys or His) if they meet the following criteria (Kumar and Nussinov 1999):

(1) The centroids of the side-chain charged groups in oppositely charged residues lie within 4.0Å of each other.

(2) At least one pair of Asp or Glu side-chain carboxyl oxygen atoms and side-chain nitrogen atoms of Arg, Lys or His are within a 4.0Å distance.

The location of residues forming salt bridges is characterised in terms of the solvent accessible surface areas (ASA) (Lee and Richards 1971; Tsai and Nussinov 1997) of their constituent residues, with a probe radius of 1.4Å. The location of a salt bridge in the protein is estimated by the average ASA of the salt bridge. The average ASA of a salt bridge is average of the ASAs of the two salt-bridging residues. A salt bridge is classified as being buried in the protein core if it has an average ASA of ≤20%, otherwise it is classified as being exposed to the solvent. A salt bridge between two charged residues is considered to be networked if at least one of these charged residues forms additional salt bridge(s) with other charged residue(s) in the protein. Otherwise, the salt bridge is considered to be isolated.

The geometry of a salt bridge is characterised in terms of the distance between the centroids of the salt-bridging residue side-chain charged groups, and the angular orientation of these groups with respect to each other. The angular orientation of the side-chain charged groups in the two salt-bridging residues is computed as the angle between two unit vectors. Each unit vector joins a Cα atom and a side-chain charged group centroid in a salt bridging residue.

(iv) Other Electrostatic Interactions

Electrostatic interactions occur between charges on protein groups. Such charges are present at the amino- and carboxy- termini and on many ionisable

side chains. Charges buried in the protein interior will interact strongly since the protein interior is considered a low-dielectric medium. Van der Waals interactions are also electrostatic that involve transient dipoles. This may also occur due to the presence of permanent dipoles and have similar effects.

π-π (aromatic-aromatic) interactions between the aromatic groups are important in protein structures. Phe-Phe interactions is a good example of these interactions. However, they occur almost exclusively in electrostatically attractive geometries. Electrostatically unfavourable regions are only sparsely populated. Electrostatics dominate the geometry of interaction, while van der Waals' interactions are less significant due to the hydrophobic environment of the protein core.

Cation-π interactions are also important for protein folding. A cation-π interaction is a noncovalent binding force of broad importance in biological systems and in supramolecular chemistry. It is defined as the attraction between a cation and the face of a simple π system, such as in benzene or ethylene. The physical origin of the cation-π interaction is primarily electrostatic, involving an attraction of the cation to a locus of negative electrostatic potential associated with the face of the π system. It is a common and pervasive contributor to protein secondary structure and to a wide range of small molecule and macromolecule binding interactions in biology. Within a protein, cation-π interactions (Gallivan and Dougherty 1999) can occur between the cationic sidechains of either lysine (Lys, K) or arginine (Arg, R) and the aromatic sidechains of phenylalanine (Phe, F), tyrosine (Tyr, Y) or tryptophan (Trp, W). But, histidine can participate in cation-π interactions as either a cation or as a π-system, depending on its protonation state.

1.2.2 Configurational Entropy

Whereas the interactions discussed above tend to stabilise the native protein structure, configurational entropy destabilises it. The gain in configurational entropy relates to the increased degrees of freedom available to the protein chain in the unfolded state relative to the native state. This gain comes from

both the side chains and the backbone. Although the peptide backbone of most residues in a globular protein is relatively fixed (i.e., has low entropy), those residues that are most buried within the core of the protein have even fewer backbone degrees of freedom. The entropic effect of burying side chains is more pronounced since they have considerable flexibility on the protein surface. As larger proteins bury more of their side chains, they will have an overall larger configurational entropy change per residue. This effect may help to set a limit on the size of a globular folding domain.

The amino acid configuration also affects the configurational entropy. For instance, proteins containing large proline residues will have lower entropy in the unfolded state and thus will be more stable. The opposite will be true for proteins containing a large proportion of glycine.

(1) Entropy Cost of Fixing a Backbone

The backbone entropy term is normally used to account for the entropy cost of fixing a residue backbone. This can be different for the residues that are present in organised secondary structure regions and in loops without secondary structure due to increased flexibility of residues in loops. Compactness measures of residues are normally used to distinguish the residues in loops (Guerois et al. 2002) to assess the backbone entropy. ASA (Accessible Surface Area) and atom packing information are widely used to classify the residues in these cases (Gromiha et al. 1999b). Hydrogen bonding efficiency of specific residues is also used. If both the nearby residues can form backbone-backbone H-bond between each other, they are considered to have lower configurational entropy. Comparatively higher configurational entropy is assumed for two nearby residues that are not involved in backbone-backbone H-bonds (Guerois et al. 2002).

(2) Entropy Cost of Fixing a Side chain

Side chain entropy depends on the mobility of the side chains, which in turn directly depends on the solvent accessibility of the residues in side chains. Decreased solvent accessibility reduces the mobility and packs the side chain.

This entropy term also depends on the ability of side chain residues to form hydrogen bonds or exhibit strong electrostatic forces with the adjacent residues (Bromberg and Dill 1994). If the entropy cost is bigger than the favourable interaction energy brought by the hydrogen bond and the electrostatic interactions, neither these interactions nor the entropy of the side-chain can be assessed efficiently (Guerois et al. 2002).

1.2.3 Role of Water

Water plays a crucial role in the stabilisation of proteins. The small molecular size of water relative to other liquids, along with its complex hydrogen bonded structure, makes it a good solvent for many functional groups (Shirley 1995). These same features also give rise to hydrophobic effect which has got more than one definition in literature (Shirley 1995):

(1) Transfer of a compound from an organic liquid to water.

(2) Transfer of apolar surface from any initial phase into water.

(3) Transfer into water accomplished by a large ΔC_p.

Considering protein stability, the hydrophobic effect refers to energetic consequences of removing apolar groups from the protein interior and exposing them to water. So, the second definition is considered to be more relevant. The term hydration is considered to be the transfer of any group from gas phase to water. Though, hydration and hydrophobic effect are described separately, some of the other interactions are described here:

(1) Atomic Solvation Parameters

Atomic solvation parameters (ASPs) can be explained as transfer energies from water to the protein interior (Lomize et al. 2002). They can be effectively used to incorporate the role of water in protein structure and stability assessment. Studies are available which suggest that the protein core can be approximated using atomic solvation parameters. The polarity of different atom types, that is, the rank order of their transfer energies from water to the different media, was identical for protein and organic solvents (Cali < Caro < S < N < O). However,

the absolute values and even the signs of ASP were strongly environment dependent. If mean force potentials (MFPs) are protein environment dependent based on solvent parameters (e.g., solvent accessibility), it will be more accurate than the environment independent potentials.

(2) Water molecules forming Hydrogen Bonds

Water molecules form hydrogen bonds, both in the folded and unfolded state with the primary and secondary structures of proteins. The calculation of the effect of hydrogen bonding water in protein stability is a complex issue. Several experimental studies show that the deletion of polar atoms that make hydrogen bonds with a partially or fully buried water molecule can have a large destabilising effect on the protein interaction. (Takano et al. 1997; Grantcharova et al. 2000; Covalt et al. 2001). It may be sensible to define a water bridge as a water molecule that makes more than two hydrogen bonds with the protein. Removing one of the polar groups involved in a water bridge may exclude the bound water from a particular site of the protein and induce the desolvation of the other polar groups partners of the water molecule. Thus, it is important to determine the water positions and its ability to form hydrogen bonds with protein structures.

1.2.4 Hydrophobic Effect

During protein folding, the transition from the unfolded state (with several short-lived intermediates) to a single native state is accompanied by the burial of solvated nonpolar side chains (and polar peptide units) into the nonsolvated core of the protein. The "hydrophobic effect" or "hydrophobic interaction" in protein structure is derived from the combined properties of H-bonds in water and van der Waals forces applied to amino acid residues with nonpolar side chains. A nonpolar side chain in water makes less favourable van der Waals interactions than if it was dissolved in an apolar solvent. In addition, the solvating water molecules cannot satisfy their four potential H-bonds while they surround an apolar solute. In contrast, a nonpolar side chain in an apolar core of a protein has gained favourable van der Waals interactions and has rid

itself of the dissatisfied solvating water. The interior of folded proteins is tightly packed. Residue specific hydrophobicity scales were derived by several people to quantify the hydrophobic effect of proteins. Sequence specific plots were also generated using these hydrophobicity scales (Table 1). The solvent accessibility of the amino acids was used in this study to classify the amino acids from structural training datasets and mutations. These hydrophobicity values are highly correlated with ASA of the amino acids. So, these values can also be used for classifying amino acids instead (Muyoung et al. 2005).

Amino Acid	Engleman-Steitz	Hopp-Woods	Kyte-Doolittle	Janin	Chothia	Eisenberg-Weiss
PHE	-3.7	-2.5	2.8	0.5	0.0	0.61
MET	-3.4	-1.3	1.9	0.4	-0.24	0.26
ILE	-3.1	-1.8	4.5	0.7	0.24	0.73
LEU	-2.8	-1.8	3.8	0.5	-0.12	0.53
VAL	-2.6	-1.5	4.2	0.6	0.09	0.54
CYS	-2.0	-1.0	2.5	0.9	0.0	0.04
TRP	-1.9	-3.4	-0.9	0.3	-0.59	0.37
ALA	-1.6	-0.5	1.8	0.3	-0.29	0.25
THR	-1.2	-0.4	-0.7	-0.2	-0.71	-0.18
GLY	-1.0	0.0	-0.4	0.3	-0.34	0.16
SER	-0.6	0.3	-0.8	-0.1	-0.75	-0.26
PRO	0.2	0.0	-1.6	-0.3	-0.9	-0.07
TYR	0.7	-2.3	-1.3	-0.4	-1.02	0.02
HIS	3.0	-0.5	-3.2	-0.1	-9.94	-0.40
GLN	4.1	0.2	-3.5	-0.7	-1.53	-0.69
ASN	4.8	0.2	-3.5	-0.5	-1.18	-0.64
GLU	8.2	3.0	-3.5	-0.7	-0.90	-0.62
LYS	8.8	3.0	-3.9	-1.8	-2.05	-1.1
ASP	9.2	3.0	-3.5	-0.6	-1.02	-0.72
ARG	12.3	3.0	-4.5	-1.4	-2.71	-1.8
Threshold Values						
Hydrophobic	-1.4	-0.75	0.70	0.10	-0.47	0.10
Hydrophilc	1.85	1.65	-2.4	-0.45	-0.98	-0.51

Table 1: Hydrophobicity (amino acid specific) scale values derived from various studies (Chothia 1974; Janin 1979; Hopp and Woods 1981; Kyte and Doolittle 1982; Eisenberg et al. 1984; Engelman et al. 1986).

1.3 Covalent Reactions and Protein Stability

The covalent modification of proteins in vivo has been proposed as a natural mechanism to designate enzymes for turnover. Both enzymatic and nonenzymatic pathways of posttranslational modification of proteins have been identified. Spontaneous, nonenzymatic reactions include the deamidation of

asparangynyl residues, racemisation of aspartyl residues, isomerisation of prolyl residues, and glycation of amino acids, as well as site specific metal catalysed oxidations. Enzymes have been identified in vivo that specifically interact with covalently modified proteins, including caboxymethyl transferase and alkaline protease. It has been proposed that covalent changes caused by in vivo protein oxidation are primarily responsible for the accumulation of catalytically compromised and structurally altered enzymes during aging. In addition, protein oxidation may play a role in several pathological states, including inflammatory disease, atherosclerosis, neurological disorders, and cataractogenesis.

The relationship between the conformational stability and covalent reactions of a protein is of particular importance to the understanding of the mechanisms of protein activation/inactivation. On exposure to changes in environmental conditions (elevated temperature, acidic/basic conditions, or the presence of structure perturbing solutes), protein molecules may undergo conformational changes (local changes in secondary and tertiary structure), or inactivation (irreversible changes in structural or chemical integrity of the molecule). Perturbation of protein structure often leads to the exposure of previously buried amino acid residues, facilitating their chemical reactivity. In fact, partial unfolding of a protein is often observed prior to the onset of irreversible chemical or conformational processes (Shirley 1995). Moreover, protein conformation generally may control the rate and extent of deleterious chemical reactions. Conversely, chemical change to the polypeptide backbone or amino acid side chains of a protein may lead to loss of conformational stability. For example, the reduction of disulphides or the oxidation of cysteine residues can induce protein unfolding and aggregation which plays a considerable role in the denaturation of proteins. Obviously, this coupled interaction between these two phenomena has the potential to complicate studies of protein folding and unfolding significantly.

During the protein engineering and other solutions related to the analysis of protein folding and stability, these reactions should invariably be considered

with importance to identify chemical degradation in proteins and minimise its occurrence.

1.3.1 Deamidation and Isoaspartate formation

The spontaneous, nonenzymatic deamidation of asparagines residues is one of the most commonly encountered chemical modifications of proteins. Deamidation can occur in acidic, neutral, or alkaline conditions, although the chemical mechanism of hydrolysis is strongly dependent on pH. The biological purpose of deamidation in vivo may involve the regulation of protein degradation and clearance, thus serving as a type of biological clock. Naturally occurring protein methyl transferases have also been identified that specifically modify deamidated by-products, perhaps by tagging damaged protein for either repair or clearance.

By examining the amide loss for a large series of synthetic pentapeptides of sequence (Gly-X-Asn-X-Gly and Gly-X-Gln-X-Gly) under physiological conditions, the enhanced lability of peptide amides compared to simple aliphatic amides was demonstrated (Robinson and Rudd 1974). The asparagines containing peptides were observed to deamidate faster than glutamine counterparts. Direct hydrolysis of amide linkages was found to be slow due to the presence of an intramolecular mechanism in which, under neutral to basic conditions , the peptide bond nitrogen attacks asparanginyl carbonyl residues, causing ring closure with concomitant release of ammonia. The resulting five-membered succinimide is unstable and susceptible to subsequent hydrolysis which, in turn leads to the formation of α- and β-aspartyl residues. Under acidic conditions, deamidation thought to proceed by direct hydrolysis, resulting in the formation of α-aspartyl residues alone. Asn-Gly and Asn-Ser sequences were found to be particularly labile owing to decreased steric hindrance of succinimide formation by C-terminal residues.

1.3.2 Cleavage of Peptide Bonds

The cleavage of a peptide disrupts the linear sequence of amino acid residues within a protein chain. This covalent modification, however, may or may not

affect higher ordered structure of a protein and its biological activity. There are numerous examples of both non-specific hydrolysis and proteolysis leading to extensive protein degradation as well as specific proteolytic clips activating precursor forms of enzymes. Conversely, since the intramolecular interactions responsible for tertiary structure formation are sufficiently strong (cooperative), the introduction of a single intrachain clip in the polypeptide backbone may have little or no effect on a protein's structure or function.

Three major mechanisms of peptide bond cleavage have been identified (Shirley 1995):

(1) Preferential hydrolysis of peptide bonds at aspartic acid residues under acidic conditions.

(2) At more physiological pH, C-terminal succinimide formation at Asn residues.

(3) Enzymatic proteolysis including autolysis.

The preferential hydrolysis of a peptide bond at Asp residues is generally believed to occur at the C-terminal side of this residue in polypeptide chains. The carboxyl group side chain of Asp catalyses the cleavage reaction by acting as a proton donor at pH values below the pKa of the carboxyl group. The Asp-Pro bond is known to be particularly labile due to more basic nature of the proline nitrogen. Cleavage of polypeptide can also occur under physiological conditions. Analogous to the deamidation reaction discussed previously, succinimide formation at asparagine residues can potentially lead to the spontaneous cleavage of polypeptide chains. In this case, the side chain amide nitrogen attacks the peptide bond to form a C-terminal succinimide residue and a newly formed amino terminus. This type of cleavage has been reported to occur in both model peptides and in proteins (Tyler-Cross and Schirch 1991). Contaminating proteases are often found to cleave recombinant proteins during both fermentation and purification. Strategies to limit proteolysis include the addition of protease inhibitors, careful selection of cell host including protease negative mutants, sequence modification of susceptible sites in target proteins,

and optimisation of fermentation and purification conditions. Storage of purified proteases under certain conditions may also lead to peptide bond cleavage (autolysis).

1.3.3 Cysteine Destruction and Thiol-Disulphide Interchange

Cysteine residues are naturally occurring crosslinks that covalently connect polypeptide chains either intra- or intermolecularly. Disulphides are formed by the oxidation of thiol groups of cysteine residues by either thiol disulphide interchange or direct oxidation. The probability of formation of a disulphide bond will depend on both the intrinsic stability of potential cysteine residues to free cysteines and the conformation of the protein molecule. Intracellular proteins usually lack such crosslinks and their atypical presence commonly reflects a role in an enzyme catalytic mechanism or involvement in the regulation of its activity. In contrast, extracellular proteins frequently contain disulphide bonds, probably reflecting the need for the increased stability of such proteins. The destruction of cysteine residues in proteins have been shown to proceed by a base catalysed (β-catalysed) reaction in alkaline media (pH 12-13). Protons on polypeptide α-carbon atoms are relatively labile at high pH, since it is attached to two electron withdrawing groups (-CONH-, -NHCO-). This β-elimination results in the formation of two unstable intermediates, dehydroalanine and thiocysteine (Whitaker and Feeney 1983). The same reaction can occur at neutral pH and elevated temperatures and has been shown to contribute irreversible thermoinactivation of ribonuclease and lysozyme at pH 6-8 and 90-100° C (Ahern and Klibanov 1988).

1.3.4 Oxidation of Cysteine Residues

The relative stability of a reduced cysteine residue and its oxidised disulphide counterpart depends on the redox potential of a protein's environment. In vivo, the electron donors and acceptors that interact with protein thiols and disulphides are primarily other thiols and disulphides (e.g., as reduced and oxidised glutathione). These compounds catalyse disulphide exchange reactions, resulting in the most thermodynamically favourable redox status of

protein's cysteine residues (free thiols vs disulphides). Redox buffer containing oxidised and reduced thiol compounds are used to catalyse the cysteine residues with the resultant reshuffling of disulphide bonds leading to the formation of the native protein. Reducing agents (eg. dithiothreitol) are also sometimes used to maintain cysteine residues in their active, reduced form. Some metal ions (e.g., copper, iron) at elevated pH also catalyse oxidation to form inter and intra-molecular disulphide bonds together with some non-molecular byproducts such as sulphenic acid. Purification and storage of proteins containing naturally reduced cysteine often produces inactivation (eg. acidic fibroblast growth factor).

1.3.5 Oxidation of Methionine Residues

The oxidation of methionine residues has been associated with the loss of biological activity in a number of peptides and proteins (Swaim and Pizzo 1988). During oxidation, this thioether is converted to its sulphoxide counterpart. This is a reversible reaction in which methionine residue can be regenerated either by reducing agents or enzymatically. Harsher oxidative conditions cause irreversible formation of methionine sulphone. In vitro, proteins are commonly treated with dilute hydrogen peroxide (H_2O_2) solution or stronger oxidisers to achieve methionine oxidation. In vivo, oxygen containing radicals, such as superoxide, hydroxyl, and H_2O_2, are generated in a variety of cells (e.g., neutrophils), leading to the oxidation of several amino acids, including methionine, with potential implications for various aging or disease related processes (Swaim and Pizzo 1988).

1.3.6 Photodegradation of Proteins

Both ionising and non-ionising radiation can cause protein inactivation. The effects of different types of ionising radiations (γ-rays, X-rays, electrons, α-particles) on a protein molecule (in both solid and solution states) have been examined in detail because of interest in the use of radiation as a potential sterilisation technique in the food industry (Shirley 1995). Both direct effects and indirect effects (radiolysis of water or buffer salts and subsequent protein

alterations) have been extensively documented and recently reviewed. Nonionising radiation, such as UV light, may also cause irreversible damage to protein molecules. These effects are of particular concern biologically in understanding the mechanism of cataract formation and sunburn damage. In addition, protein unfolding/refolding studies frequently utilise UV/visible and fluorescence spectroscopy as methods of detection in which the potential adverse effects of incident light on proteins must be controlled and minimised.

The amino acids tryptophan, tyrosine and cysteine are particularly susceptible to UV-A and UV-B photolysis. The absorption of photons leads to photoionisation and the formation of photodegradation products either by direct interaction with an amino acid or indirectly via various sensitising agents. (such as dyes, riboflavin, or oxygen).

1.3.7 Glycation and Carbamylation of Protein Amino Groups

Sugars are frequently used as stabilisers of proteins during storage in solution or as lyophilised powders. Reducing sugars can covalently react with protein amino groups (e.g., the ε-amino groups of lysine residues or the amino group of N-terminus of polypeptide chains), which may lead to irreversible changes in conformation and stability of proteins. When a reducing sugar, such as glucose is incubated over long periods, the spontaneous formation of a Schiff's base between protein amino groups and glucose is often observed. Through a series of subsequent reactions known as the Amodori rearrangement, covalent adducts are then formed. This process is frequently referred to as Maillard reaction or nonenzymatic browning. These Maillard adducts can further degrade to form so-called "advanced glycosylation end products" (AGEs), resulting in both protein crosslinking and the appearance of fluorescent byproducts. These glycation reactions are believed to be involved in degenerative processes in vivo.

Protein amino groups are also reactive with isocyanate ions leading to carbamylation of proteins (Stark 1965). Urea is in equilibrium with isocyanate ions. Therefore, protein unfolding experiments using this denaturant should be

done with freshly prepared urea and with minimised period of contact between urea and protein.

1.4 Protein Structural Descriptors

1.4.1 Role of Secondary Structure Elements

The main SSEs (Secondary Structure Elements), helices and strands, are formed by hydrogen bonds. Thus, a hydrogen bonding potential becomes very useful in empirical potentials. Helices are formed by hydrogen bonds between residues in the same helix. Three different helices exist, but only α-helix is more common than the others. The bonds forming helices restrict the torsion angles, and the idealised angles for 'geometrically correct' α-helix are ϕ = -57.8 and ψ = -47.0. However, the real angles usually deviate from these.

Strands and sheets are formed by successive hydrogen bonds between residues which can be far apart in sequence (Table 2). The backbone hydrogen bonding groups (N-H and O=C) are in the plane of the sheet, with the bonding groups from successive residues pointing in opposite directions. Let residue *i* be in one strand, and residue *j* in another. Then, the bonding of two strands can be either parallel or antiparallel. Parallel bonding is formed by each residue forming hydrogen bonds to two residues on the other strand, separated by a residue in the sequence (successive H-bonds). Antiparallel bonding is formed by each residue forming two hydrogen bonds with a single residue on the other strand (successive hydrogen bonds). Sheets can be parallel, antiparallel or mixed (with both parallel and antiparallel bondings). The idealised strand satisfying these constraints can be thought of as a helix with two residues per turn, with torsion angle of approximately ϕ = -120 and ψ = +120.

Distance matrices can be useful, either manually or automatically, to indicate where there can be SSEs. For idealised α-helices, the distance between the C_α atoms from the start of the helix can be roughly calculated to be 3.8, 5.4, 5.1, 6.3, 8.7, 9.9, 10.6, 12.5, ….. These distances are found by idealised angle pair for α-helices and the distances between the backbone atoms. Real helices

usually deviates from these due to irregularities. In a distance matrix, a helix will turn up as an area of small distances along the main diagonal.

For an idealised β-strand the successive distances from a residue *i* can be calculated to be 3.8, 6.6, 10.3, 13.5, 16.9, …… Real strands also deviate from these values.

Thus, the development of compactness (of amino acids) and SSE (secondary structure element) specific statistical potentials from radial pair distribution of atoms and torsion angles must be more accurate and their coarse grained nature produces a high definition of protein structure and stability.

SSEs		H-bond order
Helices	α-Helix	H-bond(*i*, *i*+4), H-bond(*i*+1, *i*+5), ……
	3_{10}-Helix	H-bond(*i*, *i*+3), H-bond(*i*+1, *i*+4), ……
	π-helix	H-bond(*i*, *i*+5), H-bond(*i*+1, *i*+6), ……
Sheets	Parallel	H-bond(*i*, *j*), H-bond(*j*, *i*+2), H-bond(*i*+2, *j*+2), H-bond(*j*+2, *i*+4), ……
	Antiparallel	H-bond(*j*, *i*), H-bond(*i*, *j*), H-bond(*j*+2, *i*+2), H-bond(*i*+2, *j*+2), ……

Table 2: Conservation of H-bond order in SSEs (secondary structure elements).

1.4.2 The Denatured State

For most proteins, the denatured state is insoluble and many of the physical techniques available for characterising it (in solution) are relatively insensitive for detecting its structure that has a highly flexible, dynamic character. In the absence of any evidence, it was only simple to assume that it is a featureless random coil state. It was also essential to interpret the experimental data because the energetics of protein's native state achieves a larger role only when it's assumed to be a random coil. In effect, the experimentally measured thermodynamic parameters reflect the entire process of protein folding, with the sum (Shortle 1996) of protein interactions in the native state supplying all the free energy needed to derive the formation of structure. In spite of some attempts to explain the role of the denatured state, more concrete evidences are needed to understand the definitive nature of its involvement in protein folding and stability.

The denaturing agents play a dominant role with denatured state rather than the native state. Some of the denaturing agents like SDS have a close interaction with denatured state. These amphipathic compounds interact almost exclusively with the denatured state (3), because most of the hydrophobic surface to which they bind becomes available only when the native state breaks down. Although the details of the chemistry underlying the action of solvent denaturants like urea and guanidine hydrochloride are still poorly understood at phenomenological level, their mechanism of action is thought to involve weak binding or adsorption to nonpolar surfaces (4, 5). Because much more nonpolar surface is exposed in the denatured state, urea and guanidinium ion promote the dissociation and unfolding of proteins through their more extensive association with the denatured state.

1.4.3 Protein/Amino Acid Packing Measures

The compactness of a protein can be defined as the ratio of solvent accessible area of the protein and the surface area of a sphere with equal volume to the protein. Assuming that most proteins are more or less globular in shape, a better packed protein will have a smaller ratio value. For analysing point mutations associated with single amino acids, it becomes important to analyse the compactness of a single amino acid. Some of the packing measures are given below:

(1) It can be described as the relative ASA which is derived as the ratio between the real ASA of an amino acid in native state and the constant ASA of the same amino acid in ALA-X-ALA extended state.

(2) Other measures of compactness are also available which prove to be viable in certain cases of protein structure prediction methods. It is derived as the distribution of C_β or C_α atoms around any amino acid. The number of C_β atoms in a distance of 6-8Å can be calculated, where the compactness is directly proportional to the number of selected atoms at a defined cutoff distance.

1.4.4 Protein Flexibility Measures

Dynamics of proteins plays an important role in function of proteins (Brooks et al. 1988). Stability of a protein after a point mutation, flexibility of protein environment may have a considerable role to accommodate the mutated amino acid in any specific position. However, assessing protein flexibility of the mutated region is necessary to include its effect.

One of the earliest attempts to accommodate small changes in conformation were through the use of implicit methods (Jiang and Kim 1991) for protein-ligand docking studies. The protein is held fixed, but a "soft"-scoring function is used to evaluate the fit of the ligand to the receptor. Often, scoring functions are derivatives of force fields from molecular mechanics, modified for use in a new application. Soft functions allow for some overlap between the ligand and the protein, giving a small estimate of the plasticity of the receptor. Protein structural stability or rigidity is also highly correlated with protein unfolding (Rader et al. 2002).

The ideal method to predict protein flexibility is to perform molecular dynamics simulation of proteins in aqueous solution with an accurate physical-based energy function (Brooks et al. 1983). The simulation, however, often requires long computational time. Thus, it is of interest to develop a simple efficient method to predict protein flexibility. Several methods have been developed for an efficient flexibility prediction.

(1) Gaussian and anisotropic network models (Micheletti et al. 2004; Pandey et al. 2005).

(2) Graph theory based model (Jacobs et al. 2001).

(3) A statistical mechanical distance constraint model (Jacobs et al. 2003; Livesay et al. 2004).

(4) Statistical mean-field theory based models (Micheletti et al. 2002; Pandey et al. 2005).

Gaussian and anisotropic network models (GNM and ANM) predict flexibility based on normal mode analysis of a simple representation of proteins, whereas the graph theory provides a coarse-grained estimation of flexibility based on connectivity. The Hamiltonian of an atom mean field theory is constructed using either C_α or all atoms with bonded and non-bonded terms used separately. The distance constraint model (DCM) identifies flexible regions within protein structure consistent with specified thermodynamic condition. It is based on a rigorous free energy decomposition scheme representing structure as fluctuating constraint topologies. Entropy non-additivity is problematic for naive decompositions, limiting the success of heat capacity predictions. The DCM resolves non-additivity by summing over independent entropic components determined by a network-rigidity algorithm.

1.5 Amino Acid Substitution Matrices

The divergence among sequences can be modeled with a mutation matrix. The matrix, denoted by M, describes the probabilities of amino acid mutations for a given period of evolution.

$$P_r \text{ (amino acid } i \rightarrow \text{amino acid } j) = M_{ji} \qquad (2)$$

This corresponds to a model of evolution in which amino acids mutate randomly and independently from one another but according to some predefined probabilities depending on the amino acid itself. This is a Markovian model of evolution and while simple, it is one of the best models. Intrinsic properties of amino acids, like hydrophobicity, size, charge, etc. can be modeled by appropriate mutation matrices. Dependencies which relate one amino acid characteristic to the characteristics of its neighbours are not possible to model through this mechanism. Amino acids appear in nature with different frequencies. These frequencies are denoted by *fi* and correspond to the steady state of the Markov process defined by the matrix M., i.e., the vector f is any of the columns of or the eigenvector of M whose corresponding eigenvalue is 1 (Mf=f). This model of evolution is symmetric, i.e., the probability of having an

i which mutates to a j is the same as starting with a j which mutates into an i. The following is a list of amino acid substitution models which use matrices.

1.5.1 Empirical substitution models

In contrast to DNA substitution models, amino acid replacement models have concentrated on the empirical approach. Dayhoff and co-workers developed a model of protein evolution which resulted in the development of a set of widely used replacement matrices (Dayhoff et al. 1978). In the Dayhoff approach, replacement rates are derived from alignments of protein sequences that are at least 85% identical; this constraint ensures that the likelihood of a particular mutation being the result of a set of successive mutations is low. One of the main uses of the Dayhoff matrices has been in database search methods where, for example, the matrices P(0.5), P(1) and P(2.5) (known as the PAM50, PAM100 and PAM250 matrices) are used to assess the significance of proposed matches between target and database sequences. However, the implicit rate matrix has been used for phylogenetic applications.

1.5.2 PAM matrices

In the definition of mutation the matrix M implies certain amount of mutation (measured in PAM units). A 1-PAM mutation matrix describes an amount of evolution which will change, on the average, 1% of the amino acids. In mathematical terms this is expressed as a matrix M such that

$$\sum_{i \in \sum} fi(1 - Mii) = 0.01 \qquad (3)$$

The diagonal elements of M are the probabilities that a given amino acid does not change, so $(1\text{-}M_{ii})$ is the probability of mutating away from i.

If we have a probability or frequency vector p, the product Mp gives the probability vector or the expected frequency of p after an evolution equivalent to 1-PAM unit. Or, if we start with amino acid i (a probability vector which contains a 1 in position i and 0s in all others) M*i (the ith column of M) is the corresponding probability vector after one unit of random evolution. Similarly, after k units of evolution (what is called k-PAM evolution) a frequency vector

p will be changed into the frequency vector Mk p. Notice that chronological time is not linearly dependent on PAM distance. Evolution rates may be very different for different species and different proteins.

1.5.3 Dayhoff matrices

Dayhoff and co-workers (Dayhoff et al. 1978) presented a method for estimating the matrix M from the observation of 1572 accepted mutations between 34 superfamilies of closely related sequences. Their method was pioneering in the field. A Dayhoff matrix is computed from a 250-PAM mutation matrix, used for the standard dynamic programming method of sequence alignment. The Dayhoff matrix entries are related to M250 by

$$D_{ij} = 10 \log_{10} \frac{(M^{250})ij}{f_i} \qquad (4)$$

1.5.4 JTT matrices

Recently, two groups (Gonnet et al. 1992; Jones et al. 1992) have used the same methodology as Dayhoff, but with modern databases. The Jones et al. model has been implemented for phylogenetic analyses with some success. Jones and co-workers have also calculated an amino acid replacement matrix specifically for membrane spanning segments. This matrix has remarkably different values from the Dayhoff matrices, which are known to be biased toward water-soluble globular proteins.

1.5.5 Other empirical models

Some groups (Adachi and Hasegawa 1996) have implemented a general reversible Markov model of amino acid replacement that uses a matrix derived from the inferred replacements in mitochondrial proteins of 20 vertebrate species. The authors show that this model performs better than others when dealing with mitochondrial protein phylogeny.

1.5.6 Blosum (Block substitution matrices)

Blosum is a different approach (Henikoff and Henikoff 1992) and used local, ungapped alignments of distantly related sequences to derive the BLOSUM

series of matrices. Matrices of this series are identified by a number after the matrix (e.g. BLOSUM50), which refers to the minimum percentage identity of the blocks of multiple aligned amino acids used to construct the matrix. It is noteworthy that these matrices are directly calculated without extrapolations, and are analogous to transition probability matrices P(T) for different values of T, estimated without reference to any rate matrix Q. The BLOSUM matrices often perform better than PAM matrices for local similarity searches, but have not been widely used in phylogenetics.

1.5.7 Poisson models

A simple, non-empirical model (Nei 1987) of amino acid replacement implements a Poisson distribution, and gives accurate estimates of the number of amino acid replacements when species are closely related.

1.6 Energy Functions

1.6.1 Experimental Protein Denaturation

Protein denaturation is commonly defined as any noncovalent change in the structure of a protein where organised molecular configuration is disturbed. This change may alter the secondary, tertiary or quaternary structure of the molecules. In this definition, it should be noted that what constitutes denaturation is largely dependent upon the method utilized to observe the protein molecule. Some methods can detect very slight changes in structure, while others require rather large alterations in structure before changes are observed.

$$Unfolded \underset{}{\overset{Keq}{\rightleftharpoons}} Folded$$

The main causes of denaturation can be classified into the following criteria:

(1) Changes in temperature and pH.

(2) Changes in salt concentration.

(3) Detergents

(4) H-bonding agents.

(5) Oxidants and reductants

(6) Non-polar solvents.

Increase in temperature is directly proportional to the increase in kinetic energy of the folded protein structure that eventually results in the breakage of relatively weak H-bonds, electrostatic interactions and hydrophobic interactions. Changes in pH directly alter the electric charge of acidic or basic functional groups on the protein which disrupt or create electrostatic interactions that will alter the protein structure (Table 3).

pH 2	carboxylic acid groups are not charged
pH 7	carboxylic acid groups are negatively charged ($-COO^-$) and amino groups are positively charged ($-NH_3^+$)
pH 12	amino groups are not charged

Table 3: Effect of pH in altering the charges of amino and carboxylic acid groups.

While high salt concentration tends to reduce the electrostatic interactions, low salt concentrations increase the electrostatic interactions. Extra ions in solution tend to insulate charges in protein. The Hofmeister series (Chi et al. 2003) describes the relative effects of some anions and cations in precipitating proteins which basically states that their effect is independent and additive. The effect of anions is relatively more than the cation. Anions were also further divided into chaotropic and cosmotropic in nature. The former are larger in size and considered to be water-structure breakers with high polarisability. These are mostly destabilising for the proteins. But, the latter are usually small, stabilising and considered to be polar water-structure makers with low polarisability. Protein precipitating (salting-out) experiments are also used for the purification of protein which results to the maximum of 75% removal of protein impurities normally.

(1) Cations:

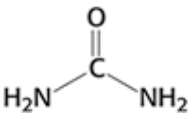
$NH4^+$, K^+, Na^+, Li^+, Mg^{2+}, Ca^{2+}, guanidium, urea, etc.

(2) Anions:

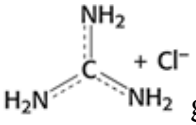
SO_4^{2-}, HPO_4^{2-}, OH^-, F^-, $CH3COO^-$, Citrate, tartrate, Cl^-, Br^-, NO_3^-, ClO_3^-, I^-, ClO_4^-, SCN^-, etc.

Detergents are amphiphilic molecules (both hydrophobic and hydrophilic parts) and disrupt hydrophobic interactions. Hydrophobic parts of the detergent associate with the hydrophobic parts of the protein (coating with detergent molecules) and hydrophilic ends of the detergent molecules interact favourably with water (nonpolar parts of the protein become coated with polar groups that allow their association with water). Hydrophobic parts of the protein no longer need to associate with each other which eventually results in dissociation of the non-polar R groups that can lead to unfolding of the protein chain. This effect is also similar to non polar solvents.

As described in this chapter, H-bonding is important in maintaining secondary, tertiary and quaternary structure of the protein. H-bonding agents compete with H-bonding between protein functional groups. This stops the H-bonding association of R groups. Dissociation can lead to unfolding of the protein chain.



Urea



guanidium chloride

Urea and Guanidine HCl are well known H-bonding agents that are frequently used in denaturation experiments to calculate the folding free energy (ΔG).

1.6.2 Oxidants and Reductants

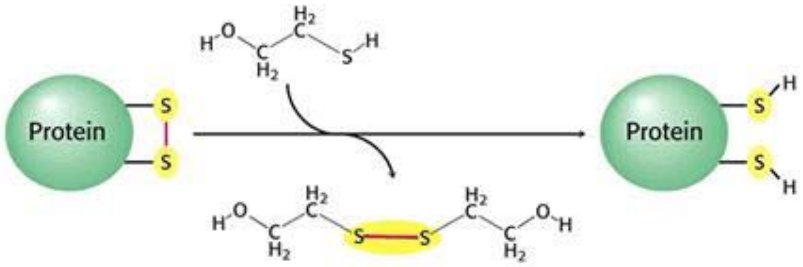


Fig. 1: Disulphide bond breakage.

Mild reductants and mild oxidants can lead to changes in protein conformation, that may alter the function of the protein. Mild reductants can break disulphide bonds (Fig. 1) and may lead to dissociation of parts of the protein chain(s) that are normally associated. Mild oxidants can cause the formation of disulphide bonds may lead to association of parts of the protein chain that are normally not associated. Stronger oxidising and reducing agents can change the nature of protein R groups most easily oxidised, if R groups next to sulphydryl groups are phenol (Tyr), hydroxyl (serine & threonine), amine (Lys, Arg, His), sulphide (Met)

Non-polar solvents disrupt hydrophobic interactions (association of non polar R groups) because non-polar R groups no longer associate, since they can now interact with the solvent. This leads to the dissociation of the non-polar R groups and results in unfolding of the protein chain.

1.6.3 Free Energy Derivation

Two basic approaches are present to study various contributions to protein stability: to study the protein stability as a function of environmental variables, such as temperature, denaturant concentration, pH, pressure, etc. with site-specific mutations in most cases. Second major approach is to study model systems where one attempts to mimic the folding process, with a system simpler than a protein so that it's easier to interpret. Though the second approach has been considerably used, the first approach's experimental data is proved to be more accurate and reliable for its use with theoretical models for predicting protein stability.

Depending on the environmental variables, different methods are employed to measure the free energy differences during protein denaturation. These methods are listed below:

(1) Differential Scanning Calorimetry (DSC) in which excess heat capacity has been used as a function of temperature.

(2) Fluorescence spectroscopy that uses intrinsic fluorescence of aromatic amino acids to monitor unfolding/refolding transitions induced by chemical denaturants, temperature, pH and pressure.

(3) UV spectroscopy that uses absorption of near UV (small shifts in wavelengths for folded and unfolded states) by amino acids to study folding/unfolding transitions.

(4) Circular Dichroism that measures the chirality of protein structures which can clearly distinguish between tertiary, secondary and unfolded structures.

In principle, apart from the widely used techniques described above, any physical technique that is capable of distinguishing the native and denatured states of a protein can be used to monitor the unfolding transition. Biological activity measurements, immunochemical techniques, hydrodynamic methods, such as viscosity, NMR, UV difference spectroscopy can all be used to follow unfolding.

1.6.4 ΔΔG and $\Delta\Delta GH_2O$

The native state of most naturally occurring proteins is only about 5-15 kcal/mol more stable than its unfolded conformations. By assuming the two state mechanism, only the folded and unfolded forms of the protein are present at significant concentrations and

$$f_F + f_U = 1 \qquad (5)$$

where f_F and f_U represent the fraction of the total protein in the folded and unfolded conformations, respectively. The observed values at any point of the transition curve is given by

$$y = y_F f_F + y_U f_U \qquad (6)$$

where y is any observable parameter chosen to follow unfolding, and y_F and y_U represent the values of y characteristic of the folded and unfolded protein. The values of y_F and y_U for any point in the transition region are obtained by extrapolation of the pre- and post-transition baselines, which is generally achieved by least squares analysis. Combining equations (5) and (6) yields

$$f_U = (y_F - y)/(y_F - y_U) \qquad (7)$$

The equilibrium constant K_U, and the free energy change, ΔG_U, for the folding/unfolding reaction can be calculated using

$$K_U = f_U/(1-f_U) = f_U/f_F = (y_F - y)/(y - y_U) \qquad (8)$$

and

$$\Delta G_U = -RT\ln K_U = -RT\ln[(y_F - y)/(y - y_U)] \qquad (9)$$

where R is the gas constant and T is the absolute temperature (K).

$\Delta G_U H_2O$, the free energy change in zero denaturant concentration is then calculated for the protein in equilibrium. To obtain an estimate of $\Delta G H_2O$ from these studies, accurately measured values of the equilibrium constant, K_U, are determined under denaturing conditions, and an attempt is made to extrapolate back to zero denaturant concentration. ΔG_U is generally found to vary linearly with denaturant concentration. The simplest and at present most widely used model assumes that the linear dependence of ΔG_U on denaturant concentration observed in the transition region continues to zero concentration of denaturant. A least square analysis can be used to fit the data to an equation of the form:

$$\Delta G_U = \Delta G_U H_2O - m[\text{denaturant}] \qquad (10)$$

The value of *m* in this equation is a measure of the dependence of free energy on denaturant concentration. Apart from the linear expolation model, there are also other methods for analysing the denaturation curves (Wyman 1964; Inoue and Timasheff 1968).

After the point mutation, the difference in ΔG_U and $\Delta G_U H_2O$ between mutant and wild type protein is then calculated ($\Delta\Delta G_U$ and $\Delta\Delta G_U H_2O$). These values are often mentioned as $\Delta\Delta G$ and $\Delta\Delta G H_2O$ in the literature.

1.6.5 Theoretical Background

Given the thermodynamic hypothesis, studies of protein folding (i.e. structure prediction, fold recognition, homology modelling and design) generally make use of some form of energy function. There are three different types of energy function that are in use.

The first is based on the true effective energy function, which can be obtained, in principle, from a fundamental analysis of the forces between the particles. These are often known as physical effective energy functions (PEEF). PEEFs typically consist of a molecular mechanics energy function and a model for the effect of solvation on the free energy. Thus, PEEFs are approximations to the (unknown) true energy function.

The second is the empirical effective energy function (EEEF) and its approaches combine a physical description of the interactions with lessons learned from experiments. Good examples of such algorithms are the helix/coil transition algorithm AGADIR (Munoz and Serrano 1997; Lacroix et al. 1998) or the SPMP (Takano et al. 1999) method. The AGADIR algorithm is accurate at predicting the helical content of peptides in solution and has been used to design mutations that increase the thermostability of a protein through local interactions (Guerois et al. 2002). A limitation of this algorithm is that it can be applied only to α-helices and cannot take tertiary interactions into account. Later, advanced variants of EEEFs were used by others for predicting changes in protein stability upon mutation.

The third is an energy function based on data derived from known protein structures (often statistics concerning pair contacts and surface area burial). These are often known as statistical effective energy functions (SEEF). These are used initially by several researchers in the prediction of protein structure and stability.

1.7 Experimental Substitution Methods

Techniques for altering protein primary structure (sequence) using point mutations fall into three major categories: site-specific mutagenesis, random point mutations and shuffling. Numerous variants of each category exist, but the principles are general.

1.7.1 Site-Specific Mutagenesis

If a protein is produced in the laboratory by expression of its gene, point mutations can be readily introduced by site-directed mutagenesis, using the Polymerase Chain Reaction (PCR). Typically, the gene has already been cloned into a plasmid.

1.7.2 Random Mutations at Specified Positions

It is often desirable to investigate the effect of more than one amino acid on protein stability and function. If there is reason to believe a particular position was critical to folding, it's essential to determine the substitutions at that position with increased stability. The most direct approach is to construct 19 site-directed mutations, each with the codon of a different amino acid at the centre of the primer, and measure the folding free energies of the wild type and all mutants. An alternative is to generate all possible mutants and screen for the most stable.

1.7.3 DNA Shuffling

DNA shuffling is used to carry out random mutations throughout the whole gene. The easiest way to construct random mutations is to do PCR with low-fidelity polymerase, which makes random mistakes during gene duplication. Such error prone PCR can be combined with DNA shuffling so that diverse

sequences can be rapidly generated and selected. The method is intended to mimic recombination used by nature to generate biological diversity. A pool of identical or closely related sequences is fragmented randomly, and these fragments are reassembled into full-length genes via self-priming PCR and extension. This process is known as "assembly PCR" and yields crossovers between related sequences due to template switching. Such shuffling allows rapid combination of positive-acting mutations and simultaneously flushes out negative-acting mutations from the sequence pool. When coupled with effective selection, and applied iteratively, such that the output of one cycle is the input of the next cycle, DNA shuffling is an efficient process for directed molecular evolution. DNA shuffling is a recent invention, with the ability to sample much larger sequence space than other mutagenesis techniques. Most of its applications have been focused on discovering mutations leading to higher activities (e.g. resistance to antibiotics, higher enzymatic activities, and stronger cell fluorescence signal). Dramatic activity improvements have been achieved using DNA shuffling, and it will not be surprising if this technique uncovers mutated proteins that are much more stable than the wild type.

1.7.4 Protein Stability Assessment

There are several methods to measure protein stability as a function of an environmental perturbant. The most fundamental measures of protein stability involve temperature as the environmental variable. Differential scanning calorimetry (DSC), in which the excess heat capacity of a protein solution is determined as a function of temperature, can provide all the thermodynamic parameters that specify the stability of the protein as a function of temperature: ΔH, ΔS and ΔC_p. The ability to make single amino acid changes has provided another means by which investigators can probe the stabilisation of proteins. The calculation free energy of unfolding was already explained in this chapter. On the other hand, denaturants are also used to measure protein stability. Spectroscopic (e.g., fluorescence spectroscopy, circular dichroism) techniques are also widely used to track the folding-unfolding transition, when these denaturants are used.

1.8 Uses of Predicting Protein Stability

1.8.1 Increased Thermostability

(1) Changes in food proteins during processing and heating, especially at extremes of temperature and pH.

(2) Enzymes as catalysts for organic synthesis and as biocatalysts for manufacturing of chemicals, sweeteners and detergents.

(3) Natural peptides/proteins as therapeutic agents: understanding of causes and mechanism of inactivation for developing rational strategies for their stabilisation.

In each of these applications, protein molecules are exposed to nonphysiological conditions resulting in stress on their structural and chemical integrity that may lead to both their covalent and noncovalent alteration.

1.8.2 Decreased Stability / Thermosensitivity

Point mutations can be used to design thermosensitive proteins. In yeast, temperature-sensitive alleles of Cycline Dependent Kinases (CDKs) have promoted the analysis of cell cycle control. Temperature sensitive alleles in plants are also very useful to study cell cycle control as well as plant development, which eventually helps obtaining a synchronisable organism. In rice, molecular analysis of functional regions was done using temperature sensitive mutants. For the mutations, which were not analysed previously for its role, the computational prediction tools are highly useful to design the initial set of point mutations. They can also be used to reduce a big set of already available point mutations to a smaller number.

1.8.3 Mutations and Drug Targets

Analysis of the stability of point mutations can be used to identify a wide spectrum of drug resistance conferring mutations. A simple experimental overview can be described as follows: Initially, the target cDNA is cloned into a retroviral vector. Then, the vector is propagated in bacteria that are deficient

in DNA repair mechanisms, creating an exhaustive library of mutations in the target genes. The drug sensitive cells are transfected or infected with mutated vector and dispersed in soft agar in the presence of drug. The resistant colonies can be isolated and then the target cDNA is recovered and sequenced to identify mutations. In next step of confirmation, point mutations are recreated in the native cDNA by site-directed mutagenesis and resistance is measured by proliferation assays and/or immunoblotting.

Mutations can be analysed for their structural consequences by mapping onto a model of the protein crystal structure. The prediction tool can determine the mutations that are altering the functional stability of the target protein, thus altering the resistance phenomena against the drug. Anticancer variants as well as resistance to drugs exhibited by HIV1 proteases have been studied previously. New prediction tools with promising results and higher accuracy can be used to analyse mutations and speed up the drug development cycle.

2 LITERATURE REVIEW

Force fields and energy functions are generic in usage to some extent, and can be used in analyzing many of the properties of molecular structures like proteins, nucleic acids and their complexes. Protein and nucleic acid structural properties are studied for several decades using the energy functions and several other methods. Likewise, predicting protein mutant stability is also a subject of critical interest and several methods apart from energy functions were used. Thus, the purpose of this review is two-fold: to develop methods for predicting protein stability changes upon mutations and to examine the use of statistical energy functions with methods other than protein stability predictions.

This chapter covers mostly the protein stability prediction models that try to predict the changes in mutant stability. Many context specific experiments were carried out for evaluating the stability of particular proteins. But, the most generic models that predict wide range of mutations with empirical or statistical energy functions, neural network based models, support vector machines are reviewed. Advantages and problems of all the current approaches for predicting protein stability were studied and the challenges were analysed for a newly developed prediction model. Possible solutions are constructed for a new model and the already existing methods are modified to improve the prediction efficiency and reliability of the new model.

The energy functions, both empirical and statistical, that are used to predict the protein structure, protein-protein interactions, protein flexibility, enzyme reaction mechanisms, protein-nucleic acid complexes and protein-drug interactions are all closely related with the prediction of currently developed statistical energy functions in this work (Gohlke et al. 2000; Micheletti et al. 2002). But, the evolution and application of statistical mean force potentials differs considerably for its use in predicting protein structural stability. The evolution of mean force potentials range from the simple sequence and residue level models to complicated atom level models and coarse grained orientational

potentials, boasting improved prediction efficiency compared to each other. Distinguishing the amino acids between each other also differs depending on its application. The electrostatic charges, size, polarity, etc. play different roles in terms of structure, function, and all other inter residue interactions. Proteins with similar structure and different function are good examples for these interactions.

2.1 Use of Empirical and Statistical Energy Functions

2.1.1 Protein Structure Solutions

Protein structure prediction from sequence remains fundamentally unsolved despite more than three decades of intensive research effort. Fold recognition, homology modelling and design were carried out using mean force potentials derived from protein structures. The possibility of predicting a protein's structure from its amino acid sequence is limited by errors in the energy parameters (Finkelstein et al. 1995a; Finkelstein et al. 1995b) and by the astronomical number of possible alternative structures. Prediction is a feasible task only with energy functions that allow fast and efficient sorting over many conformations. To this end, a residue–residue approximation is usually used, which attributes all atomic interactions between residues to a single point within each residue. Physically, such simplified potentials should result from some averaging of the atomic interactions over various positions and conformations of the interacting amino acid residues and atoms in addition to the surrounding solvent molecules (Reva et al. 1997). Residue level potentials were developed initially for the structure prediction models. Later, mean force potentials were developed in atomic level (Melo and Feytmans 1997; 1998).

Residue level potentials used electrostatic charges (Zhu and Karlin 1996; Karlin et al. 1999) for the calculation of energy functions. Mean force potentials at atomic level were involved and replaced residue level potentials to some extent, which enabled increased accuracy of protein structure definition by approximating non-bonded atomic interactions (Colovos and Yeates 1993). Solvent accessible contact energies (Delarue and Koehl 1995) were used to

derive the atomic environment energies. Atom densities were also studies with two types of environments, where one based on side-chain atom contacts and the other based on all atom contacts (Karlin et al. 1999). Several classification systems for the amino acid atoms (Cline et al. 2002; Mintseris and Weng 2004) were given by different investigators. They were compared and reviewed for their ability to represent the protein structure parameters. A novel atom type model was proposed based on chemical nature, location and connectivity of atoms to describe the non-local interactions in protein structures (Melo and Feytmans 1997; 1998). Measures of residues packing densities (Baud and Karlin 1999; Fleming and Richards 2000) were analysed and reviewed (Levitt et al. 1997) to distinguish the protein environments. Protein environment specificity has also been used recently to dissect the matrices of contact potentials using hydrophobicity and secondary structure (Muyoung et al. 2005).

2.1.2 Protein Folding

Investigations of proteins that fold in a two-state manner, i.e. where no partially folded intermediates accumulate during folding (Paci et al. 2005), have led to major advances in our understanding of the elementary steps of protein folding (baker 2000, Dobson 2003). The introduction of the protein engineering method to obtain residue specific information (Serrano et al. 1992) was useful in this context.

One of the early developments of statistical potentials for protein folding problem was developed from the theory of spin glasses which described the process as a polymer collapse of a homopolymer that has no latent during the transition. Three-dimensional models of folding intermediates were created by keeping the portions of the protein in the native geometry, allowing other regions to relax into random conformations. In this study, the $\Delta\Delta G$ of site-directed bi-histidine mutants were used and the peripheral regions that are differentially populated according to their relative stability were analysed.

2.2 Stability Assessment

Several experimental hurdles exist in analysing protein stability changes upon point mutations. The measurement of free energy change ($\Delta\Delta G$ and $\Delta\Delta GH_2O$ explained in Introduction) is not a straightforward process and derived from different experimental techniques with several assumptions. These assumptions include,

(1) The two-state mechanism of protein folding that include only native (folded) and denatured (unfolded) states without intermediates. For many proteins, the amount of intermediates is assumed to be negligible.

(2) When measuring $\Delta\Delta GH_2O$, the free energy of unfolding at zero denaturant concentration, a linear extrapolation is assumed in most of the cases (Shirley 1995).

There are several experimental characteristics expected for a two-state process. Two of these characteristics are observed when determining urea or guHCl denaturation curves. To avoid the incorrect interpretation of experimental data, it is best to apply these tests before attempting a thermodynamic analysis. These include:

(1) The transition from native to denatured protein should be characterized by an abrupt, single step and should not contain a plateau or even a shoulder.

(2) When unfolding is followed with several different techniques, f_{obs}, the observable fraction of unfolded protein, should be independent of the observable parameter. This is sometimes referred to as multiple-variable test. When multiple parameters are used, one parameter may distinguish the unfolding of tertiary structure alone, while the other(s) may involve the denaturation observation at secondary structure level.

Even, if these characteristics are observed, the folding mechanism under a given set of conditions is not necessarily two-state. However, if either of these characteristics is not observed, the unfolding mechanism cannot be described as two-state.

In ProTherm (Bava et al. 2004) web database, careful distinction has been made to include $\Delta\Delta G$ and $\Delta\Delta GH_2O$ separately from literature. Besides, all the auxiliary data (techniques, publication, year, pH and temperature) have been given for the majority of the point mutations.

2.2.1 Protein Structure Quality

The quality of protein structures deposited in the Protein Data Bank (PDB) (Berman et al. 2000) plays an important role for the statistical potentials to derive a good quality. Deposition of high resolution structures increase drastically as the PDB grows larger now-a-days. The Aug 2005 update contains 32149 structures totally which includes 25416 proteins (peptides and viruses) from X-ray diffraction studies and 3916 proteins from NMR.

2.3 Theoretical Prediction Models

For predicting protein stability changes, several methods were used which can be divided into energy function based methods using force-fields, neural network based methods and SVM (support vector machines) based methods.

The development of a fast and reliable protein force-field is a complex task, given the delicate balance between the different energy terms that contribute to protein stability (Lazaridis and Karplus 2000). Many different force-fields have been constructed for predicting protein stability changes. These range from energy functions based on pure statistical analysis of structural sequence preferences (O'Sullivan et al. 2004), and force-fields based on multiple sequence alignments (Munoz and Serrano 1997), detailed molecular dynamics force-fields (Kollman et al. 2000). These energy functions can be divided into three major categories:

(1) Physical effective energy functions (PEEF).

(2) Empirical effective energy functions (EEEF).

(3) Statistical effective energy functions (SEEF).

Physical effective energy functions are computationally very expensive and they can therefore be used only on small sets of protein mutants. The computation time can be reduced somewhat by using implicit terms for solvation energies and side-chain entropies, but the time required to get a reliable estimate of a free energy difference between a wild-type and mutant protein is still significant (Guerois et al. 2002).

EEEF approaches combine a physical description of the interactions with lessons learned from experiments. Good examples of such algorithms are the helix/coil transition algorithm AGADIR (Munoz and Serrano 1995; 1997) or the SPMP method (Takano et al. 1999). The AGADIR algorithm is accurate at predicting the helical content of peptides in solution and has been used to design mutations that increase the thermostability of a protein through local interactions (Lacroix et al. 1998). A limitation of this algorithm is that it can be applied only to a-helices and cannot take tertiary interactions into account.

The power of SEEFs is that they contain terms that account for complex effects that are difficult to describe separately, and they contain empirical approximations for the denatured state. A drawback of this approach is that once an SEEF potential has been constructed, improvements cannot be added easily without introducing overlaps in the underlying energies.

2.3.1 Empirical Energy Functions and Prediction Models

One of the early implementations of empirical energy functions for the predicting mutant stability changes were implemented by AGADIR and SPMP potentials followed by FOLDEF (energy function).

AGADIR method (Munoz and Serrano 1994) was created to analyse the stability of point mutations in α-helices using a helix-coil transition algorithm. Using an empirical analysis of experimental data the study estimated a set of energy contributions which accounts for the stability of isolated α-helices. With this database and an algorithm based on statistical mechanics, it describes the average helical behaviour in solution of 323 peptides and the helicity per residue of those peptides analysed by NMR. Moreover the algorithm

successfully detects the α-helical tendency, in solution, of a peptide corresponding to a β-strand of ubiquitin.

SPMP (Stability Profile of Mutant Protein) method (Takano et al. 1999) calculates $\Delta\Delta G_{SPMP}$, the predicted values of the experimental $\Delta\Delta G$. In this method, a pseudo-energy potential developed for evaluating structure-sequence compatibility in the structure prediction method was employed, consisting of four elements: side-chain packing, hydration, local conformation and hydrogen bonding efficiency of the backbone. The side-chain packing function is a Sippl-type pairwise function (Sippl 1990; 1993), considering the distance between the side-chains and the interacting directions, but not considering in detail the conformation of side-chains. The hydration function is based on the partitioning of the amino acid residue type into the surface or interior of a globular protein. The local conformation function is a potential estimated from the frequencies of an amino acid residue observed in a conformational state. The hydrogen-bonding efficiency of the backbone function is given to the pair of proton donors (oxygen atoms) and acceptors (nitrogen atoms) in the backbone atoms, depending on the preference on hydrogen bond formation between two amino acid residues. Nine lysozyme mutants were selected to verify the SPMP's reliability in predicting mutations. All these mutants had stabilising effects according to SPMP, but DSC studies suggested that only one out of the selected nine mutants had stabilising effect, and the others had either destabilizing or unaltered effects in the mutants. It was concluded that this empirical potential overestimates the increase in stability or underestimates negative effects due to substitution.

In FOLEDEF (Guerois et al. 2002), they used solvent exposure, van der Waals, solvation energies, hydrogen bonding efficiency, electrostatistics, backbone and side chain entropy terms, effect of water bridges to model the mutants and predict their stability. The solvent accessibility is estimated using the atomic occupancy method (Occ), which sums the volumes of the atoms j surrounding a given atom i. The van der Waals and the solvation energies were obtained from the free energy of transfer of the amino acids from vapour to water and from

organic solvents to water. Electrostatic energies are calculated between charged atoms of the N and C termini, and between the charged atoms of Asp, Glu, Arg, Lys and His residues only if they are closer than 20Å. The backbone entropy term is used to account for the entropy cost of fixing a residue backbone. The water bridge is defined as a water molecule that makes more than two hydrogen bonds with the protein. In the FOLDEF, the energy assigned to a water bridge interaction allows us to reduce the solvation penalty for buried polar atoms when they are involved in such an interaction. There were also some additional features taken into account for the predictions. In some structures of the protein database, van der Waals clashes are observed and can be due to the resolution of the structures. For structures with resolution lower than 2 A°, ΔG_{clash}, the free energy correction for a clash, is usually zero and does not exceed 1.0 kcal mol for one residue in a protein. N-caps of α-helices were also dealt carefully, which projects its involvement in stabilisation of a protein due to the water molecules bound to the NH terminal. These empirical energy functions showed a correlation coefficient of 0.75 between the experimental and predicted energy values for 1088 mutants. After the removal of outliers, the correlation coefficient improved to 0.83 for a dataset of 1030 mutants. For the training dataset of 339 mutants, it was observed to be 0.70 (Guerois et al. 2002). Only $\Delta\Delta GH_2O$ were used and no further validation tests were done.

Another method (Bordner and Abagyan 2004) used similar empirical energy functions with free energy contributions of hydrophobic effect and that of unfolded state in addition to hydrogen bonding, van der Waals forces, electrostatistics and conformational entropy. This study used a dataset of 1816 mutants only with $\Delta\Delta GH_2O$. A split sample validation with 908 selected mutants is used as training with a correlation of 0.79 and the remaining mutants were for validation with a covariance of 0.68. After the removal of 23 outliers, correlation increased to 0.82. No other validation tests were carried out.

2.3.2 Statistical Energy Functions

One of the earliest prediction models (Gilis and Rooman 1997) derived distance and torsion potentials using 10 proteins with mutations at the buried

and solvent accessible regions of protein. The correlation coefficient between the predicted and experimental ΔΔG was observed to be 0.80 and 0.67 for 121 buried (training) and 106 surface mutations (test) respectively.

Another group of investigators (Khatun et al. 2004) developed contact potentials and took 3 datasets of 2317 mutations totally from 13 proteins. Those contact potentials used a simplified model of amino acid interactions by approximating the potential energy of amino acid interactions, which is derived as a sum of two- and three-body interactions, together with the contribution to the protein potential energy from the solvation of amino acid residues. For a big dataset of 1356 mutations, the correlation was 0.66 and 0.46 during the training and testing of the split sample validation respectively. The correlation coefficient for the jack-knife test was 0.45. These results (correlation with test dataset and jack-knife with all mutants) were insufficient for the accuracy and transferability of the prediction model. So, they suggested the use of an atomistic form of potentials with ASA differences in wild type and mutant residues for future improvement of protein stability prediction upon point mutation.

Another study (Hoppe and Schomburg 2002) used similar statistical potentials with a training dataset of 546 mutations with a correlation of 0.75 and applied the parameters to a test dataset of 866 mutants with a correlation of 0.62. But, the ΔΔG and ΔΔGH2O values were mixed in the prediction system. Apart from the split sample validation, no other validation tests were carried out.

Recent methods that use statistical mechanics potentials (Zhou and Zhou 2002a) focus on distance-dependent, residue-specific, all-atom assumption. The common approximation (reference state) made by a standard contact potential is the approximation over all the amino acid distributions of the folded proteins from which the potentials were derived. This approximation has its origin in the "uniform density" reference state used by a previous study (Sippl 1990) to derive the residue-based, distance-dependent potential. In this approximation, the total number of pairs in any given distance shell for a reference state is the same as that for folded proteins. In other words, the distance dependence of the

pair probability distribution of the reference state is an averaged distribution over all residue or atomic pairs. This reference state is a non-interacting ideal-gas reference state only if the average interaction of all residue or atomic pairs is zero (i.e., attractive and repulsive interactions cancel each other). However, it is highly unlikely that attractive and repulsive interactions could cancel each other exactly. To explore these missing residual interactions, they established a non-interacting reference state without using the above mentioned assumption. This is done by using uniformly distributed non-interacting points in finite spheres. The reference state coupled with a simple distance scaling method employed to derive an all-atom potential of mean force from a structural training database of 1011 non redundant protein structures. They reported a correlation of 0.55 for 1023 mutants in 35 proteins. But, the mutations that have decreased number of atoms were only used to avoid strains associated small-to-large mutations (Zhou and Zhou 2002b).

2.3.3 Neural Networks

Caprioti and investigators developed two methods: a neural network based method and a support vector machine based method to predict protein stability changes upon mutation (Capriotti et al. 2004; 2005b). The neural network method was used to discriminate the stabilising and destabilizing mutations and has an accuracy of 80%. When coupled with empirical energy values of FOLDEF with known experimental pH and temperature conditions, the prediction can raise up to 90%. Though the experimental conditions are present for many mutations selected for their analysis, it becomes difficult to correlate and predict the experimental conditions of new mutations in site-directed mutagenesis and other similar methods.

2.3.4 Support Vector Machines

This method utilizes both the structure as well as the sequence information separately in two models for the prediction. When predicting $\Delta\Delta G$ values associated with the mutation, the correlation coefficient between predicted and

observed values was 0.71 and 0.62, depending on the structure- and sequence-based prediction, respectively (Capriotti et al. 2005b).

2.4 Application Note

2.4.1 PopMuSIC

PoPMuSiC (Gilis and Rooman 2000) is one of the early tools for rational computer-aided design of single-site mutations in proteins and peptides. It's based on the algorithm developed by Gillis and Rooman using the statistical potentials. Two types of queries can be submitted. The first option allows estimating the changes in folding free energy for specific point mutations given by the user. In the second option, all possible point mutations in a given protein or protein region are performed and the most stabilizing or destabilizing mutations, or the neutral mutations with respect to thermodynamic stability, are selected. For each sequence position or secondary structure the deviation from the most stable sequence is moreover evaluated, which helps to identify the most suitable sites for the introduction of mutations. It is optimized mostly for the human prion proteins and trained with less number of mutations that are insufficient for a prediction model. It's available from the URL below:

http://babylone.ulb.ac.be/popmusic/

2.4.2 Fold-X

Fold-X (Guerois et al. 2002) is based on the FOLDEX empirical energy functions developed by Guerois et al. as discussed in the previous section. It was tested using 1088 mutations with a correlation of 0.75. It is available from the URL below:

http://foldx.embl.de/

2.4.3 I-Mutant (version 1 and 2)

I-Mutant version 1.0 and 2.0 are developed using neural networks and support vector machines (Capriotti et al. 2004; 2005b). It is available from the URL below:

http://gpcr2.biocomp.unibo.it/~emidio/I-Mutant/I-Mutant.htm

2.4.4 DMutant

DMutant (Zhou and Zhou 2002a) was developed using statistical potentials using distance-dependent finite ideal gas reference state (DFire). It is based on the algorithm used by Zhou et al. and available from the URL below:

http://phyyz4.med.buffalo.edu/hzhou/mutation.html

3 MATERIALS AND METHODS

The entire prediction model for predicting changes in protein stability upon point mutations was developed computationally. No commercial software was used for the development process. DSSP was used for calculating secondary structure parameters. Web applications were used for generating structural training datasets.

3.1 Structural Training Datasets

In order to derive statistical mechanics potentials, a non-redundant (non-homologous) dataset of protein structures must be used. In this work, only the structures from X-ray crystallography were used from Protein Data Bank (PDB) and all other methods like NMR were avoided. NMR structures have multiple models for the same structures and results in inaccuracy of the statistical potentials. Old PDB structures with only Cα atoms were removed. There are several available algorithms from which these non-redundant protein structures can be derived. There are several variables which influence the construction of non-redundant datasets (Wang and Dunbrack 2003). These include:

(1) Maximum percentage identity between protein structures.

(2) Minimum resolution.

(3) Maximum resolution.

(4) Maximum R-value.

(5) Minimum and maximum chain lengths.

There are several ways with which the above variables can be adjusted before deriving the non-redundant datasets. In this work, protein structures ranging from 25% to 50% maximum sequence identity are used. There are several new structures with high resolution. Typically, the resolutions from 2 to 2.5Å were used in the datasets (Mintseris and Weng 2004). The minimum length was set to 40 residues to the maximum of 10,000 residues.

3.1.1 Selection

Selection criteria and the number of proteins in the non-redundant dataset initially depend on the method used to derive the structural training dataset. Different algorithms in the internet are slightly different in implementing non-redundancy check. These algorithms and their web serves are listed below:

(1) PISCES

PISCES (Wang and Dunbrack 2003) (Protein Sequence Culling Server) is a public server for culling sets of protein sequences from the Protein Data Bank (PDB) by sequence identity and structural quality criteria. PISCES can provide lists culled from the entire PDB or from lists of PDB entries or chains provided by the user. The sequence identities are obtained from PSI-BLAST alignments with position-specific substitution matrices derived from the non-redundant protein sequence database. PISCES therefore provides better lists than servers that use BLAST, which is unable to identify many relationships below 40% sequence identity and often overestimates sequence identity by aligning only well-conserved fragments. PDB sequences are updated weekly. PISCES can also cull non-PDB sequences provided by the user as a list of GenBank identifiers, a FASTA format file, or BLAST/PSI-BLAST output.

(2) SCOP – ASTRAL

ASTRAL web server (Chandonia et al. 2004) uses SCOP database of protein domain for its selection of non-redundant list of protein structures using AEROSPACI (Aberrant Entry Re-Ordered SPACI scores) scores. This algorithm uses a method for rapid multiple sequence alignment based on fast Fourier transform. The non X-ray structures are not automatically removed. So, separate custom filters were used to select only X-ray structures. The AEROSPACI scores are derived from SPACI (Summary PDB ASTRAL Check Index) scores which incorporates three different quantities: resolution of the structure, R-factor and stereochemical check parameters that indicate how well the model complies with the standard molecular geometry. The protein

structures with 50% of maximum sequence identity has been selected and used for torsion angle potentials for comparison.

(3) TOP500

A list of 500 proteins were compiled for Ramachandran plot distributions (Lovell et al. 2003), and this can be used for deriving torsion angle potentials for the main torsion angles (ϕ and ψ) that were used for the prediction model. It was slated towards the usage of less number of proteins, which were enough to assess the torsion angle distribution in proteins and minimising the noise in distribution function. This list was compiled with proteins with higher resolution (1.8Å or better). Clash scores (for atoms B<40) were observed to be less than 22/1000 atoms with fewer than 10/1000 atoms whose main chain bond angles (including to C_β) having standard devation more than 5 (Engh and Huber 1991). Structures that contain unusual amino acids with main chain substitutions are avoided (e.g., 1mroA, 1rtu). No free atom refinements are included. Wild type is preferred over mutant. If proteins, related but not same, the ones with best combination of resolution and clash score are taken. If a dataset with 500 structures is enough for the torsion angle distribution, the noise rendered by other structures can be greatly reduced. Conversely, this dataset of 500 proteins may not include the torsion angle distribution of some structures.

3.1.2 Filters

The energy functions are basically used to determine the most favourable energy values contributed by the amino acids for native protein structures. However, the values contributed by amino acids are greatly influenced by other factors. These include:

(1) Proteins containing heavy metal ions.

(2) RNA binding proteins (protein-RNA complexes).

(3) DNA binding proteins (protein-DNA complexes).

(4) Virus coating proteins.

(5) Co-factor complexes and prosthetic groups.

(6) Membrane proteins.

(7) Transcription factors.

The heavy metals contain highly electrostatic charges and stabilise the protein native structure by keeping it intact. The same is observed when proteins bind to nucleic acids (DNA or RNA). The structure of the complex is stabilised not only by the amino acid sequence/structure features but also by the free energy contribution of nucleotides. The transmembrane proteins also have structure stabilised by external factors, where the membrane keeps the native structure intact. Including these proteins in structural training dataset for the energy functions will generate noise for the predictions. So, these structures (Table 4) are filtered out from the initial non-redundant dataset. Filtering of these structures was made easy with PDB's utility known as "PDB at a glance". It has been hosted in the NIH server given below.

http://cmm.info.nih.gov/modeling/pdb_at_a_glance.html

Proteins Filtered Out of Non-redundant Dataset:	
Membrane proteins	Bacteriochlorophyll-A, Bacteriorhodopsin, G Proteins, Hemolysin, Porin/Phosphoporin, Reaction Center, Transducin, Vitelline
Virus Coating proteins	Hemagglutinin, HIV Molecules, Inovirus Proteins, Papillomavirus Proteins, Rhinovirus Proteins
Cofactor complexes	AMP-Bound, ADP-Bound, ATP-Bound, NAD-Bound, NADH-Bound, NADP-Bound, NADPH-Bound
Nucleic Acid complexes	Protein-RNA complexes, protein-DNA complexes
Heavy Metals (found in PDB entries):	
Magnesium, Copper, Zinc, Iron, Molybdenum, Manganese	

Table 4: List of proteins that are filtered out of structural training datasets to reduce the noise in statistical potentials.

3.2 Mutation Datasets

Amino acid single mutations were taken from Protherm database (Bava et al. 2004) and literature (Alber et al. 1987; Yutani et al. 1987; Shih et al. 1995; Shoichet et al. 1995; Topham et al. 1997; Xu et al. 1998) whose stability relative to the wild-type ($\Delta\Delta G$ or $\Delta\Delta GH_2O$) were determined experimentally.

Mutants range between the core and periphery with highly variable solvent accessibility and secondary structure specificity (given as supplementary material). At the same time, the proteins also vary widely in their sequence identity and functional aspects.

3.3 Statistical Potentials

Two versions of statistical potentials were derived for the prediction model: Firstly, distance dependent pair potentials were extracted from the atom distribution using a radial pair distribution function. Secondly, torsion angle potentials were derived from the distribution of main torsion angles ϕ and ψ. These potentials were then unified using linear regression methods to construct the prediction model.

3.4 Distance Dependent Pair Potential

The basic statistical mechanics setup include mean force potentials that are established using radial distribution of 40 atom types (Melo and Feytmans 1997; 1998) and main torsion angles of amino acids. The atomic level organisation of potentials based on the radial distribution is an extended version of conventional amino acid potentials and exhibits a wide coverage of local and non-local interactions, and hence benefits with an evolution of accuracy in predictions. In addition, the data extracted from torsion angles also help improving the above predictions. The structural training dataset that initially furnishes the information for the extraction of these potentials consists of a dataset of 4024 non-redundant protein structures extracted from a recent PDB repository using the PISCES algorithm with 50% sequence identity and resolution less than 2.5Å.

3.4.1 Radial Distribution of atoms

The energy functions are predominantly derived from mean force potentials (Sippl 1993) based on the inverse Boltzmann's principle which essentially states that probability densities and energies are closely related quantities. Thus, the radial pair distribution function ($\Delta G_{ij}(r_d)$) has been derived:

$$\Delta G_a = -kT \ln\left(\frac{g_{ij}(r_d)}{g(r_d)}\right) \qquad (11)$$

where $g_{ij}(r_d)$ is the radial pair distribution function of a pair *i, j* separated by a distance r_d. $g(r_d)$ is the description of the reference state. The distribution of all 40 heavy atoms is taken with the radial coverage of 2.5-20Å and bin size of 0.5Å for the mean force potentials. Though different groups have tested various forms of reference states (Betancourt and Thirumalai 1999; Pandey et al. 2005; Ruvinsky and Kozintsev 2005), we used the standard method in which it is calculated as the approximation over all the amino acids together (Sippl 1990; 1993).

3.4.2 Distance Cutoff

The atom potentials were derived around the central amino acids with minimum and maximum cutoff distances of 2.5Å and 20Å respectively. Some atoms are observed to be below 3Å, especially in the cases of CIS prolines. Besides, most of the amino acids that are present in the loop regions are observed to have long range interactions, since there is considerably a high population of atoms in the distances between 18Å and 19.5Å. Thus, the distance cutoff values for the pair potentials are judged.

3.4.3 Atom Classification Models (Atom Types)

Different atom classification schemes were used for the atoms distributed around any central amino acid. These schemes include:

(1) Basic organic atoms ($C_{aromatic}$, $C_{aliphatic}$, N, O, S).

(2) Amino acid C_α atoms (as a central point of residues).

(3) Li-Nussinov atom model: 24 amino acid atoms.

(4) SATIS model: 28 amino acid atoms.

(5) Melo-Feytmans atom model: 40 amino acid atoms.

All these atom models were compared and the best prediction model has been established that uses the optimised combination of the selected atom model.

These atom classification models were analysed and the validity of their use in protein stability predictions is listed:

(1) Basic Organic Atoms (5 atoms):

The simplest way of classifying atoms (Hoppe and Schomburg 2002) distributed around the central amino acid is to classify atoms using the presence of basic organic atoms of the amino acids. This includes carbon, nitrogen and oxygen. Since the behaviour of aromatic carbon atoms is different from the aliphatic carbon atoms, they are classified separately. Though this description of the atom model does not explain the complicated structural or functional features of a protein structure, minimal coverage of interactions and conservation of the atom distribution are covered to certain extent. This model was useful for the predictions which used less non-redundant protein structures during early periods of Protein Data Bank.

(2) Amino acid C_α atoms (20 atoms):

One of the classical ways of describing the distribution of atoms around the central amino acid is to consider the C_α atoms of amino acids alone for deriving the pair potentials. Here, any distributed C_α atom acts as the centre of interactions exhibited of all atoms of that specific amino acid. Since the amino acids can significantly explain structural and functional role of protein structures, they have established good prediction models in several cases of protein structure prediction (Melo et al. 2002).

(3) Li-Nussinov atom model (24 atoms):

Li and Nussinov classified the atom distribution into 24 amino acid atom types (Li and Nussinov 1998). The 25^{th} atom type is assigned for H_2O. In the present work, the water molecules were excluded from the atom definition and only 24 amino acid atom types were used for developing the prediction system. The classification criterion in this model is according to the number of hydrogen bonding.

The first 14 types were classified as carbon and sulphur atoms with a varying number of bonded hydrogens and/or different covalent bonding environments. These atoms can be considered as apolar or hydrophobic. The last 11 types were classified as nitrogen and oxygen atoms, with a varying number of bonded hydrogens that are either polar or charged. The placement of the boundary between polar and apolar is somewhat arbitrary. Some of the apolar atoms, like the carbon atoms that are covalently bonded to either polar or charged atoms, may have substantial polar character. Some atoms were not classified into any of the 25 types in table 5. Those were Cz of Arg, Cg of His, and N of Pro, since the number of these atoms in the dataset was too small for each of them to be considered as a separate type and their effect would be negligible for the prediction model.

(4) SATIS atom model (28 atoms):

SATIS (Simple Atom Type Information System) is a protocol (Mitchell et al. 1999) for the definition and automatic assignment of atom types and the classification of atoms according to their covalent connectivity. Its distinctive feature is that no bond type information is involved. Rather, the classification of each atom is based on a connectivity code describing the atom and its covalent partners. It is particularly useful when handling coordinate-based molecular representations with no bond order information, such as the PDB format.

This model seeks a method of categorising and indexing atoms with a connectivity code, which depends only on the identities of their covalently bonded partners and being independent of any subjective definitions, either of functional groups or of bond orders. In this atom type definition a set of connectivity codes was defined that is dependent only on the atomic number of an atom and on the number and identity of its bonded partners. Thus, there was no subjectivity in the assignment of the connectivity codes, except in those rare cases where the existence of covalent bonds is open to dispute.

It has been reported that when applying SATIS to atom typing, either each connectivity code can be used as an atom type in its own right or atom types can be defined as sets of (one or more) connectivity codes describing chemically similar atoms. In principle, any computational representation of chemical structure can be used to generate connectivity codes automatically and SATIS converts these different computational representations of chemical structure into connectivity codes. The connectivity information for each atom was formulated as 10-digit connectivity code. The first two digits are the atom's atomic number (e.g., 06 for carbon or 16 for sulphur). The remainder of the code consists of four two-digit numbers, representing the atomic numbers of the atom's covalently bonded partners in ascending numerical order. If an atom has fewer than four bonded partners, the remaining positions in the connectivity code were filled with 99. This classification of the atom type definitions used in this work is listed in the table 5.

This method provides a simple scheme for categorising the atoms found in any covalently bonded molecule. It can be extended to the definition of atom types for either potential energy functions or analysis of spatial distribution. The analysis of the atoms found in the 20 common amino acid residues shows that SATIS automatically implements a classification scheme comparable with others devised for these atoms. One may choose to refine the scheme by grouping some connectivity codes together into atom types. SATIS is applicable to all covalently bonded atoms, so the possible problem of having no relevant atom type defined for unusual covalent connectivity is avoided. An advantage is that the connectivity codes do not depend on a subjective assessment of bond orders. Hydrogens were also particularly well classified. But, they were not considered separately as an individual atom type.

(5) Melo-Feytmans atom model (40 atom types):

Melo and Feytmans used 40 distinct atom types for their potential of mean force describing interactions within proteins. They used a mean force potential at atomic level, which is grounded on a particular definition of atom types. In total there are 167 heavy atoms in all the 20 existing amino acids, and they were classified into 40

different atom types (Fig. 3). From a physico-chemical point of view, all the atoms would be different with different environments and the atom type definition is based on its connectivity, chemical nature and location level (side chain or backbone).

Li-Nussinov Atom Types (24 types)		SATIS atom types (28 types)	
Number	**Atom Symbol**	**Number**	**MF Atom types**
1	CA	1	5, 34
2	C	2	4, 33
3	CH	3	3, 25, 36, 39
4	CH2	4	8
5	CH2b	5	1
6	CH2ch	6	12
7	CH3	7	6
8	Char	8	28
9	Car	9	2, 32, 35, 37
10	CHim	10	11, 13
11	Cco	11	18, 22
12	Ccoo	12	7
13	SH	13	16, 40
14	S	14	29
15	N	15	24
16	NH	16	14, 23
17	NH+	17	27
18	NH2	18	30
19	NH2+	19	15
20	NH3+	20	17
21	O	21	26
22	Oco	22	31
23	Ocoo	23	21
24	OH	24	20
		25	10
		26	38
		27	19
		28	9

Table 5: The Li-Nussinon amino acid atom types (LN24) and SATIS amino acid atom types. SATIS atom types are cross-referred with 40 atoms of MF40 atom classification model (Fig. 2).

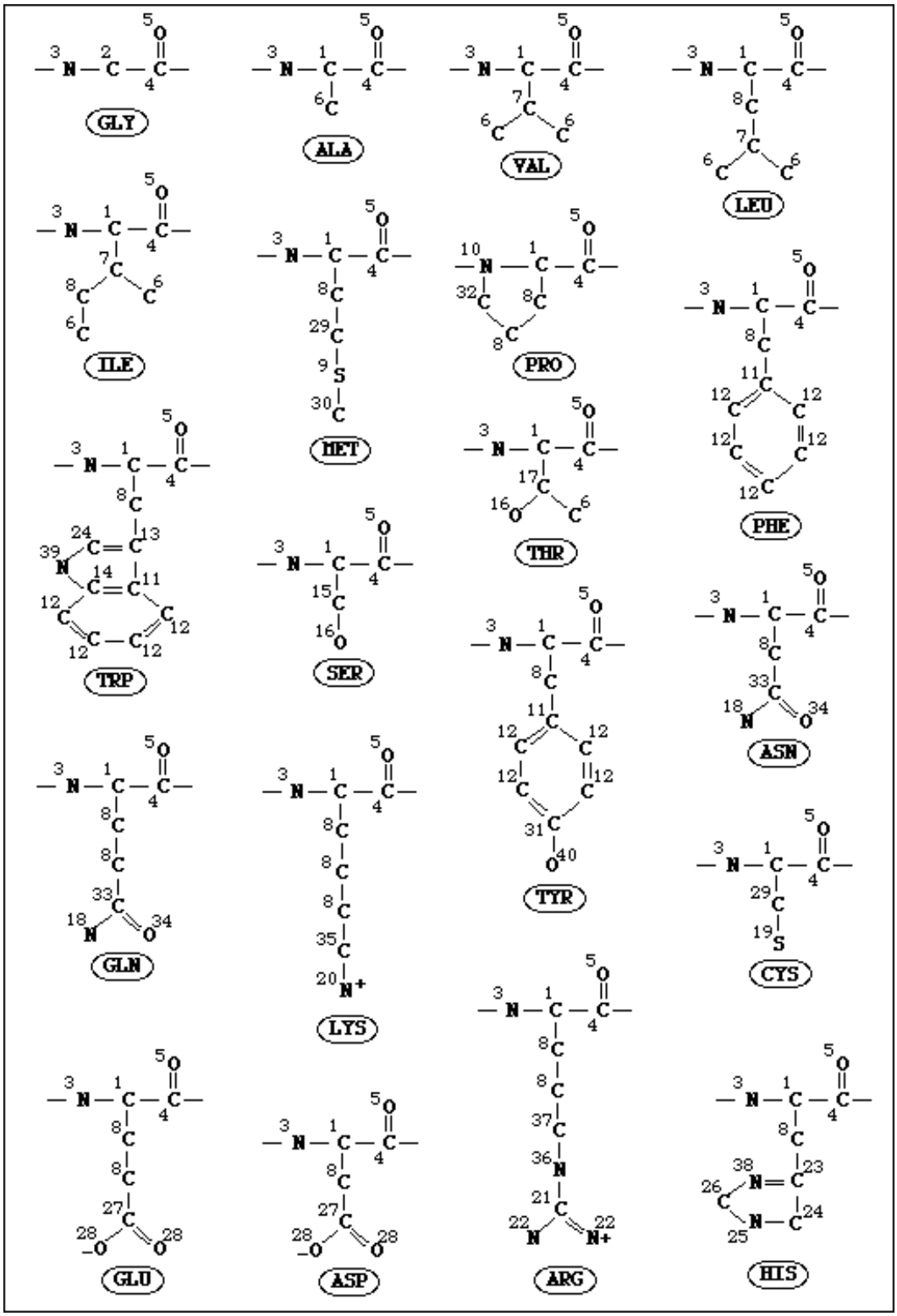


Fig. 2: Melo-Feytmans atom classification model (MF40). Amino acid atoms were classified into 40 types according to their location, covalent connectivity and chemical nature.

3.5 Torsion Angle Potential

3.5.1 Basic Construction

The same dataset of 4024 non-redundant structures was used to derive the torsion angles ϕ and ψ, after running DSSP for the whole dataset. The 'top500' was also used for the comparison of the efficiency of the torsion angle potential between the two different datasets. The minimum bin size for the torsion angles was set to 1^o comprising the bins ranging from -180 to 180 for both the torsion angles. Before the potential was developed, the torsion angle bins were initialised with a constant to avoid null values for the development of Boltzmann energy values. Then, the bins were normalised with a standard procedure using the circular Gaussian function for ϕ and ψ having the bivariate normal distribution (Niefind and Schomburg 1991):

$$f(\phi,\psi) = \frac{1}{2\pi\sigma^2}.A(\phi,\psi) \qquad (12)$$

Here, σ is the standard deviation and $A(\phi,\psi)$ is the Gaussian apodisation function for the torsion angles ϕ and ψ where the distribution of torsion angle potential is tapered around the peaks to accommodate torsion angle perturbation in the mutants.

$$A(\phi,\psi) = e^{-\left(\frac{(\phi-\mu_\phi)^2+(\psi-\mu_\psi)^2}{2\sigma^2}\right)} \qquad (13)$$

The torsion angle count exhibits different frequencies and the population of angles bins differ from one amino acid to other. In order to avoid this problem, the torsion angle bins for all 20 amino acids were further normalised individually for the angles ϕ and ψ with a scaling factor H satisfying the following condition:

$$\frac{1}{H}\sum_{i=1}^{n} f(\phi_i,\psi_i) = 10^2 \qquad (14)$$

The normalised torsion angle distribution was then used to derive the Boltzmann energy values for mean force potentials of all amino acids individually with and without their classification based on accessible surface area and secondary structure specificity.

$$\Delta G_{tor} = -kT \ln\left(\frac{g(\phi_i,\psi_i)}{g_{ref}(\phi,\psi)}\right) \qquad (15)$$

Here, $g(\phi_i,\psi_i)$ and $g_{ref}(\phi,\psi)$ are the normalised torsion angle distribution of a specific amino acid and the average distribution over all the amino acids respectively.

3.5.2 Optimisation

Several parameters should be optimised for the torsion angle potential. Before deriving the torsion angle distribution, the angle bins are initialised with 0.001 (1/100) for all ϕ and ψ combinations. Other initialisation variables can also be used: $1/360^2$ for all ϕ and ψ combinations (Dönitz 2001). Instead, the total number of amino acids (n) in the structural training dataset can also be replaced with 360^2 to depict the amino acid specific initialisation scores (Dengler et al. 1997; Dengler 1998) for the torsion angle bins.

Apodisation is carried out by the Gaussian function, though the other variants (Blackman, Hamming or Connes apodisation functions) can also be used to render the tapering of torsion angle distribution. When large numbers of protein structures are used from the structural training datasets, torsion angle distribution is observed accurately by having enough counts in many torsion angle bins. This also results in increased noise for the Gaussian apodisation by having edge effect, where the tapering of two or more adjacent peaks in the distribution result in clashes between themselves in the edges. To reduce this phenomenon to a considerable extent, the maximum values of μ_ϕ and μ_ψ were optimised accordingly so that the effect is minimised. A maximum angle of 10° for μ_ϕ and μ_ψ was used initially which allows the Gaussian function to normalise more than 400 combinations of ϕ and ψ totally. Later, the maximum

angle was reduced to 7° to minimise the edge effect. In this case, around 200 combinations of ϕ and ψ were normalised. The energy distribution curves were compared to visualise the difference.

3.6 Protein Environment Specificity

Mean force potentials are usually derived using the common approximation over all the amino acids from a selected list of non-redundant protein structures. This method has a long standing history of accuracy for many statistical mechanics based prediction models for protein structure predictions and many other cases where the involvement of mean force potentials can be applied. However, this would be optimal for relatively smaller amount of structures. But, the models that use protein structures of highly variable sequence, structural and functional diversity demand an increased accuracy from the statistical potentials. For this reason, the generic model that makes no distinction between the amino acid environments can be dissected using the physical features that prevail in molecular structures. These features are anticipated to distinguish the regions within a protein structure. Besides, they can also logically associate the similar features of multiple protein structures. For this work, two main characteristics that are known to dominate protein structures are considered: amino acid compactness and secondary structure specificity.

3.6.1 Amino Acid Compactness

(1) Solvent Accessibility

Solvent accessibility of the amino acids have been used in measuring compactness in many cases, and proved to be one of the most important features for protein structure and stability. Accessible surface area (ASA) of amino acids was used to determine the solvent accessibility. For this, constant values of ASA, when they exist in ALA-X-ALA extended state, were used for all 20 amino acids (Appendix C) to derive the relative ASA of amino acids in protein structures. The relative ASA is determined by dividing the observed

ASA for a given residue by the constant ASA of that residue in Ala-X-Ala extended state, and multiplied by 100 to obtain a percentage (eqn. 16).

$$Relative\ ASA = \frac{observed\ ASA}{const.\ ASA} \times 100 \qquad (16)$$

(2) Packing of C_α or C_β atoms

Alternatively, the packing of the backbone C_α atoms can be used for measuring compactness of the amino acid. However, C_β atoms can also be used instead. As in the previous case, constant values are determined for all 20 amino acids by observing the maximum possible distribution of C_α atoms around the 20 amino acids over all the proteins available in the structural training dataset. Once the values are determined, relative packing of atoms was calculated in percentage (eqn. (17)) for all the amino acids in the dataset which will later help distinguishing them in statistical potentials.

$$Relative\ packing = \frac{obs.\ atoms}{max.\ atoms} \times 100 \qquad (17)$$

3.6.2 Secondary Structure Specificity

Secondary structure specificity of amino acids was used together with their compactness ratio for distinguishing the amino acids for statistical potentials. For this, the amino acids that belong helices (α-helices, 3_{10}-helices, 5-helices) and sheets (isolated β-bridges, β-ladders) are grouped into two groups respectively. All the other amino acids were classified into a third group. This method is kept simple, even though the different helices, sheets, coils or loops can further be broken down into separate groups. Especially, the parallel and anti-parallel beta sheets' radial pair distribution is different from each other. But, they were still maintained in the same group, since it was believed that the inclusion of torsion angle potential can demarcate their characteristic features.

To distinguish the structural regions, the secondary structure specificity and solvent accessibility were used to classify the amino acids of point mutations

and proteins from structural training dataset. Amino acid specific statistical potentials were then derived for the structural regions separately and used for the respective point mutations that were present in the same regions. Values of accessible surface area are flexibly used to classify the mutations.

Several classification methods were used to optimise the number of mutations and their subsequent prediction efficiency. The methods are given in table 6a. Initially, the mutations were classified different structural regions using the secondary structure (helices, sheets and others). Later, the mutations were classified using the ASA of the amino acids. In table 6a, the numbers indicate the total number of structural regions classified using ASA in a specific secondary structure element. In CL9 (Table 6b), 9 different structural regions were defined. Then, an extended classification method was used where 12 structural regions were defined. In CL12A_1 or CL12_B (Table 6c), ASA range was further divided to extend the CL9 method. Conversely, in CL12B_1 and CL12B_2 (Table 6d), the secondary structure specificity was extended by classifying the 'turns' as a separate structural region. In spite of not having enough point mutations in turns, these classification methods were used for the purpose of comparison.

SS / ASA	**CL9**	**CL12A_1**	**CL12A_2**	**CL12B_1**	**CL12B_2**	**CL11**
Helices	3	4	4	3	3	4
Sheets	3	4	4	3	3	3
Turns	X	X	X	3	3	X
Others	3	4	4	3	3	4
Bins	9	12	12	12	12	11

(a)

	CL9		
SS/ASA	0 – 2	2 – 50	50 +
Helices	1	4	7
Sheets	2	5	8
Others	3	6	9

(b)

	CL12A_1				**CL12A_2**			
SS/ASA	0 – 2	2 – 40	40 – 70	70 +	0 – 2	2 – 30	30 – 60	60 +
Helices	1	4	7	10	1	4	7	10
Sheets	2	5	8	11	2	5	8	11

Others	3	6	9	12	3	6	9	12

(c)

	CL12B_1			CL12B_2		
SS/ASA	0 – 2	2 – 50	50 +	0 – 2	2 – 50	50 +
Helices	1	5	9	1	5	9
Sheets	2	6	10	2	6	10
Turns	3	7	11	3	7	11
Others	4	8	12	4	8	12

(d)

Table 6: (a) Classification of structural regions using various methods for amino acids in structural training datasets and mutation datasets. (b) CL9 method involves 9 structural regions. (c)(d) CL12A and CL12B methods involve 12 structural regions using ASA (Accessible Surface Area) and SS (secondary structure) specificity.

3.7 Statistical Methods

Theoretically derived energy values of all the 40 atom types and torsion angles were used as independent variables for regression with experimental energy values and the prediction equation was derived. The analysis is initially carried out individually and then classified into 12 different potentials based on solvent accessibility and secondary structure and the equations were derived separately. A linear model is assumed to conduct multiple and stepwise linear regression between the experimental energy values and those of atoms and torsion angles.

3.7.1 Simple Linear Regression

Simple linear regression (SLR) or the unweighted linear regression gave less correlation with the experimental $\Delta\Delta G$ from thermal denaturation experiments. For 159 T4 lysozyme mutants (Topham et al. 1997), unweighted linear regression gave a correlation coefficient ranging from 0.36 to 0.64 depending on different methods used to derive the equation. For the same mutants, robust weighted linear regression gave a correlation coefficient ranging 0.56 to 0.77. This is evident from the fact that the empirical (hydrogen bonds, electrostatistics, etc.) or statistical potentials (atom potentials, torsion potentials, etc.) have relatively higher or lower impact (weights) between each

other. The coarse grained statistical atom potentials can even assign weights to specific atom types to make the linear regression robust. However, simple linear regression was carried to cross check the results with other linear regression models.

3.7.2 Multiple Linear Regression

Multiple linear regression model was implemented using atom and torsion potentials together. Here, the regression coefficients are derived from the variance observed in the experimental $\Delta\Delta G$ or $\Delta\Delta GH_2O$. Models for these two experimental values were developed separately, since $\Delta\Delta G$ and $\Delta\Delta GH_2O$ are from thermal and chemical denaturation experiments respectively, and should not be mixed.

In the multiple regression model, atom types were used as independent variables, and regressed against experimental $\Delta\Delta G$ or $\Delta\Delta GH_2O$ as dependent variable. Since the pair distribution is classified using ASA and secondary structure specificity of central amino acids, separate regression models were developed for each of them. Initially, the secondary structure specificity was used and three different groups were obtained as explained previously. But, to extend these three groups using compactness, the relative ASA is used flexibly depending on the number of mutations present in different solvent accessible regions of protein secondary structures. This is because the multiple regression model requires enough variables to be present in each groups so that there will not be a problem of overfitting of variables with the prediction model. If there are many independent variables (many atoms, torsion angle) employed for multiple regression, the ability to explain the amount of variance that exist in experimental $\Delta\Delta G$ or $\Delta\Delta GH_2O$ can be maximised. Thus, it results in high values of R^2 (covariance) or R (correlation). But, when more variables are used for the prediction model, the ability to test the model for its reliability (validation tests) is minimised which eventually leads to poor probability of getting the same correlation accuracy for new mutations that will be predicted by the model in future. To reduce this problem, further statistical analyses were carried out to test the efficiency and reliability of atom and torsion potentials.

This results in using robust multiple linear regression model (eqn. (18)) where the influence of atom potentials and torsion potential can be dynamically added by regression coefficients.

$$\Delta\Delta G = b_0 + \sum_{a=1}^{a=n} b_a.\Delta\Delta G_a + b_{tor}.\Delta\Delta G_{tor} \qquad (18)$$

Here, n is the number of atoms taken for the prediction model which is initially 40 and reduced using either colinearity diagnostics or stepwise regression. Predicted stabilising energy values $\Delta\Delta G_a$ and $\Delta\Delta Gtor$ from atom potentials and torsion angle potentials respectively were added together after multiplying with appropriate regression coefficients to derive the final predicted stabilising energy values.

3.7.3 Multicolinearity Diagnostics

Multicolinearity diagnostics determine the inter-relationships between independent variables, where the atoms and torsion angles are analysed and their relationships are studied. With variable levels of colinearity diagnostics, it was essential to analyse the influence of atoms that were highly correlated to one or many of the other atom types. Colinearity diagnostics were done statistically using the Variance Inflation Factor (VIF) which is the inverse of tolerance (eqn. (19)). VIF was derived for all the atoms separately where *n-1* atoms were taken as predictors and regressed with the remaining atom type to diagnose its correlation with all other atom types. Atoms that showed a specific VIF cutoff were selected and their distribution was unified and used in multiple regression as a single distribution.

$$Tol = 1 - R^2 \; (VIF = 1/Tol) \qquad (19)$$

Here, R^2 is the coefficient of determination (squared correlation coefficient) between the atom i with all other atoms together. Though the atoms with a statistical *VIF* cutoff of more than 10 represent colinearity in the model, various cutoff values of *VIF* such as 20, 30 and 40 were used and the results were tested. If colinearity is detected in the model, atoms that represent a VIF of more than the selected VIF cutoff are considered to be highly correlated with

other atom types and become eligible for unification of their distribution. The selected VIF cutoff plays a major role in the selection of correlated atoms where more atoms were selected for unification with decreasing *VIF* cutoff, while the distribution of all other atoms were unaltered and used together with this unified distribution.

3.7.4 Stepwise Linear Regression

Stepwise linear regression was also performed based on the forward and backward selection methods to detect and analyse the atoms with high and low influence. The atoms are dynamically selected using the statistical significance (*p)* values. This is performed separately for the 12 different datasets of CL12A and CL12B classification system, since the radial distribution may be different for the structurally variable regions in proteins.

The potentials were extracted separately from the proteins using the ASA and SS based classifier. Thus, the potentials were classified into 12 different types and the Boltzmann energy values were derived individually. The distribution of 40 atoms and the torsion angles were used to derive the stabilisation energy values which were derived by fitting them with experimental stabilisation energy using the linear regression models.

Statistical significance (*p*) was set to be 0.05, which allows the 95% confidence interval. Statistically, *p* values exhibit the probability of explaining the difference between the mutations. This can also be explained as the ability to successfully distinguish the available mutations for the prediction model. Here, p value of 0.05 means that the model will be able to explain 95% of variance exhibited by the selected mutations.

3.7.5 Final Prediction Model

The final prediction model used the stepwise linear regression to predict the mutations. Overall prediction efficiency was calculated for the two versions of models that use $\Delta\Delta G$ from thermal and chemical denaturation experiments respectively.

3.7.6 Assessment of Overall Prediction Efficiency

(1) Correlation

Pearson's correlation coefficient has been calculated for the predicted and experimental values of ΔΔG and $\Delta\Delta GH_2O$ separately through two different prediction models to distinguish the thermal and denaturant denaturation experiments respectively. Initially, the correlation coefficient has been calculated for all the mutations selected from different regions of proteins so that the maximum prediction efficiency of the model can be observed. Additionally, the correlation coefficient has been calculated for all the validation tests.

(2) Prediction Accuracy

Predicted accuracy depicts the amount of mutations to be correctly predicted as stabilising or destabilising. This is also calculated as a percentage value for the models based on signs of $\Delta\Delta G$ and $\Delta\Delta GH_2O$.

(3) Sensitivity

Sensitivity of the prediction model depicts its ability to correctly identify the stabilising (negative values of ΔΔG or $\Delta\Delta GH_2O$) mutations. Statistically, sensitivity is given in the equation 20. Higher values of sensitivity reflect in smaller number of mutations detected as false negatives.

$$Sensitivity = \left(\frac{TP}{TP + FN} \right) \qquad (20)$$

(4) Specificity

Specificity of the prediction model depicts its ability to correctly identify the destabilising (positive values of ΔΔG or $\Delta\Delta GH_2O$) mutations. Statistically, specificity is given in the equation 21. Higher values of sensitivity reflect in smaller number mutations detected as false positives.

$$Specificity = \left(\frac{TN}{TN + FP} \right) \qquad (21)$$

(5) Standard Error

Standard error (σ_{est}) is a statistical measure of the accuracy of the predictions made with a regression line. In the prediction model, the standard error is observed for the linear regression fit between experimental and predicted $\Delta\Delta G$. Standard error is calculated using the equation 22. Here, Y and Y' are the experimental and predicted $\Delta\Delta G$ respectively and $(Y- Y')$ is the error of prediction.

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}} \qquad (22)$$

3.7.7 Validation of Prediction Model

(1) Split-sample validation

The most commonly used validation method during generalisation of prediction equations is to split the mutations into two sets: training and test. Selected mutation data was used for training the model, while a unique test set was created with remaining mutations that could act as a representative set for training set mutations. Both datasets consist of proteins with highly variable sequence identity and the selection criteria for training and test datasets is similar to solvent accessibility and secondary structure based classifier. After training the model, the regression coefficients were applied to test, and the observed error is used to interpret an unbiased estimate of generalisation. The disadvantage of split-sample validation was that it minimised the availability of mutations for training and validation, so the statistical model was highly optimised to handle all validation tests.

(2) k-fold cross-validation

In k-fold cross validation, the mutation dataset was divided into k subsets of same size approximately. Each k-1 subsets were used for training and the remaining subset is used to compute prediction and error statistics for generalising the model equation.

(3) Jack-knife Test and Outliers

Jack-knife test is used to estimate the accuracy or the bias of the statistic. Here two datasets were developed: When N is the total number of mutations, one training dataset consisting of *N-1* mutations and a validation test of the remaining 1 mutation are taken and the same process is repeated for all the single mutations. Some prediction statistics were developed based on the above process and compared with that of the all mutations together.

4 RESULTS AND DISCUSSION

The structural training datasets were obtained from PISCES server and used to derive the radial pair distribution of atoms (amino acid environment) around all the amino acids present in the dataset. Several atom classification models were involved to classify the radial pair distribution of atoms. Then, the boltzmann's energy values were derived for 20 amino acids averaging the amino acid environments. Later, these energy values were applied for amino acid environments to be mutated. Stabilisation energy values were calculated from these Boltzmann's energy values for all the mutations in the mutation dataset. Torsion potentials were also calculated individually from the same structural training datasets. Statistical models with assumed linear relationship with experimental energy values ($\Delta\Delta G$) were used to develop the linear regression methods. Atom and torsion potentials were then unified using these linear regression methods to construct the prediction model that can predict the $\Delta\Delta G$. Experimental and predicted $\Delta\Delta G$ values were compared to analyse the statistics of prediction efficiency. Furthermore, various validation tests for the prediction model were also carried out to ensure its reliability.

4.1 Construction of Statistical Potentials

4.1.1 Structural Training Datasets

Structural training datasets are used to derive the mean force potentials from a list of protein structures. The details for the PISCES datasets are given in the table 7. PISCES dataset from Jan 2004 with 50% maximum sequence identity was used for all prediction models. For the purpose of comparison between the datasets, different datasets with sequence identity ranging from 25% to 50% (Aug 2005) was then used and the influence of the structural training datasets were compared.

Datasets	Date	Sequence identity	Selected PDB Chains	Filtered Proteins	Final List of Chains
PI-1	Aug 2005	25%	2387	515	1872
PI-2	Aug 2005	30%	2828	613	2215
PI-3	Aug 2005	35%	3201	714	2495
PI-4	Aug 2005	40%	3535	780	2755
PI-5	Aug 2005	45%	3761	833	2928
PI-6	Aug 2005	50%	3993	883	3110
PI-7	Jan 2004	50%	4127	104	4023

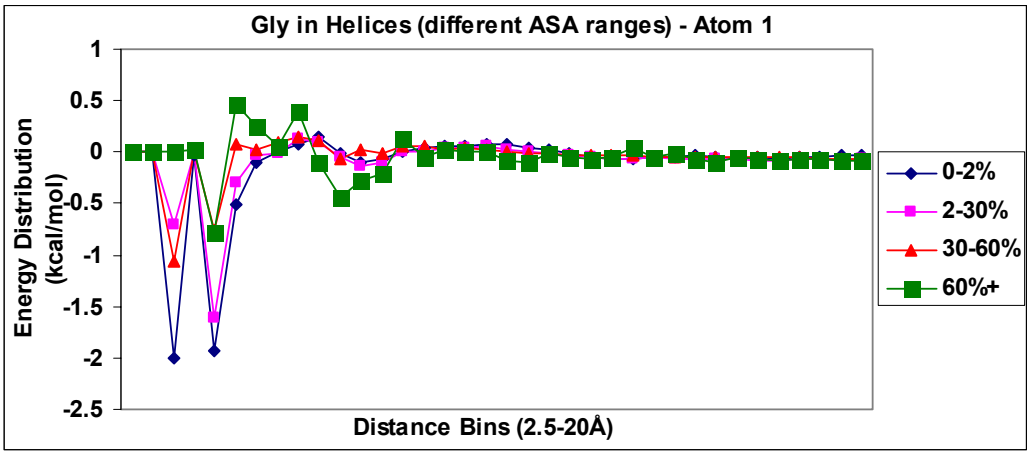
Table 7: Selection Criteria: All non-redundant datasets were derived with R-factor 0.3, and sequence chain length of 40 to 10,000. For PI-7, the resolution cutoff was 2.5Å. For other datasets, the resolution was set at 2Å. Non-X-ray entries and C_α-only entries were excluded from the dataset. Chain-wise selection was performed. PI-7 dataset was used for almost all prediction models. Other datasets were only used for the purpose of comparison.

4.1.2 Distance Dependent Pair Potential

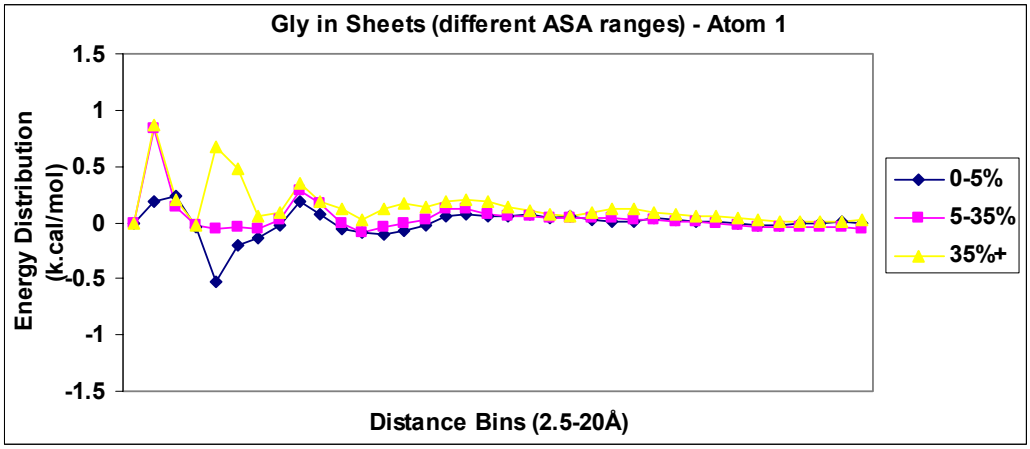
The pair potentials were derived according to the methods explained in the materials and methods (section 3.4) for 20 amino acids. Different atom classification models were used to develop the radial pair distribution function.

(i) Boltzmann Energy Distribution of Atom Types

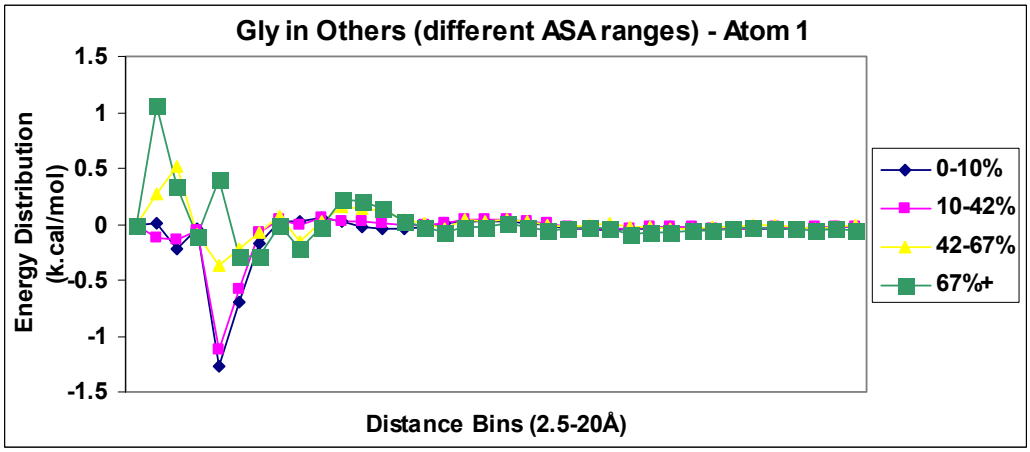
The energy distribution was calculated using eqn. 11. Each atom type's energy distribution was calculated in the distance range from 2.5Å to 20Å. Thus, 800 curves (20 amino acids × 40 atom types) of energy distribution were calculated for MF40 atom model. Moreover, the Basic5, AAC_α, LN24 and SA28 atom models also generate 200, 400, 480 and 560 curves respectively. For the purpose of comparison and visual distinction, the environments of amino acids with contradictory nature are compared (Fig. 3, 4 and 5).



(a)



(b)



(c)

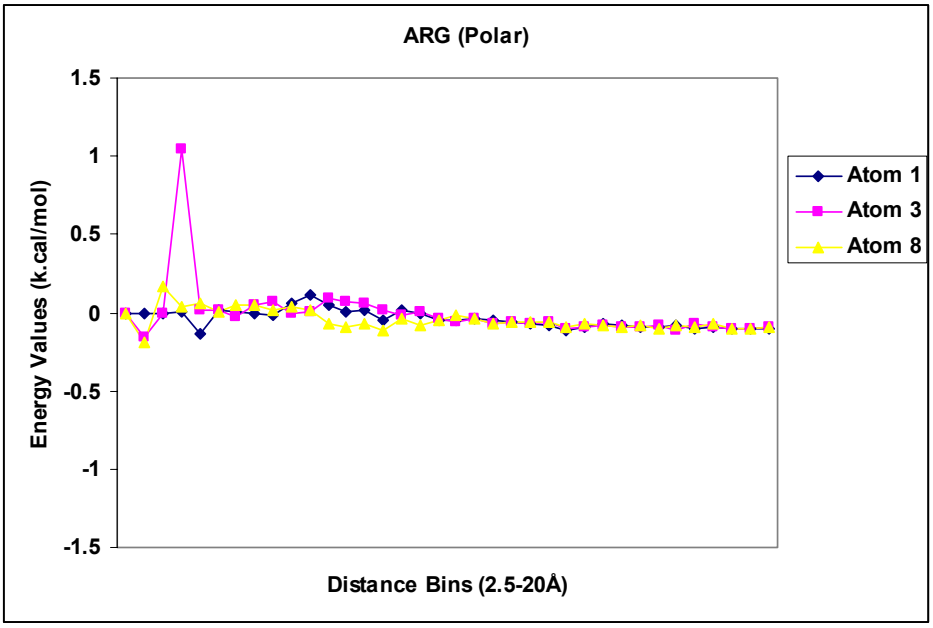
Fig. 3: Distribution of 'atom 1' (atom 1 is C_α atom of amino acids except Gly's C_α atom) in Gly's environment from MF40 atom model. (a) Gly in Helices. (b) Gly in sheets (c) Gly in others (turns, coils, etc.). Relative ASA ranges (legends) for different structural regions are classified from 1% to 100% within the secondary structure elements.

Following cases are discussed:

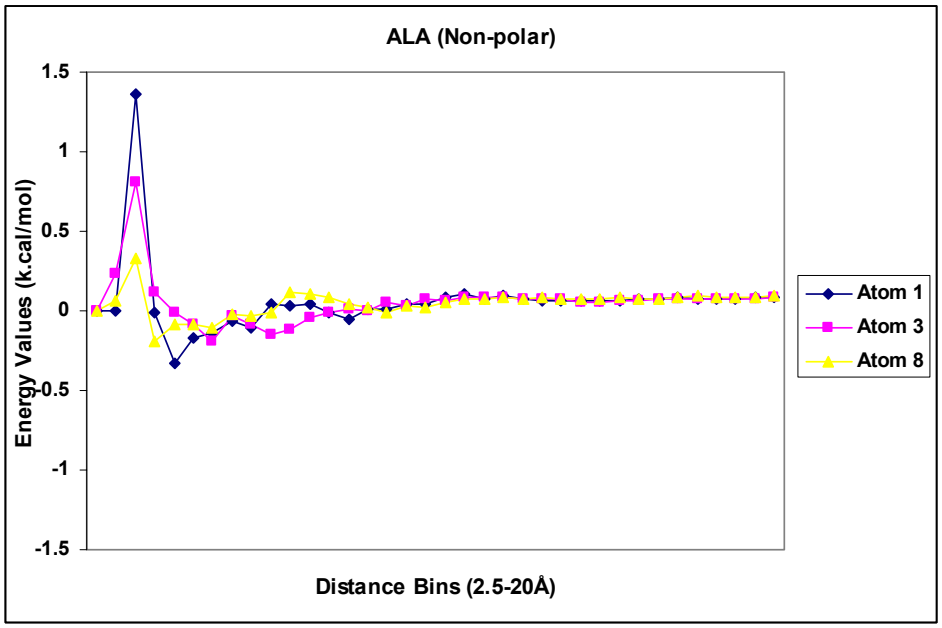
(1) To compare the distribution of a specific atom (atom 1 from MF40 atom model: C_α atom of amino acids except Gly's C_α atom) around a specific amino acid (Gly) in different structural regions, energy distribution plot of Gly's environment in helices, sheets and others (turns, coils, etc) are compared (Fig. 3). The structural regions were further distinguished using relative ASA (for compactness).

(2) Environments of amino acids that are different in nature are compared. Initially, polar (Arg) and non-polar (Ala) amino acid environments are compared (Fig. 4). Later, aliphatic (Val) and aromatic (Phe) amino acid environments were compared (Fig. 5).

The results of energy distribution clearly show that the amino acid environments can be distinguished using an atom level coarse grained model. The distributions of atom 1 in different secondary structure elements (helices, sheets and others) are compared. If a specific atom is not observed in a particular distance bin (Fig. 3) or a distance bin contains no atom counts for all amino acids, energy values become either ∞ (-log 0) or undefined (0/0). When a new amino acid is substituted in a protein mutant, the mutant amino acid may come across atoms in these distance bins that are not naturally observed in its optimal distribution calculated from structural training datasets. In that case, penalty values for energy distribution in mutants for those specific distance bins are assigned. These penalty values denote the presence of a destabilising environment in protein mutant.

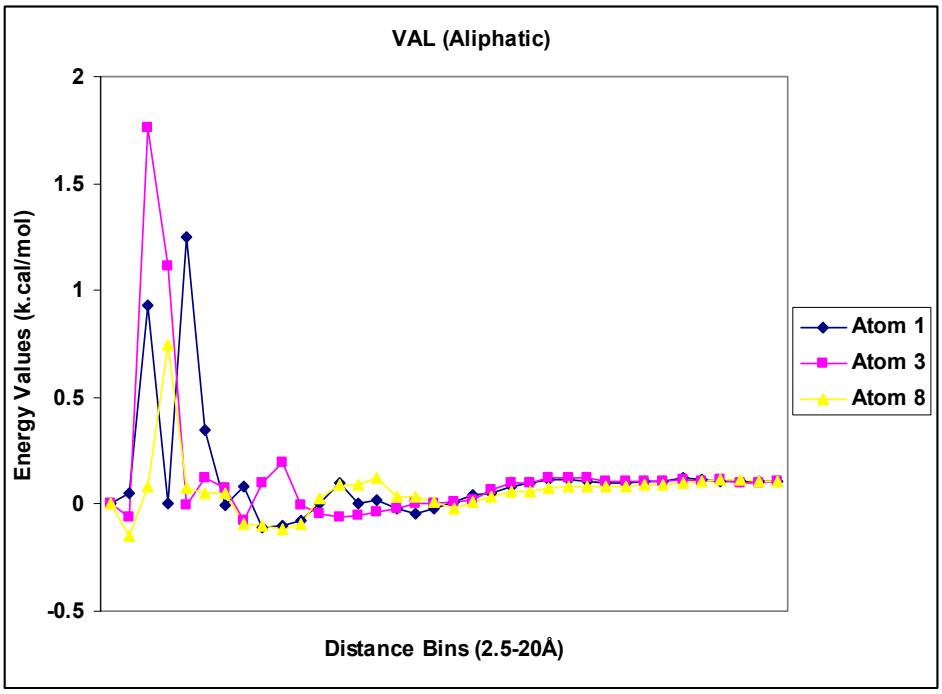


(a)

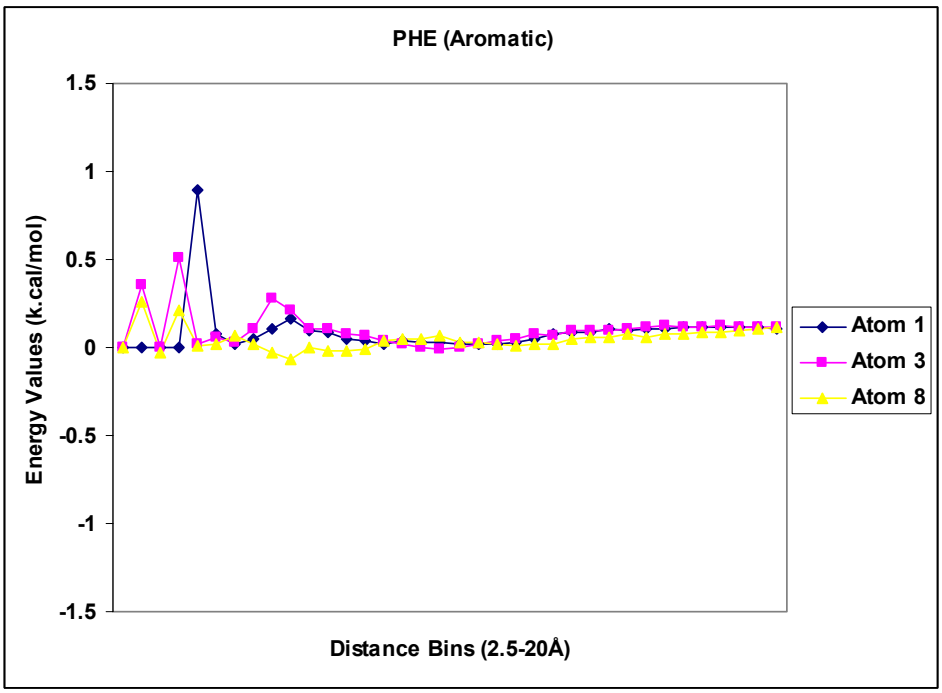


(b)

Fig. 4: Comparison between polar (Arg) and non-polar (Ala) amino acid environments. Boltzmann's energy distributions of atom 1 (C_α atom of amino acids except Gly's C_α atom), atom 3 (N- terminal nitrogen atom of amino acids except Pro) and atom 8 (some of the Cβ and its neighbouring atoms) of the MF40 atom model are plotted. These Arg and Ala exist in helices.



(a)



(b)

Fig. 5: Comparison between aliphatic (Val) and aromatic (Phe) amino acid environments that exist. Boltzmann's energy distributions of atom 1 (C_α atom of amino acids except Gly's C_α atom), atom 3 (N- terminal nitrogen atom of amino acids except Pro) and atom 8 (some of the C_β and its neighbouring atoms) of the MF40 atom model are plotted. These Val and Phe exist in helices.

Energy distribution also differs between amino acids that are different in nature. Distribution of atom 1, atom 2 and atom 8 differs (Fig. 4) at short distance ranges for amino acids Arg and Ala. Atom 1 has peaks in positive energy in Ala, whereas Arg shows null values in similar distance ranges. Similarly, Val shows high positive values at short distance ranges (3-3.5Å), but Phe shows slightly favourable values in the same ranges. This differs due to the size and volume of these amino acids (Fig. 5; Appendix C). Presence of the aromatic ring increases the packing of Phe better than Val in many cases.

(ii) Optimisation of Pair Potentials

Optimisation of statistical pair potentials is an important step to select the best possible combination of parameters to construct the prediction model. Apart from the protein environment and the use of atom types, other important parameters that are supposed to be optimised are given below:

(1) Distances cutoff for radial distribution (Minimum & Maximum).

(2) Distance bin size.

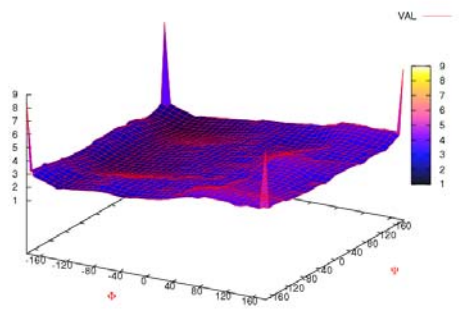
(3) Reference state for Boltzmann's function.

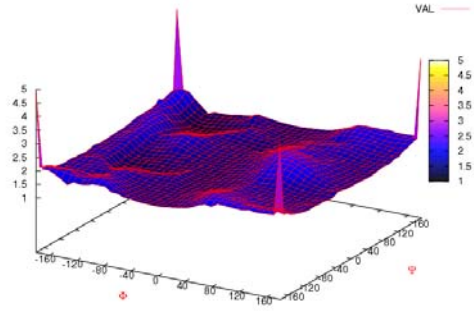
As already stated in methods, the standard way (Sippl 1990) of calculating the reference state (approximation over all the amino acids) was used. On the other hand, optimisation was carried out to set the minimum and maximum distance cutoff for pair potentials. To make the prediction more accurate, bin size of 0.5Å was taken. All types of atoms were present in almost all ranges, except in some cases of short distance ranges. One of the shortest possibilities of any neighbouring atoms (present in one of the atom models) to come close may be between 2.5 to 3Å (e.g., Prolines in CIS conformation). At the same time, some interactions that could act in long distance ranges were also observed in the distance ranges of 19 to 19.5Å. Thus, the minimum and maximum cutoff distances for the radial pair distribution function was set to 2.5Å and 20Å respectively so that many of the atomic interactions to assess the protein structural stability can be captured by this function.

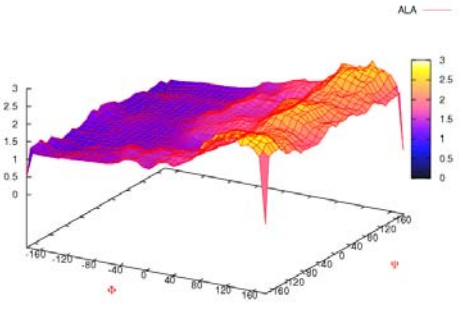
(iii) Boltzmann Energy Distribution of Torsion Angles

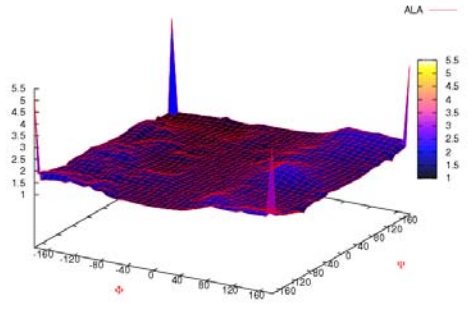
As described in Methods, the distribution of torsion angles ϕ and ψ was derived and Gaussian apodisation function was applied. For the 20 amino acids, 3D plots were drawn from the energy values of 180×180 ϕ and ψ combinations. The torsion angle potential has been mainly developed from two structural training datasets: PISCES (Jan 2004) and top500 (Mar 2004). SCOP-ASTAL dataset was also used for comparison. The energy distributions for both the datasets are given in fig. 6. The energy distribution was also compared with and without the Gaussian apodisation. The effect of altering the maximum value of tapering angle for the Gaussian apodisation was also compared between two angles: 7° and 10°. The difference in smoothing and the edge effect can be observed from Thr (Fig. 6) at a ϕ-ψ combination of 160×160, since the maximum value of 7° shows comparatively smooth effect due to reduction of edge effect (section 3.5.2). This resulted in using 7° as the maximum value for the Boltzmann energy values in the prediction model.

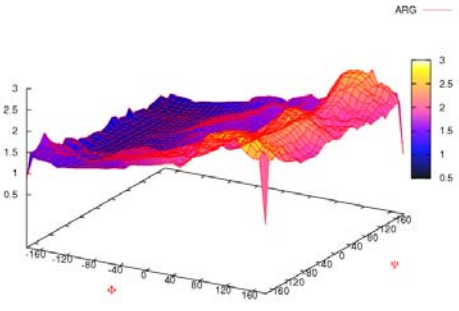
Though the 'top500' dataset includes enough amino acids to cover many of the possible distribution of ϕ and ψ combinations, some of the combinations were still not covered. This is evident from the fact that the energy distribution of the same amino acids (Fig. 6: e.g., Arg, Gly, etc.) derived from the two datasets, PISCES and top500, were noticeably different from each other. Additional counts of ϕ and ψ combinations were observed from PISCES (PI-7) dataset. Thus, PI-7 dataset was selected for the prediction model. However, the correlation coefficient and prediction accuracy between the experimental and predicted energy values must be compared after the implementation of the statistical regression models.

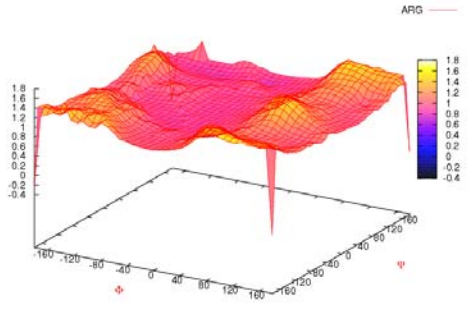

VAL
Φ
Ψ

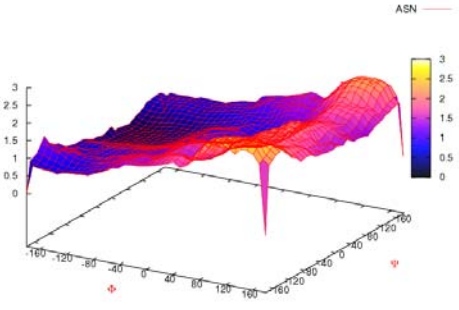

VAL
Φ
Ψ

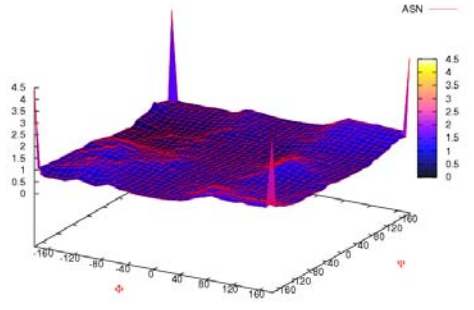

ALA
Φ
Ψ

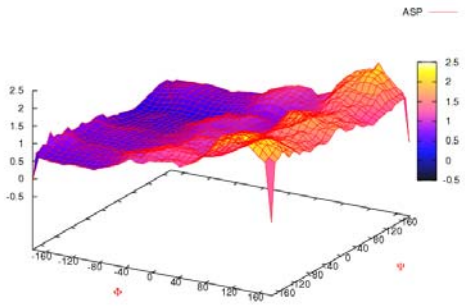

ALA
Φ
Ψ

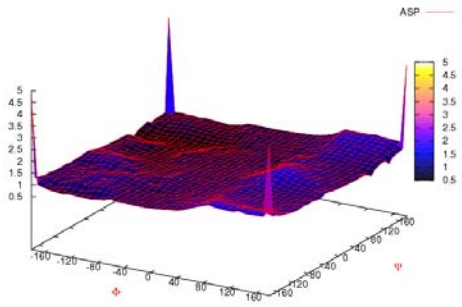

ARG
Φ
Ψ

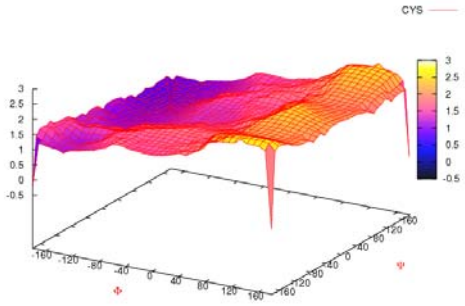

ARG
Φ
Ψ

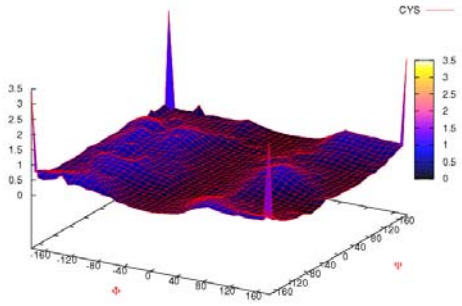

ASN
Φ
Ψ

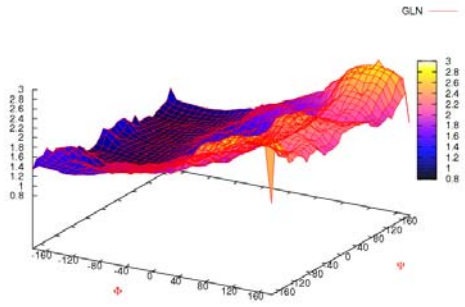

ASN
Φ
Ψ

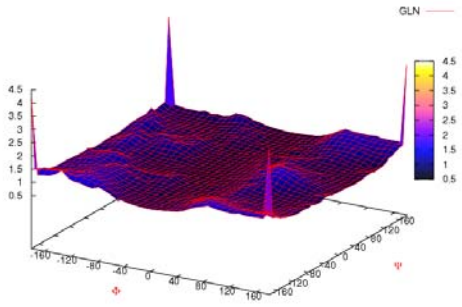

ASP

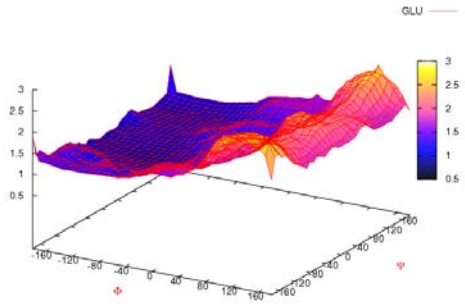

ASP

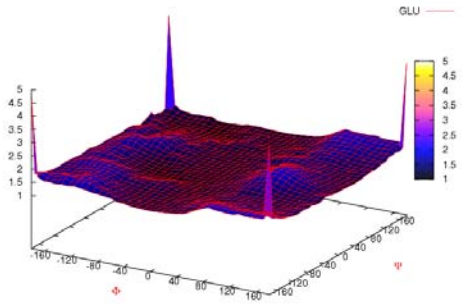

CYS

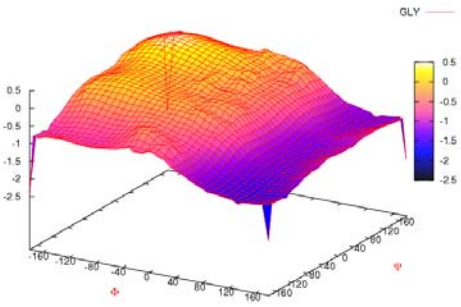

CYS

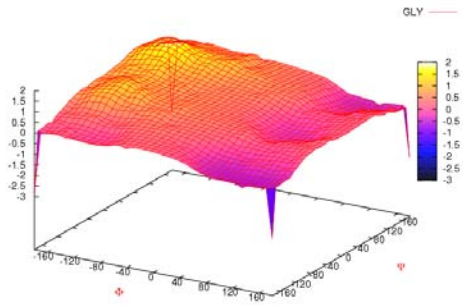

GLN

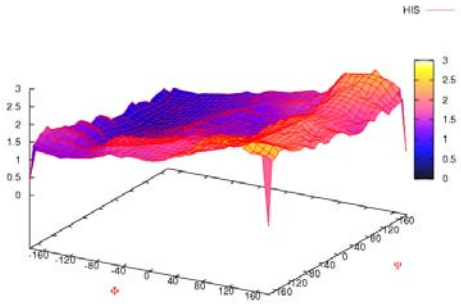

GLN

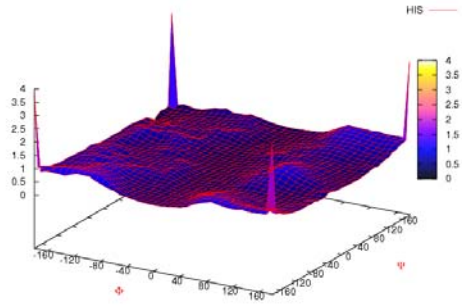

GLU

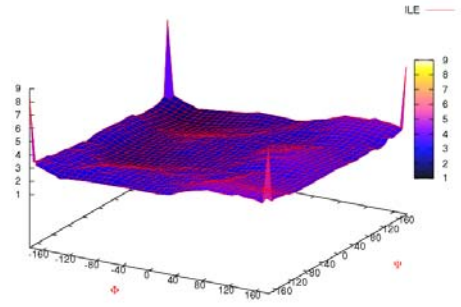

GLU

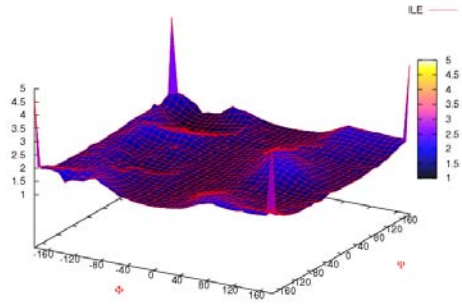

GLY

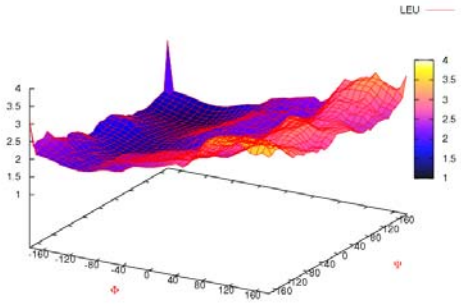

GLY

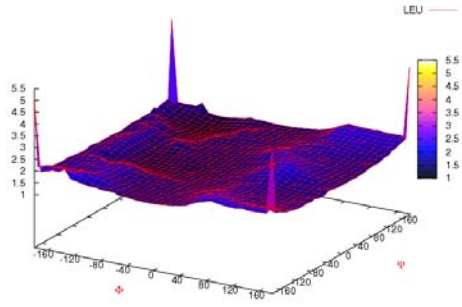

HIS

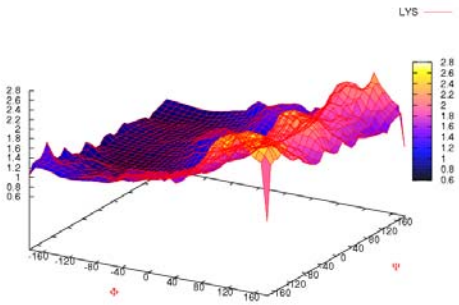

HIS

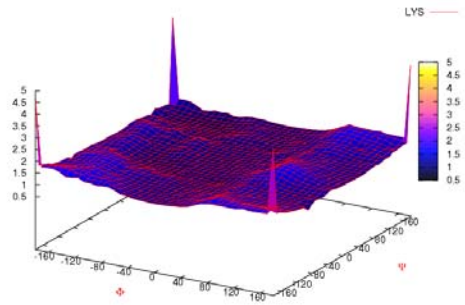

ILE

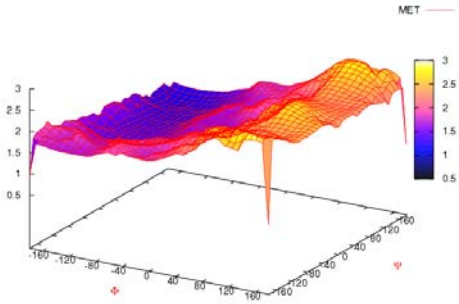

ILE

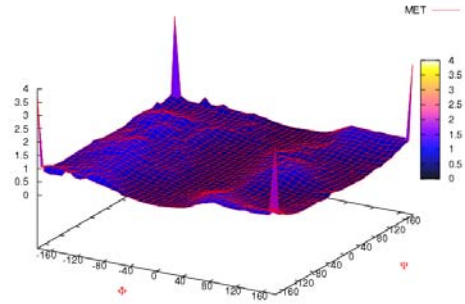

LEU

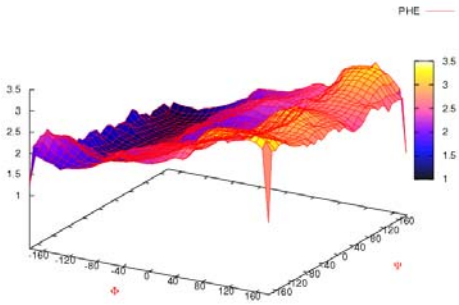

LEU

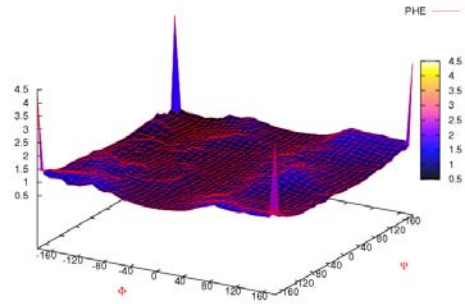

LYS

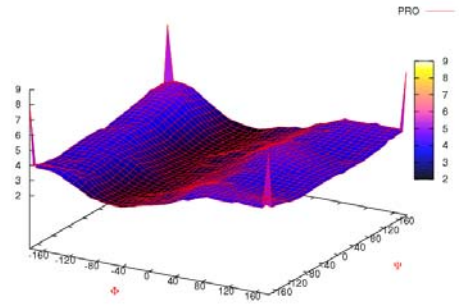

LYS

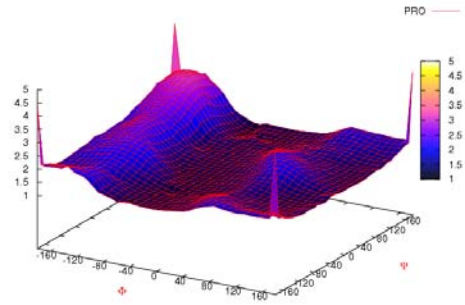

MET


MET


PHE


PHE


PRO


PRO

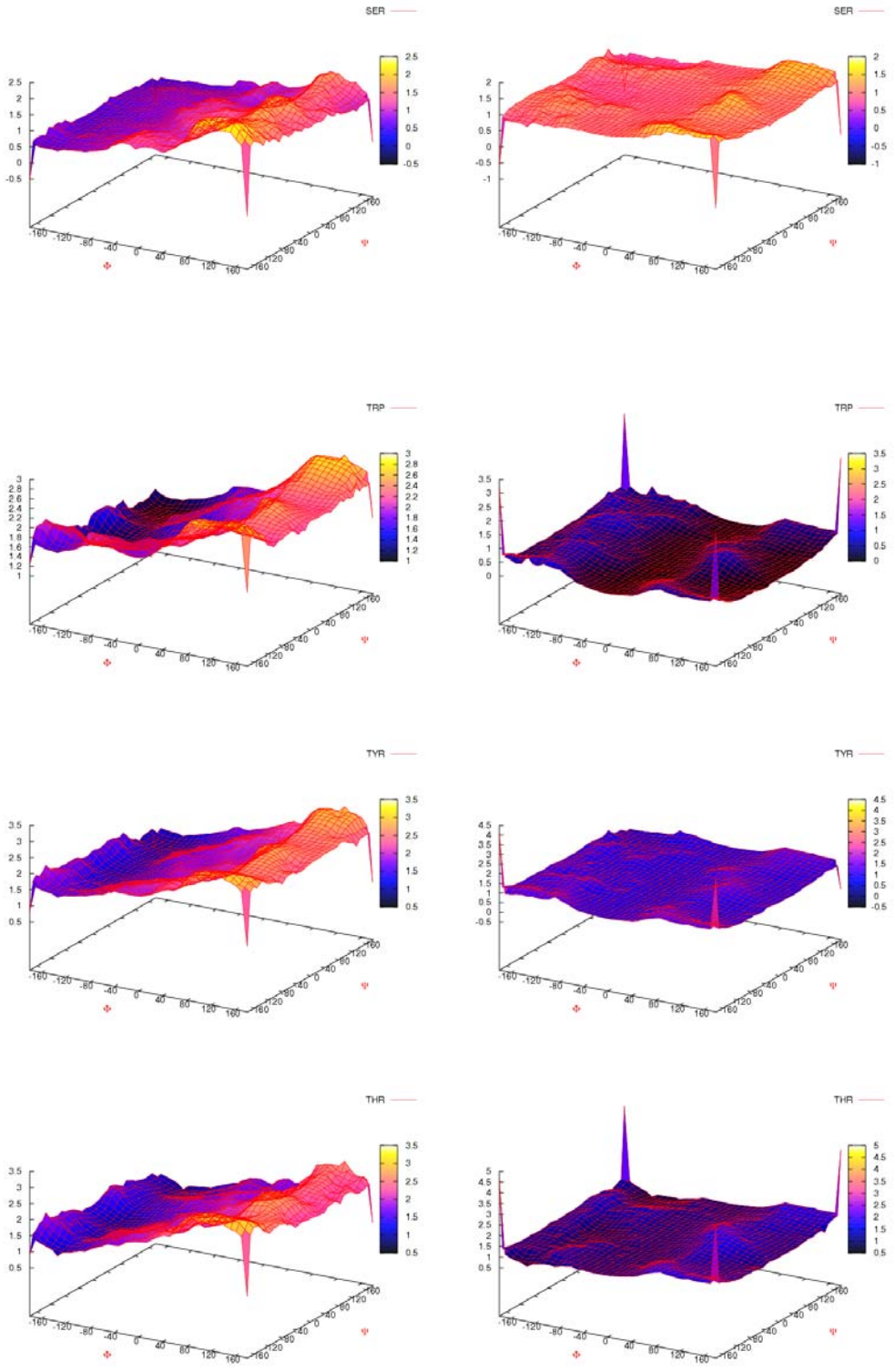


Fig. 6: Boltzmann's energy distribution derived from torsion angles ϕ and ψ for 20 amino acids. Plots from left and right columns are derived from PI-7 and 'top500' datasets respectively and compared. Corners of the distribution graphs are denoted with sharp legs/edges (shown up or down depending on general distribution data value): These are not energy values.

4.2 The Prediction Model

4.2.1 Mutation Datasets

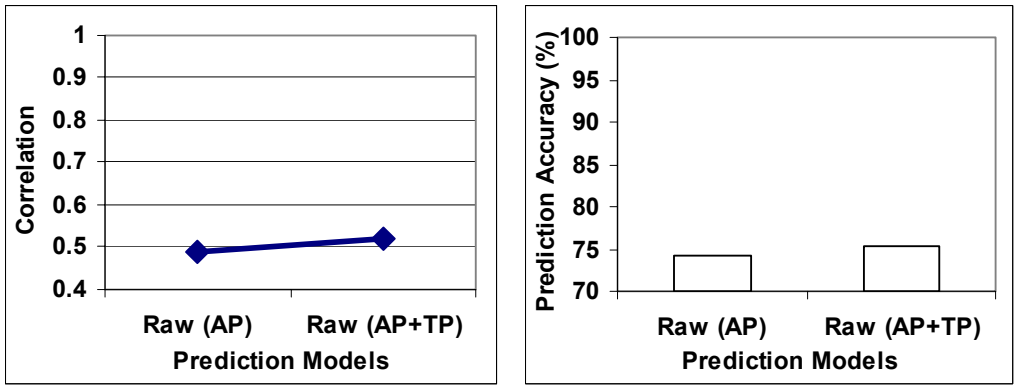
Amino acid single mutations were taken from Protherm database (Gromiha et al. 1999a; Bava et al. 2004) and literature (Alber et al. 1987; Yutani et al. 1987; Shih et al. 1995; Shoichet et al. 1995; Topham et al. 1997; Xu et al. 1998) whose stability relative to the wild-type ($\Delta\Delta G$ or $\Delta\Delta GH_2O$) were determined experimentally. Mutants range between the core and periphery with highly variable solvent accessibility and secondary structure specificity (Tables 8, 9). At the same time, the proteins also vary widely in their sequence identity and functional aspects.

4.2.2 Simple Linear Regression

In the simple linear regression model, the atoms' stabilisation energy values were calculated from the Bolzmann's energy values. Depending on the wild type and mutant amino acid combinations, these stabilisation energy values were calculated and added together to form the final stabilisation energy of a specific mutation. However, simple regression gave a correction coefficient of 0.29 for 1543 mutations with thermal $\Delta\Delta G$ values. For 1603 mutations with $\Delta\Delta GH_2O$ values, a correlation coefficient of 0.30 was observed. Thus, the simple linear regression developed a low correlation coefficient, and failed to formulate a relationship between predicted and experimental stabilisation values.

4.2.3 Multiple Linear Regression

Since the simple regression did not provide enough contribution towards the linear relationship and prediction equation design, a multiple regression model was chosen as a viable solution. Here, the atoms were fit with experimental data using dynamic regression coefficients. These regression coefficients were calculated for all the atoms, by regressing the stabilisation energy values with the experimental $\Delta\Delta G$.



(a) **(b)**

Fig. 7: (a) Correlation between predicted and experimental ΔΔG from thermal denaturation and (b) prediction accuracy for mutations to be correctly predicted as stabilising or destabilising. Raw uses multiple linear regression for 1538 mutations without classifying them into different structural regions. Raw (AP) uses only atom potentials for prediction. Raw (AP+TP) uses both atom and torsion angle potentials for prediction.

The correlation coefficient was observed to be 0.49 with ΔΔG values of 1538 mutations (Fig. 7a). Mutations that were correctly predicted to be stabilising or destabilising was observed to be 74.21% (Fig. 7b). After the inclusion of torsion angle potentials, the correlation coefficient increased to 0.52 (Fig. 7a) with 75.31% of the mutations correctly predicted (Fig. 7b). This phenomenon was already observed by some of the investigators previously for a small set of mutations. However, the prediction efficiency of the model with large amount of mutations decreases dramatically. In the simple linear regression, the regression coefficient was determined only for the final stabilisation energy calculated from the atoms. But, in multiple linear regression, unique regression coefficients were calculated for all the atoms. Then, the stabilisation energy values were added, after multiplying with respective regression coefficients.

Torsion angle potentials can be included with pair potentials in two ways. In the first method, the pair potentials (stabilisation energy values provided by individual atoms) can be regressed separately with experimental ΔΔG to calculate the total stabilisation energy contributed by pair potentials alone. Later, the stabilisation energy values of both torsion angle and pair potentials can be added using the weighting factors and the final ΔΔG can be calculated.

List of Protein Mutants with thermal ΔΔG values					
PDB ID	**All**	**Helices**	**Sheets**	**Turns**	**Others**
1lz10	137	29	26	44	38
1bpi0	45	12	13	0	20
1g6nA	2	2	0	0	0
1ycc0	37	24	0	0	13
1bniA	26	4	16	0	6
1mbg0	3	3	0	0	0
1hfzA	23	14	3	0	6
2aky0	4	2	1	0	1
1pga0	5	0	2	3	0
1rn1B	26	3	14	6	3
2lzm0	403	316	16	11	60
1onc0	9	0	0	9	0
1rtb0	15	0	7	0	8
2ci2I	85	18	17	6	44
4lyz0	42	13	3	10	16
1ropA	21	2	0	19	0
1bta0	1	0	0	0	1
1stn0	17	5	3	0	9
1c9oA	14	1	9	1	3
1csp0	7	0	5	0	2
1el1A	4	2	0	2	0
2rn20	110	58	23	7	22
1ankA	4	0	0	0	4
1bvc0	35	33	0	0	2
5croO	11	2	0	0	9
1arrA	1	0	1	0	0
3ssi0	49	14	11	0	24
1poh0	13	13	0	0	0
1sup0	6	0	2	0	4
1clwA	5	0	2	0	3
2trxA	2	0	2	0	0
1cyo0	3	3	0	0	0
1sarA	3	0	0	3	0
1tpkA	6	0	6	0	0
1am7A	2	0	0	1	1
1em7A	12	0	12	0	0
1chkA	9	6	3	0	0
2hpr0	3	3	0	0	0
1bgsA	59	17	4	2	36
1lyd0	45	29	13	2	1
1rnbA	79	14	18	10	37
4mbn0	8	6	1	0	1
3lzm0	57	44	2	7	4
1l630	89	69	0	4	16
1a230	1	1	0	0	0

Table 8: List of proteins and number of mutations with thermal ΔΔG values: 'All' indicates all the mutations of a specific protein.

List of Protein Mutants with $\Delta\Delta GH_2O$ values from chemical denaturation					
PDB ID	All	Helices	Sheets	Turns	Others
1bniA	139	62	25	5	47
1bvc0	31	24	0	2	5
1stn0	511	137	157	97	120
1arrA	62	35	9	0	18
2ci2I	69	26	17	11	15
1hfyA	8	8	0	0	0
1axb0	1	1	0	0	0
1yea0	2	1	0	0	1
1ubq0	4	4	0	0	0
1fkj0	36	9	20	0	7
1ycc0	33	30	0	0	3
1amq0	6	2	1	2	1
2lzm0	25	15	3	3	4
1a230	2	1	0	0	1
1cah0	2	1	0	0	1
3pgk0	9	3	3	0	3
1sakA	27	19	0	0	8
1igvA	14	8	4	1	1
1bta0	6	3	2	0	1
1rhgA	7	7	0	0	0
1rx40	47	7	16	1	23
1bp20	9	9	0	0	0
2mm10	3	3	0	0	0
3hhrA	10	6	0	0	4
4lyz0	36	15	5	2	14
1cyo0	2	2	0	0	0
2trxA	6	1	5	0	0
1c2rA	5	4	0	0	1
1akk0	3	3	0	0	0
2wsyA	65	2	56	0	7
1i5tA	8	3	0	1	4
451c0	5	4	0	1	0
2hmb0	9	1	7	1	0
1qlpA	13	2	8	1	2
1c9oA	10	1	7	0	2
2hpr0	3	3	0	0	0
1dktA	19	1	9	4	5
2afgA	7	1	0	4	2
1iob0	15	3	9	0	3
1lz10	10	2	0	6	2
2rn20	18	4	8	1	5
1rn1A	18	1	11	6	0
1htiA	4	3	0	0	1
3mbp0	2	1	1	0	0
1fxaA	11	4	0	6	1
1oiaA	6	1	4	1	0
1a430	5	2	0	1	2
1poh0	9	8	0	0	1
1cyc0	4	4	0	0	0
1znjA	4	3	0	0	1
1sceA	14	1	0	3	10

1vqb0	12	0	12	0	0
1shfA	30	0	13	15	2
1znjB	3	0	3	0	0
2ifb0	16	0	6	3	7
1flv0	6	0	4	0	2
1hngA	1	0	1	0	0
2imm0	11	0	6	1	4
1ttg0	37	0	27	1	9
1ten0	34	0	24	0	10
1dil0	2	0	0	0	2
1bpi0	22	0	12	0	10
5azuA	2	0	0	0	2
1lve0	16	0	7	5	4
1aarA	1	0	1	0	0
1pga0	5	0	5	0	0
1fepA	6	0	5	1	0
1tupA	5	0	1	0	4
1csp0	6	0	4	0	2
1mjc0	6	0	6	0	0
1div0	1	0	1	0	0
1idsA	1	0	0	1	0
3blsA	1	0	0	0	1
1b0o0	1	0	0	0	1
1onc0	1	0	0	1	0
1tit0	1	0	0	0	1
1av1A	1	0	0	1	0
1frd0	1	0	0	1	0

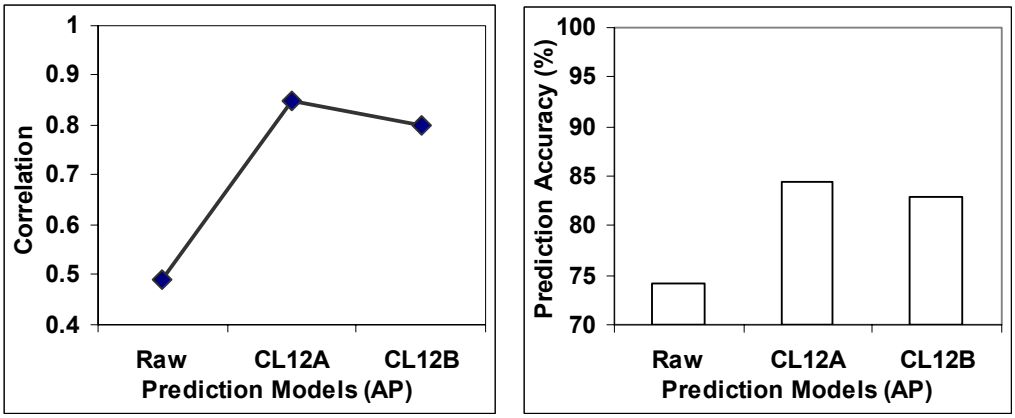
Table 9: List of proteins and number of mutations with $\Delta\Delta GH_2O$ values: 'All' indicates all the mutations of a specific protein.

In the second method, the stabilisation energy values contributed by individual atoms and torsion angle potential can be regressed together with experimental $\Delta\Delta G$ in a single step. It has benefits for the dynamic stepwise selection of independent regression variables (atoms and torsion angles).

4.2.4 Classifying the Protein Environment

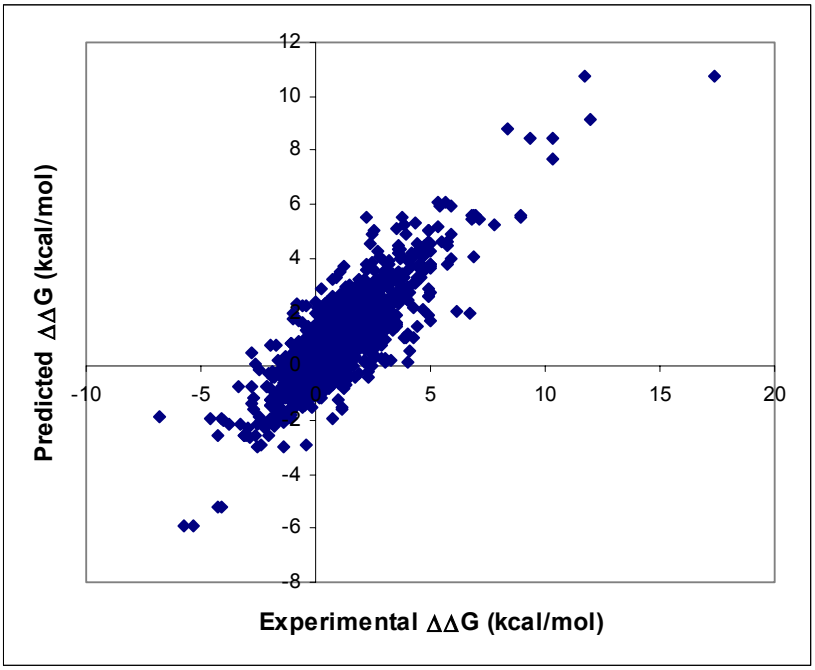
Correlation and prediction efficiency provided by multiple regression was not enough for a prediction model because one generalised model that covers all the structural regions in proteins cannot completely distinguish and predict the changes in protein mutants. Classification of structural regions using the ASA and secondary structure specificity was then implemented to classify mean force potentials and mutation dataset into smaller subsets. Optimisation of total number of structural regions was necessary (section 4.3.6) to obtain reliable results for the correlation and prediction accuracy (%).

The correlation and prediction accuracy of mutant stability changes compared to the experimental ΔΔG for all the mutations together before and after using the two methods of classification are shown in fig. 8. Prediction efficiency increases dramatically with the classification of potentials and mutations. A single generic model to fit all mutations without classification is definitely less accurate. It has been reported that the factors influencing the stability of protein mutants depend on the location of mutants based on secondary structure and solvent accessibility (Gromiha et al. 1999b). Furthermore, an optimised usage of relative ASA (section 4.3.6) and secondary structure specificity together to distinguish the prediction model showed a correlation of 0.85 with experimental ΔΔG for all 1538 mutations (Fig. 8a, 10a) with 84.8% of mutations correctly predicted (Fig. 8b, 10b). These results used the CL12A dataset which uses 3 secondary structures and 4 ranges of relative ASA to distinguish the mutants. CL12B (Fig. 8) was not considered in the final model, since the experimental data included a smaller number of mutations in turns, which are too small to develop a reliable prediction model.

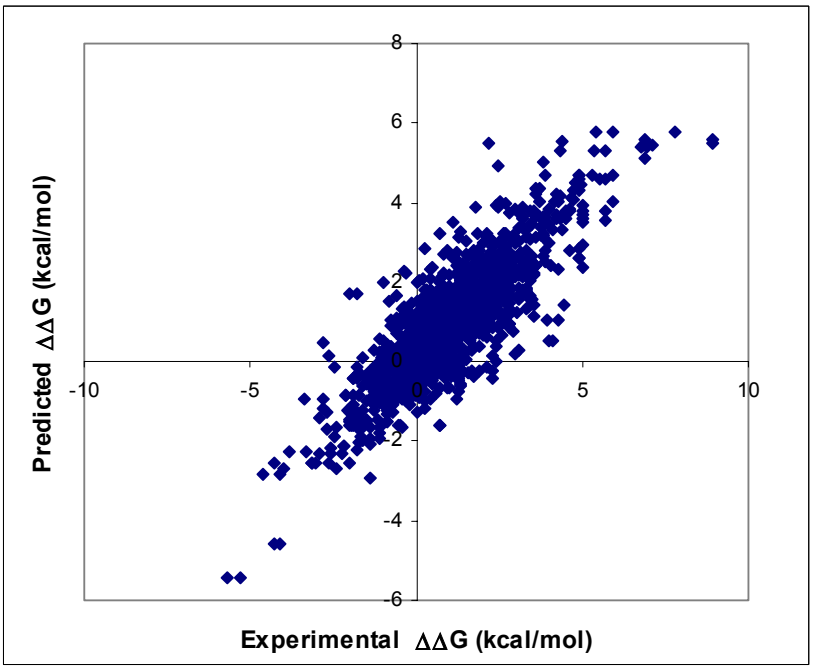


(a) **(b)**

Fig. 8: Prediction of protein mutant stability in a set of 1538 mutations using atom potentials (AP) with different classifications: (a) correlation and (b) prediction accuracy. Raw: all mutations are taken without classification, CL12A: classification of mutants into three secondary structures and four ranges of solvent accessibility [helices (0-2, 2-30, 30-60, 60+), sheets (0-5,5-35,35+) and others (0-10,10-42,42-67,67+)]. CL12B: Classification into four secondary structures and three ranges of relative ASA with each secondary structure [helices, sheets, turns and others with ASA ranges 0-2, 2-50, 50+ for each secondary structure].

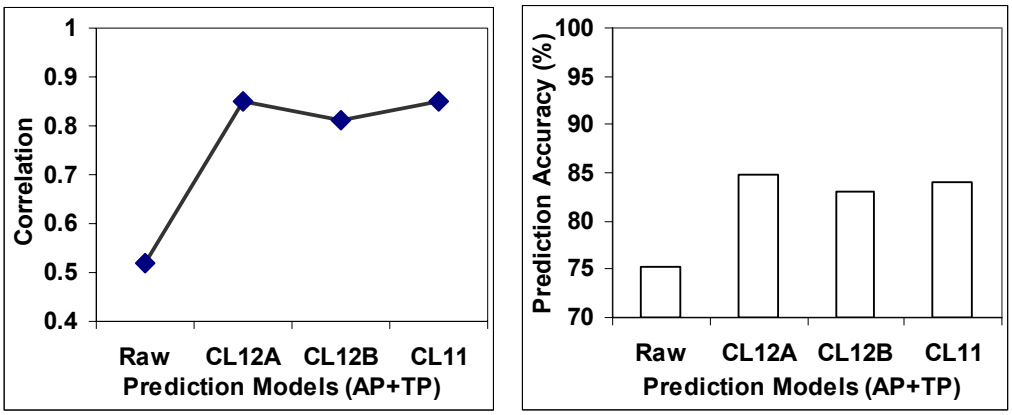


(a)



(b)

Fig. 9: Scatterplots explaining the experimental and predicted ΔΔG values for 1538 (a) and 1518 (b) mutations. These mutation datasets were used with and without 20 outliers respectively. Outliers were removed to improve the prediction efficiency of the multiple regression model.



(a) **(b)**

Fig. 10: Prediction improvement after the inclusion of torsion potentials (TP) with atom potentials (AP): (a) Correlation and (b) prediction accuracy for 1538 mutations. CL11 [helices (0-2, 2-30, 30-60, 60+), sheets (0-5,5-35,35+) and others (0-10,10-42,42-67,67+)] indicates a new classification system into 11 structural regions in order to reduce over fitting of variables that may occur in the previous classification system.

Scatter plots explaining the experimental and predicted ΔΔG are shown with (1538 mutations) and without (1518 mutations) the removal of outliers (Fig. 9a, 9b). After the removal of 20 outliers (0.1%), correlation increased to 0.87 with 85.3% of mutations correctly predicted (Fig. 11).

4.2.5 Multicollinearity Diagnostics

Multicollinearity was detected, since some atoms in all the secondary structure and ASA range showed a VIF (Variance Inflation Factor) more than the selected cutoff. Tables 10 and 11 show correlated atoms with their VIF values. VIF values more than 10 indicate symptoms of multicolinearity. To resolve this issue, their distribution was unified and used together with other atoms.

Prediction models that implement VIF cutoff of 20, 30 and 50 showed to have values of correlation coefficient of 0.78, 0.81, 0.82 respectively (Fig. 11a), though variable amount of atoms (Table 14) were used and over-fitting problem was minimised. Prediction accuracy of 81.11% for VIF cutoff value 20 and 81.97% for VIF cutoff values 30 and 50 were observed (Fig. 11b). Since the results show highly similar correlation coefficient and prediction accuracy, presence of multicollinearity in the model was clearly visible. However, more number of atoms should be reduced, as the validation of mutation data might be

carried out with reduced amount of mutations if they are broken into training and test sets respectively.

4.2.6 Stepwise Linear Regression

Alternatively, stepwise regression was also used where forward selection was employed to select the atoms dynamically. Prediction equations were derived using this statistical model, and the results were compared. Tables 12 and 13 indicate selected atoms and their regression coefficients calculated for 1518 mutations. The correlation coefficient remained at 0.84 (Fig. 11c) for all the 1538 mutations together with reduced number of atoms (Table 14). Correlation increased to 0.86 after the removal of 20 outliers. Interestingly, the prediction accuracy slightly increased to 84.79% (Fig. 11d). This model with reduced atoms showed only a small difference in prediction efficiency against the initial model with all 40 atoms. Thus, the stepwise model has been used for all the validation tests and employed as final statistical prediction model.

Since the multiple (MLR1) and stepwise regression model (SRM1) were proved better for $\Delta\Delta G$ based on thermal experiments, the same were applied to $\Delta\Delta GH_2O$ values (MLR2 and SRM2 in Table 20). Correlation coefficients were observed to be 0.81 and 0.79 for the models based on multiple (MLR2) and stepwise (SRM2) regression respectively. Prediction accuracies were observed to be 86.02% and 85.07% for these models respectively. Later, validation tests for the prediction models were carried out and results were analysed in one of the next sections (section 4.3).

Outliers were observed in the scatterplot and removed from the both the datasets. Instead, standard deviation (σ) can also be used to remove outliers. For example, mutants that have values more than $3\times\sigma$ or $4\times\sigma$ can be removed. The selection of outliers was almost same between the former and latter methods. For thermal $\Delta\Delta G$ values, some mutations of protein kinase inhibitors (PDBID: 3SSI) and tail-spike Protein from Phage P22 (1CLW) were removed. These mutations had $\Delta\Delta G$ values more than 10 kcal/mol. Usually, the values of $\Delta\Delta G$ range from 5 to 15 kcal/mol through the whole folding-unfolding

transition (Shirley 1995). Extreme positive values may already favour the denatured state in the solution, and possibility of having equilibrium is minimised. Besides, statistical regression also often employs extreme values of regression coefficients to accommodate the extreme values of ΔΔG. This leads to an unreliable prediction model. For $\Delta\Delta GH_2O$, only extreme positive values were removed.

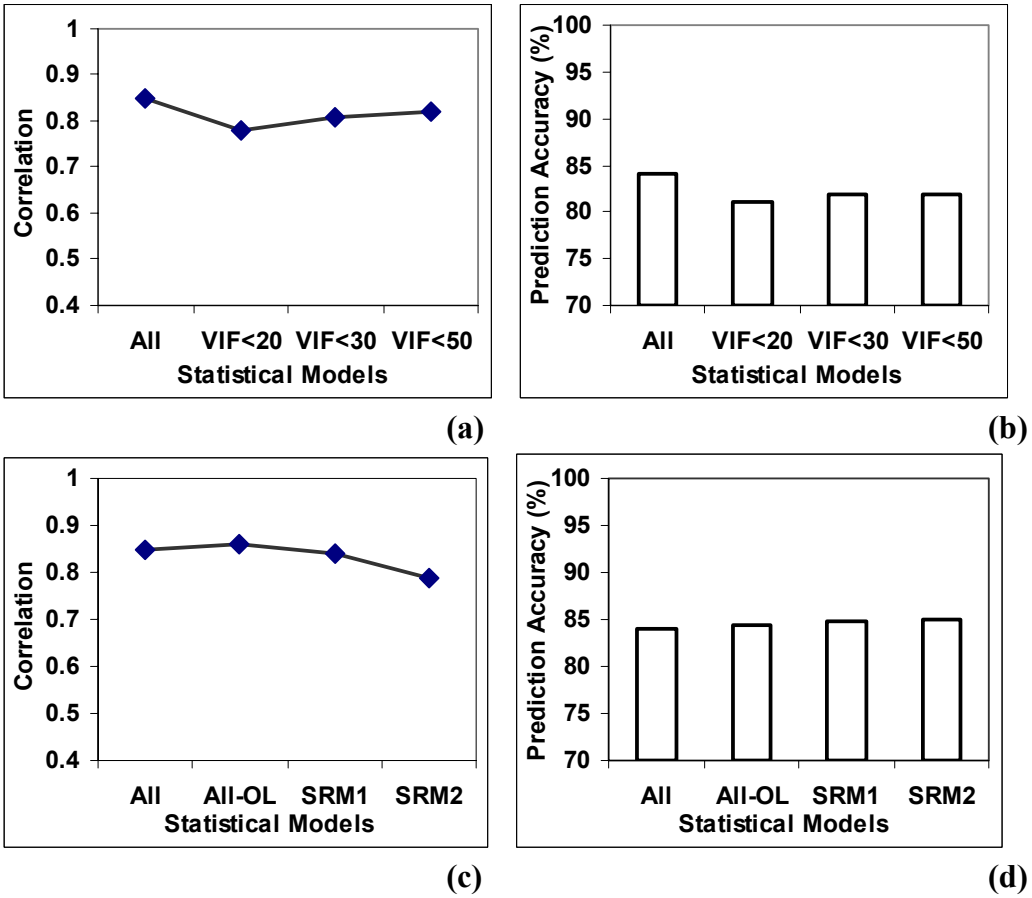


Fig. 11: Optimisation of atom types using various statistical regression models: (a) correlation and (b) prediction efficiency based on the analysis and reduction of atom types after multicolinearity diagnostics. 'All' indicates the usage of all atoms for the statistical model. VIF<20, VIF<30 and VIF<50 indicate the statistical models that use atoms with VIF values less than 20, 30 and 50 respectively. (c) Correlation and (d) prediction accuracy based on the reduction of atoms with stepwise regression selection methods. 'All' and 'All-OL' indicate the datasets of mutations before and after the removal of outliers using normal multiple regression. 'SRM1' indicates the stepwise regression selection model using ΔΔG after the removal of outliers for the prediction of protein mutant stability. 'SRM2' indicates stepwise regression model using $\Delta\Delta GH_2O$ for 1603 mutations.

Atom / Tor	ASA+SS Classified Statistical Potentials for Various Structural Regions										
	Helices				Sheets			*Others*			
	0-2	2-30	30-60	60+	0-5	5-35	35+	0-10	10-42	42-67	67+
AT1	432.5	208.6	172.7	259.1	121.1	148.0	175.7	324.0	144.2	92.9	80.5
AT2	15.8	10.1	5.4	7.0	21.3	8.4	5.2	14.6	3.5	4.3	16.0
AT3	370.0	262.7	260.9	280.3	224.8	247.5	227.3	327.7	81.5	121.5	132.1
AT4	674.7	241.7	215.9	373.0	273.8	136.0	315.7	603.1	114.8	101.4	315.8
AT5	448.0	126.3	61.5	117.8	214.1	184.9	298.2	303.5	123.5	108.3	187.0
AT6	29.4	29.0	49.9	88.7	10.4	40.9	78.3	45.3	26.3	40.6	68.9
AT7	9.0	18.2	24.3	42.4	5.4	29.9	48.0	13.1	35.5	28.8	59.5
AT8	105.1	55.6	45.0	84.3	42.9	35.7	94.9	114.7	36.1	28.9	74.7
AT9	5.9	5.1	5.2	9.0	21.9	5.8	5.1	13.1	5.0	2.7	7.4
AT10	5.6	5.7	2.5	7.4	8.4	3.2	20.6	6.0	3.4	3.8	4.0
AT11	26.1	55.1	23.6	27.0	49.9	68.0	29.3	34.0	34.7	24.6	67.1
AT12	71.9	57.8	28.4	55.1	70.2	88.7	40.8	78.0	41.2	30.3	60.1
AT13	12.9	10.1	3.1	4.7	14.2	7.0	16.2	14.0	4.5	2.9	12.6
AT14	12.7	15.5	3.6	17.7	43.7	9.0	3.6	13.5	4.1	3.2	6.6
AT15	7.1	12.1	6.2	11.0	17.8	9.5	7.0	6.7	3.9	5.0	5.8
AT16	32.5	16.6	13.0	12.3	30.8	12.4	13.5	25.3	6.0	5.8	21.7
AT17	20.1	13.8	6.1	11.2	6.5	9.0	10.2	17.4	5.0	6.4	11.8
AT18	39.3	16.8	20.5	17.4	84.6	33.2	17.1	32.8	8.2	13.6	13.1
AT19	3.8	4.7	3.4	10.3	6.1	2.7	6.0	11.0	4.2	10.3	9.0
AT20	8.6	6.5	7.4	29.7	13.9	15.8	17.7	6.1	5.7	5.5	8.5
AT21	13.8	22.0	17.2	27.6	27.6	14.9	15.8	15.0	15.1	11.6	17.2
AT22	17.6	22.8	32.9	44.0	30.4	16.0	15.5	26.9	14.3	15.6	31.2
AT23	9.0	5.9	7.5	5.2	15.9	5.9	9.6	11.1	4.3	2.9	11.0
AT24	25.4	14.5	5.8	4.6	38.1	13.7	15.6	30.7	8.7	4.0	12.7
AT25	13.4	4.0	7.5	4.9	17.0	5.6	9.3	17.8	3.9	3.4	10.0
AT26	9.9	5.8	9.5	6.6	14.4	7.8	12.6	43.7	5.2	4.2	5.5
AT27	18.4	9.9	13.3	35.2	16.4	11.1	23.8	30.3	6.0	16.2	59.2
AT28	20.1	13.0	14.2	33.7	27.3	13.3	25.2	31.9	5.2	21.1	37.6
AT29	10.4	6.6	6.1	12.0	10.7	4.1	18.2	17.1	2.7	9.3	5.0
AT30	7.3	5.3	5.4	4.3	39.2	6.5	9.3	16.2	4.5	3.1	9.6
AT31	12.9	15.9	7.7	16.5	44.9	11.2	10.5	38.5	7.1	3.5	9.7
AT32	6.2	2.8	2.3	7.5	11.9	2.3	13.9	7.1	3.4	2.6	8.3
AT33	22.2	26.7	22.4	16.6	66.3	37.7	16.7	34.9	13.9	10.3	7.3
AT34	51.8	23.6	18.0	23.8	123.8	12.4	21.1	42.2	12.8	11.1	8.9
AT35	7.6	8.1	8.2	34.0	16.6	14.2	22.8	7.6	6.3	6.2	6.9
AT36	5.9	18.5	27.8	22.4	17.6	14.1	15.6	17.6	17.2	8.0	26.1
AT37	10.8	7.8	21.0	6.1	7.6	8.9	8.5	18.6	13.6	7.6	8.8
AT38	3.6	5.6	4.1	3.1	20.9	6.0	11.0	18.5	4.5	2.8	7.2
AT39	13.1	11.4	3.8	17.7	32.1	15.0	10.4	16.7	5.1	7.3	10.7
AT40	14.5	17.7	10.0	13.2	39.9	12.4	12.8	20.5	7.7	3.5	20.1
Tor.	2.7	3.3	3.6	7.8	3.0	2.8	4.5	3.9	2.3	3.5	2.1

Table 10: VIF (Variance Inflation Factor) values of 40 atoms and torsion angle derived using multicolinearity diagnostics for the experimental $\Delta\Delta G$ values from thermal denaturation experiments.

Atom / Tor	ASA+SS Classified Statistical Potentials for Various Structural Regions										
	Helices				Sheets			Others			
	0-2	2-30	30-55	55+	0-3	3-25	25+	0-15	15-42	42-67	67+
AT1	402.0	185.7	107.4	90.7	377.4	199.7	200.1	180.3	50.5	18.9	96.3
AT2	26.6	6.6	5.6	4.5	7.2	6.2	4.4	10.8	3.5	2.8	4.4
AT3	769.2	144.9	195.2	124.2	548.6	220.4	156.5	275.6	45.0	32.8	54.9
AT4	571.6	171.8	193.9	128.3	718.1	400.0	111.3	304.6	59.1	57.1	117.9
AT5	356.6	99.6	78.6	60.4	601.2	232.4	178.2	181.5	45.1	28.1	60.3
AT6	67.0	26.2	40.1	63.2	48.0	25.9	53.1	28.1	19.0	35.9	96.9
AT7	17.7	11.6	27.3	26.9	21.8	13.1	34.6	13.5	14.2	21.1	60.0
AT8	73.5	38.4	35.8	54.2	105.8	43.3	55.6	63.7	23.2	32.5	39.5
AT9	8.1	5.5	5.8	4.5	20.1	4.2	2.8	6.2	2.6	2.8	2.5
AT10	10.5	3.1	2.9	2.8	45.6	4.9	4.0	4.3	2.3	2.8	2.1
AT11	100.6	33.4	17.0	16.8	31.1	25.9	24.9	47.4	12.7	17.9	13.9
AT12	67.3	35.3	22.7	18.3	44.7	33.2	34.1	44.0	17.1	20.4	20.5
AT13	42.8	3.5	4.4	2.6	6.0	5.3	3.4	17.1	2.5	2.9	2.7
AT14	8.0	7.7	3.6	3.1	7.2	5.5	3.9	8.4	2.7	8.1	4.6
AT15	8.0	5.9	5.9	6.4	8.4	5.0	4.8	8.9	1.8	3.0	5.3
AT16	40.4	8.1	15.3	5.7	11.6	11.9	14.4	30.9	4.8	3.3	4.9
AT17	14.6	6.1	6.2	2.8	3.7	4.4	10.0	14.3	3.8	3.4	4.4
AT18	29.2	29.4	8.1	8.3	23.9	14.5	12.9	10.8	6.0	5.2	3.9
AT19	9.4	13.7	2.3	4.1	2.9	1.5	2.7	4.7	1.7	3.3	4.8
AT20	12.6	5.7	9.6	10.2	7.1	7.5	7.6	4.5	6.7	18.2	8.1
AT21	28.9	10.6	10.7	11.3	43.5	11.3	10.9	9.5	5.3	8.9	7.9
AT22	26.8	14.9	9.7	22.2	26.9	14.5	12.5	14.5	6.6	13.9	6.6
AT23	33.7	5.1	8.5	5.6	6.4	5.1	4.2	10.9	3.6	2.2	6.2
AT24	46.3	10.7	3.8	4.3	12.1	10.3	6.6	25.7	5.4	3.0	5.5
AT25	16.5	4.6	7.5	6.0	5.6	5.3	5.0	12.8	3.2	4.4	4.2
AT26	21.4	6.8	8.1	5.5	9.1	6.4	5.4	11.4	3.7	4.4	3.6
AT27	16.7	13.9	25.2	17.6	20.0	25.8	19.2	13.5	8.4	13.9	17.0
AT28	27.4	15.8	24.3	24.5	20.8	29.9	16.3	17.8	10.5	14.2	23.6
AT29	9.5	19.0	5.7	6.9	8.8	2.6	4.0	7.0	2.3	3.8	3.5
AT30	12.0	4.9	5.4	4.2	25.5	3.9	3.4	6.8	2.8	2.2	2.3
AT31	29.6	8.1	9.1	8.0	12.0	9.7	8.2	20.8	5.7	5.4	4.7
AT32	9.7	2.7	2.7	3.3	26.3	3.6	4.7	4.6	2.4	3.1	2.4
AT33	30.0	25.5	14.2	6.6	38.7	21.1	17.1	19.9	6.9	7.3	4.7
AT34	37.4	17.4	10.8	4.9	43.3	12.8	16.4	19.9	8.5	7.7	5.0
AT35	12.4	6.8	11.5	12.4	7.8	8.1	14.5	5.6	6.3	16.6	9.6
AT36	14.7	14.0	12.2	15.1	23.4	10.8	17.0	11.0	5.2	5.9	4.9
AT37	14.2	8.2	8.1	4.2	14.8	6.2	13.4	9.3	4.2	5.5	5.1
AT38	46.7	5.1	4.6	5.0	7.7	6.9	6.4	10.3	4.7	3.3	3.8
AT39	10.7	9.3	4.7	2.7	7.1	7.4	3.3	11.6	3.5	3.1	5.0
AT40	34.4	9.8	6.1	8.7	12.8	13.7	9.0	17.7	5.7	3.3	3.8
Tor.	2.4	2.1	2.9	1.7	1.6	1.6	1.7	1.8	1.7	2.4	2.9

Table 11: VIF (Variance Inflation Factor) values of 40 atoms and torsion angle derived using multicolinearity diagnostics for the experimental $\Delta\Delta GH_2O$ values from chemical denaturation experiments.

Atom/ Torsion	ASA+SS Classified Statistical Potentials for Various Structural Regions										
	Helices				Sheets			Others			
	0-2	2-30	30-60	60+	0-5	5-35	35+	0-10	10-42	42-67	67+
AT1	0.655		0.871		0.420	0.746			0.504	0.421	0.184
AT2		-1.149		-0.456	1.168	2.305	-0.361		1.404		-0.486
AT3					-0.745		-0.310	0.794	-0.497	-0.892	
AT4		0.790				-1.536	1.173				
AT5	-0.225	-0.775			1.115	0.323	-0.493	-0.643			-0.272
AT6					-0.160	0.562	-0.151	0.448		0.498	
AT7		0.613	0.436		-0.424	-0.366		-0.979	-0.380	-0.809	
AT8	-0.472		-0.786		-0.594		-0.250	-0.226	0.362	0.418	0.101
AT9	-1.337	2.540	1.129			0.765		-1.549	-3.405	0.327	
AT10		-0.904		1.243		-1.430			-1.308		0.626
AT11	-1.317	2.661	0.547		-1.420	-0.384	-0.557		-0.701	1.267	-0.856
AT12	0.336	-0.849	-0.237		0.184					-0.586	
AT13	-4.278			1.037	2.441		1.038		1.224		-0.598
AT14		3.929			1.259	-0.476	-1.797	2.749	-2.341	0.845	-1.001
AT15		1.460			-1.487		0.835				2.103
AT16		-0.690	-0.992			0.575	-1.037	-0.801	-1.038		-0.829
AT17		1.176									0.583
AT18					-1.125		-0.770		0.618		
AT19	-0.932		-0.815			-1.446	0.768		1.127	2.067	0.778
AT20	0.987	-2.315			-2.169	-2.009	0.454	1.186	0.983	-0.395	
AT21		1.784							-2.052		
AT22	1.677	-0.808	-0.592		-0.609	0.800	-1.102	1.836	-1.094		
AT23								1.310	-0.665	0.632	
AT24			-0.492	-0.642	0.771		-2.157		1.464	-0.906	
AT25	-4.445				-1.918	-2.035				1.765	0.379
AT26	1.949	2.915	0.693	-0.892			2.242	-1.033		1.615	
AT27	-3.259	-1.543	1.559		-2.047	-0.712			-0.569	-0.541	
AT28	1.315		-0.996	-0.282	0.576	0.921				0.470	
AT29	3.706				-1.704	1.060	0.253	0.617		-1.167	
AT30	-2.015	-1.052	-0.724	1.582	2.872	-1.380		2.128	1.197		
AT31		1.092		-0.449		0.593				1.443	0.821
AT32			0.531			0.976			0.983	1.498	
AT33	2.308		0.459	0.558			1.239				
AT34	-1.643				1.846	1.027		1.585		0.541	
AT35		1.800	1.197		2.159	1.276			-1.368		
AT36	-2.064		1.153		1.154			-1.350	1.089		
AT37			0.686	0.403	-1.728	-0.648	1.613		1.256	-1.004	-0.485
AT38	4.729			0.841		1.546	-1.053	-1.827	-2.110	-1.327	
AT39		-3.194		-0.328	-2.593		2.254	-3.190		-1.258	1.693
AT40					1.226		0.485				-0.237
TOR	-0.850		-0.261	-0.453	-0.337	-0.690	-0.328			-0.106	-0.312
Int	-1.921	-1.280	-0.483	-0.192	-1.252	-1.401	-0.476	-1.244	-1.094	-0.921	-0.268

Table 12: Atoms selected (for thermal $\Delta\Delta G$) using stepwise regression and their regression coefficients. Int: Intercept. TOR: Torsion.

Atom/ Torsion	ASA+SS Classified Statistical Potentials for Various Structural Regions										
	Helices				Sheets			Others			
	0-2	2-30	30-55	55+	0-3	3-25	25+	0-15	15-42	42-67	67+
AT1	-0.552	-0.256	0.491			-1.064	-0.219		-0.220	-0.098	0.613
AT2		1.174		-0.558	-1.624		1.272				
AT3		1.178					0.842		-0.247		-0.793
AT4		-1.960	-0.665	-0.558	0.043		-1.358				-0.351
AT5		0.363	0.145	0.539			0.335	-0.464			0.556
AT6		0.243	-0.128	0.185	0.427	0.561	0.305	0.612	0.305		
AT7	0.564	-0.320			-0.688						-0.674
AT8		0.284		-0.126				0.265	0.186		0.199
AT9			-0.288	-0.358	-2.596	-2.674	-1.142				3.989
AT10					1.284	1.321	4.196	-1.713	1.090	-0.973	-0.725
AT11	0.807		-0.539	0.233	1.406		1.262		0.523		-1.607
AT12			0.252		-0.212	0.448		0.597			
AT13	-1.275		-0.426		-1.415	1.805			0.546		2.014
AT14	1.096	1.413	-0.574	-0.532					0.648	-0.573	
AT15				-0.947		-1.946	-1.550	2.303	0.996		1.015
AT16	-1.183	0.602	0.711			0.960		-2.656	1.362		
AT17		-1.913		-0.515	-1.982	-0.839	1.941		-1.534		
AT18	1.870		-0.541				-1.110	-1.286	-0.782		
AT19	-0.151		-0.374		-1.705	2.014	1.625	1.228	0.627		0.499
AT20		0.551			-1.246			-1.952		0.807	
AT21			-0.675			-1.412			-0.469	-0.366	1.887
AT22			-0.309				-0.659	-1.650	-0.463		-1.271
AT23	1.390						-2.992	-5.258	2.189		-1.211
AT24	0.788	1.027				-2.038	1.347	1.556		-0.505	-0.999
AT25	2.833		1.951	0.271			-1.676	0.662		-2.170	
AT26			-0.968				3.164		-1.650	-1.128	1.109
AT27			0.812		-2.495		3.013		-0.532	1.763	1.060
AT28	-0.255			0.107	1.091	1.422	-1.543		0.189	-0.735	-0.337
AT29		0.850		0.268	1.485						0.499
AT30	1.429	-0.437	0.497						-1.185		-1.440
AT31	1.131	0.969			1.596		1.212	-3.057		-0.343	
AT32	-1.731	1.534	-0.511			1.235	-2.494	1.948			2.536
AT33	-0.700	0.495	0.984				0.917	1.567			-0.561
AT34	1.003	0.956	-0.736	0.335	0.897	-0.973	-1.256	-2.007			
AT35			-0.781	0.205	2.204	2.828		2.135			0.594
AT36			0.439	0.393		2.823	1.346	3.953	0.795	-0.387	1.067
AT37	1.739		1.129		1.985	-1.775	-0.737	-2.237		0.941	0.462
AT38	-2.781				-1.243			5.780		2.004	0.748
AT39	-1.147	-1.987		0.726	1.250			-1.315	-1.223	0.683	
AT40	-1.503	-1.392			-2.181	1.565	-2.411	1.769			0.493
TOR			1.012		0.617	0.173	0.772	0.419	0.241	0.241	0.131
Int	0.810	0.719	0.139	0.186	1.950	-0.192	-0.070	1.129	0.518	0.097	-0.323

Table 13: Atoms selected (for $\Delta\Delta GH_2O$ values) using stepwise regression and their regression coefficients. Int: Intercept. TOR: Torsion.

Model	Helices				Sheets			Others			
	0-2	2-30	30-60	60+	0-5	5-35	35+	0-10	10-42	42-67	67+
VIF50	35	35	38	36	34	36	36	36	38	38	32
VIF30	33	35	35	30	24	32	34	26	34	35	31
VIF20	28	30	28	24	19	31	29	20	33	32	27
SRM1	21	21	25	17	23	20	28	25	23	16	29
Model	0-2	2-30	30-55	55+	0-3	3-25	25+	0-15	15-42	42-67	67+
SRM2	21	21	20	13	28	26	24	18	25	25	18

Table 14: Reduction of atoms using statistical models: VIF50, VIF30 and VIF20 indicate selection of atoms with multicolinearity diagnostics. These three models use VIF cutoff values of 50, 30 and 20 respectively. SRM1 and SRM2 indicate the selection of atoms with stepwise regression model for the $\Delta\Delta G$ and $\Delta\Delta GH_2O$ mutation datasets respectively.

	Helices				**Sheets**				**Others**			
CL12A	0-2	2-30	30-60	60+	0-2	2-30	30-60	60+	0-2	2-30	30-60	60+
Mutants	238	207	205	117	68	98	106	22	49	103	168	162
CL11	0-2	2-30	30-60	60+	0-5	5-35		35+	0-10	10-42	42-67	67+
Mutants	238	207	162	160	79	80		130	97	139	138	108

(a)

	Helices			**Sheets**			**Turns**			**Others**		
CL12B	0-2	2-40	40+	0-2	2-40	40+	0-2	2-40	40+	0-2	2-40	40+
Mutants	238	285	244	68	162	64	26	35	86	23	108	204

(b)

Table 15: No. of mutations (with thermal $\Delta\Delta G$) allocated according to the classification system using accessible surface area and secondary structure specificity. (a) CL12A and CL11 classify the mean force potentials and mutations into 12 and 11 structural regions respectively. (b) CL12B classifies turns separately with reduced number of ASA ranges.

	Helices				**Sheets**			**Others**			
CL11D	0-2	2-30	30-55	55+	0-3	3-25	25+	0-15	15-42	42-67	67+
Mutants	105	162	115	113	196	213	121	175	153	129	99

Table 16: No. of mutations (with $\Delta\Delta GH_2O$ from chemical denaturation) allocated according to the classification system using accessible surface area and secondary structure specificity. CL11D classifies the mean force potentials and mutations into 11 structural regions.

4.3 Prediction Model Analyses

4.3.1 Comparison of Structural Training Datasets

Six datasets (PI-1 to PI-6) were derived from PISCES only for the purpose of comparison. As described in the methods (section 3.1), it is mandatory to include at least one to few representatives for all possible protein structures so that their structural information is available for analysing the structural changes upon point mutations to a maximum extent. If the structural training dataset contains more than required homologous representatives for any specific protein, it may end up in generating unnecessary noise for the prediction model.

Correlation coefficient and prediction efficiency for the mutations (1518 mutations from thermal experiments) with $\Delta\Delta GH_2O$ were observed (Table 17a). CL11 method was used for classifying the structural regions. For all the mutations together, correlation coefficient ranged from 0.72 to 0.73 for all the datasets. Similarly, the prediction accuracy also showed values ranging from 83.84% to 84.09% for all the datasets (Table 17a). Since the observed overall prediction efficiency could not distinguish the datasets, the results were observed separately for different structural regions in proteins. The correlation coefficient and prediction efficiency were shown in the fig 12. In different regions of the proteins, different datasets exhibited improved correlation over the others. So, it was decided to check the prediction efficiency of datasets on all the 11 bins to know how many times the correlation of a specific dataset has been overtaken by any of the remaining datasets (Table 17b). PI-1 and PI-2 datasets were overtaken 10 times in 11 bins by the remaining datasets. PI-3 dataset was overtaken only 7 times by other datasets. Interestingly, PI-4, PI-5 and PI-6 datasets were overtaken only 3 times by any other datasets consistently. This means that the datasets PI-1, PI-2 and PI-3 were not enough to furnish maximum structural information for some regions in proteins. On the other hand, having excess amount of structural information provides noise in the prediction model developed from bigger datasets. This noise was consistently maintained at the same level in the datasets PI-4, PI-5 and PI-6

having 40%, 45% and 50% maximum sequence identity cutoff respectively. Thus, the usage of a structural training dataset with 50% maximum sequence identity for the whole prediction model is validated. Alternatively, different structural training datasets with variable maximum sequence identity among themselves can also be used for different structural regions in proteins. However, only one dataset was used in order to reduce the complexity of the prediction model construction. Some additional proteins (other molecule binding proteins) were removed from these datasets to reduce the noise to a maximum extent. However, PI-7 was used in all validation tests where only less number of proteins (Table 4) were removed.

Datasets	CC	PA (%)	TP	TN	FP	FN	Sens	Spec	Std. Err.
PI-1	0.72	84.09	1204	144	188	67	0.86	0.95	0.979
PI-2	0.72	83.91	1201	144	188	70	0.86	0.94	0.979
PI-3	0.73	83.91	1196	149	183	75	0.87	0.94	0.978
PI-4	0.73	83.84	1196	148	184	75	0.87	0.94	0.977
PI-5	0.73	83.84	1196	148	184	75	0.87	0.94	0.977
PI-6	0.73	83.84	1196	148	184	75	0.87	0.94	0.977

(a)

Structural Training Datasets	No. of times (in 11 bins) overtaken by remaining datasets
PI-1 (25%)	10 times
PI-2 (30%)	10 times
PI-3 (35%)	7 times
PI-4 (40%)	3 times
PI-5 (45%)	3 times
PI-6 (50%)	3 times

(b)

Table 17: Comparison of structural training datasets: (a) Correlation coefficient (CC) and overall prediction accuracy (PA%) were compared. (b) No. of times a specific structural training dataset is overtaken by other datasets in providing better a correlation with experimental $\Delta\Delta GH_2O$ values. Sens and Spec mean Sensitivity and Specificity; TP- True Positive; TN – True Negatives; FP – False Positives; FN – False Negatives (Appendix A).

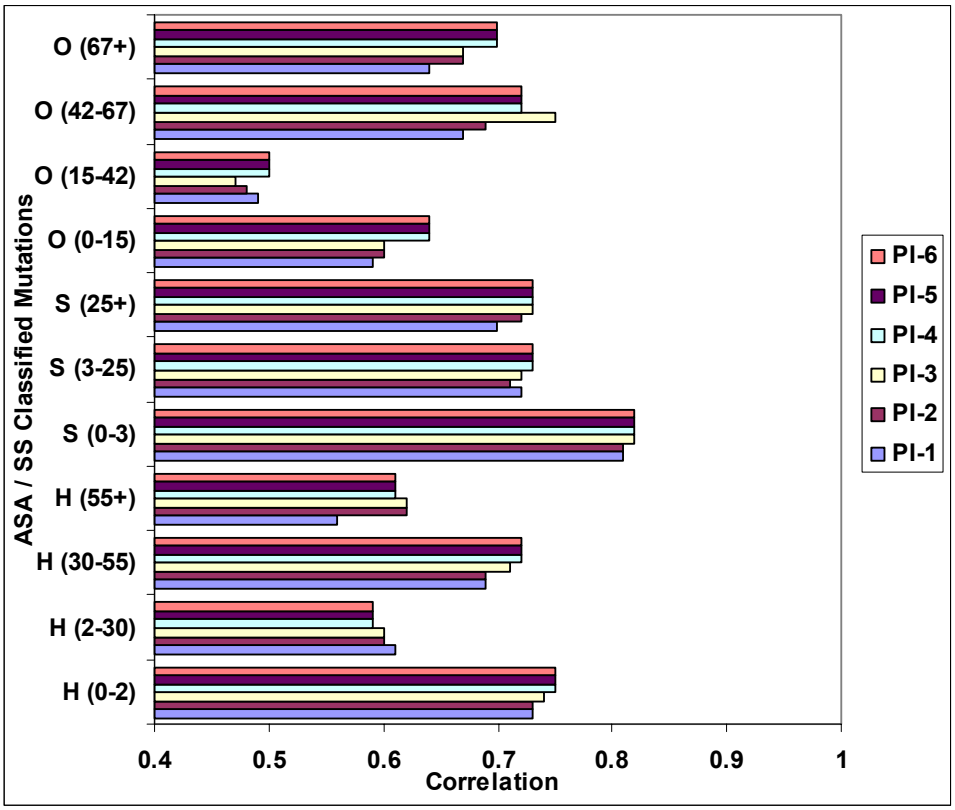


Fig. 12: Comparison of maximum sequence identity cutoff for the structural training datasets. PI-1, PI-2, PI-3, PI-4, PI-5and PI-6 indicate datasets with 25%, 30%, 35%, 40%, 45% and 50% maximum sequence identity respectively.

4.3.2 Comparison of Atom Classification Models

For the purpose of using the best atom classification model for prediction, five different atom classification models were taken and the statistical mechanics setup was implemented with all these models separately.

This comparison was one of the early implementations to select a specific atom classification system for the prediction model. The ability to provide good correlation and prediction accuracy with experimental $\Delta\Delta G$ values was tested with all the mutations together. Multiple linear regression model was used to apply the derived potentials to the subsequent point mutations selected from Protherm (Gromiha et al. 1999a; Bava et al. 2004) and literature. Correlation and prediction accuracy were predicted using a dataset of 1543 mutations for all the five atom classification models. For the mean force potentials and mutations, the classification CL11 was used.

The correlation coefficient (Fig. 13a) and prediction accuracy (Fig. 13b) of all the atom models were derived separately and compared (Table 18). The MF40 showed the best results among all the atom classification models followed by the SA28 atom model. The former showed a correlation of 0.85 with 84.06% of the mutations correctly predicted out of 1543 mutations. SA28 model showed a slightly reduced correlation of 0.82 with 82.96% of mutations correctly predicted. Correlation and prediction accuracy gradually reduced for other atom models that had less number of atoms classified.

It can be interpreted that the size of the atom model is directly proportional to the increase in correlation. This is due to the elaborate definition of protein environment of any bigger atom model. However, a statistical problem of overfitting of the atom types cannot be averted for a bigger atom model definition, since the multiple regression has too many parameters (predictors or atoms) offered by a bigger atom model. An absurd and false model may fit perfectly if the model has enough complexity by comparison to the amount of available mutation data.

Atom Models	Correlation	PA (%)	TP	TN	FP	FN	Sens	Spec
Basic-5	0.55	75.7	173	995	88	287	0.66	0.38
AACα	0.76	79.46	258	968	115	202	0.69	0.56
LN24	0.78	81.08	272	979	104	188	0.72	0.59
SA28	0.82	82.96	301	979	104	159	0.74	0.6
MF40	0.85	84.06	311	986	97	149	0.76	0.68

Table 18: Comparison of atom classification models: Correlation and prediction efficiency of 5 different atom types.

So, the solution for this problem was to use a smaller atom classification model like SA28 atom model, since the correlation from this model was nearly close to the correlation of the MF40 atom model. This may still exhibit insufficiency in finalising the model, because this model reduces the ability of getting higher correlation with experimental ΔΔG. However, several issues still remain unresolved, such as the cutoff value for the size of the atom model that can

produce the best correlation with experimental ΔΔG without producing overfitting effect for the statistical multiple regression model.

Conversely, a statistically reduced atom model can be used where the correlated atom types of the bigger atom model are clubbed using a statistical criterion. This concept was implemented using the multicollinearity diagnostics (sections 3.7.3, 4.2.5). Besides, selection of specific atoms can also be carried out using the statistical significance of atom types, as described in the stepwise linear regression and selection model (sections 3.7.4, 4.2.6). Both of these statistical models provided good correlation with experimental ΔΔG, where the reduced MF40 atom classification system performed better than the other atom models (Fig. 11). Thus, the dimensionality reduction of the atom classification model with minimised overfitting effect proved to be efficient for final prediction of protein stability.

To get further insight on the ability of these atom models, protein environment specific prediction efficiency was also analysed. The prediction algorithm using these atom models showed a good correlation for the mutations in the buried and exposed region compared to partially buried region of the protein. For the MF40 type model, a correlation of 0.84 was observed for the mutations in the buried helix regions (Fig. 14a ASA/SS classified structural region 1). The correlation slightly decreased to 0.81 and 0.71 for the mutations in the partially buried region of protein (Fig. 14a: ASA/SS classified structural regions 2 and 3 respectively). However, the correlation increased in the exposed region of the helices (Fig. 14a: structural region 4). Similar effect was observed for all the other atom models in different structural regions.

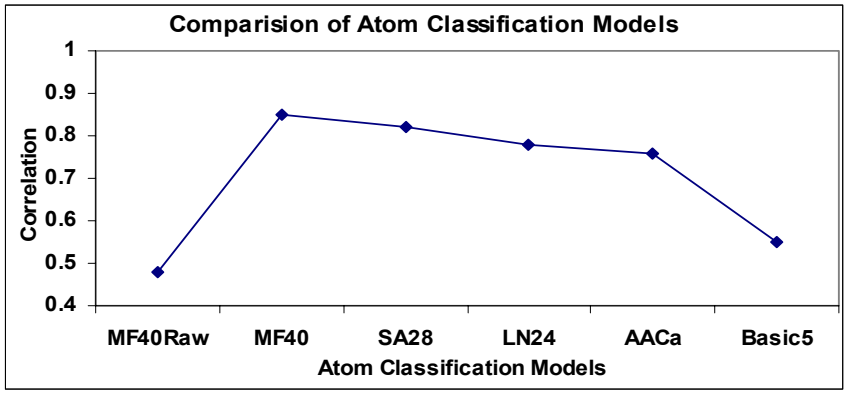
A decrease in the correlation between experimental and theoretical ΔΔG was observed in the partially buried regions of the protein for all the models (Fig. 14a) where the correlation coefficients were analysed based on the secondary structure specificity. It can be clearly seen that all the atom type models predict mutations in buried and exposed regions very well compared to the partially buried region. Due to high conservation of atom distribution in compact structural regions of proteins, the prediction model showed consistently good

results in buried structural regions. In the partially buried helix residues, conservation of atom distribution is comparatively low. Yet, the number of stabilising residues were more in number than the destabilising residues. The prediction model shows slightly decreased correlation and prediction accuracy in that region because it could not assess the stabilising effect of some of these residues, since the parameters from atom potentials were not effective.

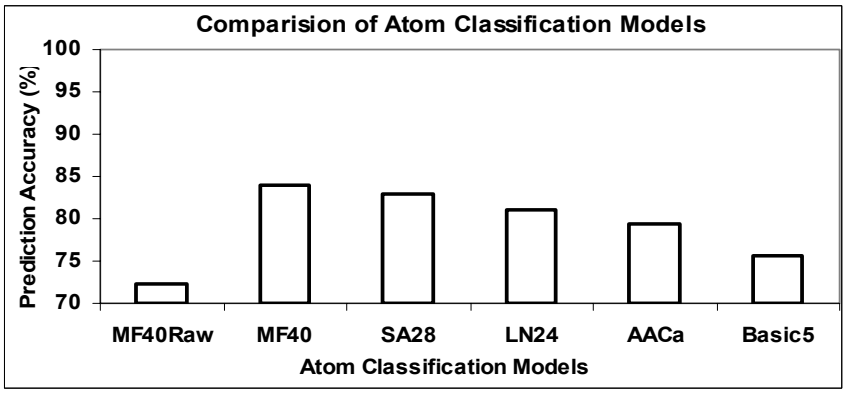
Parallel and antiparallel β-sheets are significantly different in their hydrogen bonding patterns. They were not distinguished because there were only fewer mutations in sheets. Distribution of these mutations into different structural regions that can be identified clearly as partially buried and exposed was quite difficult due to this reason. As observed in helices, there were more stabilising residues in turns and coils that exist in the partially buried region. These residues achieve stability due to the formation of favourable new interactions due to flexibility in the partially buried region.

Meanwhile, statistical potentials are better than empirical energy functions in assessing the long range interactions. Exposed turns and coils are highly flexible regions in proteins. These are mainly stabilised by long range interactions. Due to this reason, they mainly initiate unfolding process even in slightly changes in environmental conditions. Stabilising and destabilising mutations were equal in number and easily distinguished in this region.

The correlation for the mutations in sheets in the partially buried region (ASA/SS classified structural region 6 in fig. 14) for the MF40 model is low (correlation coefficient = 0.78), compared to the correlation (0.82) observed from the SA28 model. This was the only exceptional case and may be due to overfitting of the data in statistical techniques. This behaviour further supports the necessity of dimensionality reduction techniques to optimise the size of atom models. Prediction accuracy (%) was found to be similar to the observed correlation coefficient between predicted and experimental ΔΔG, except in some structural regions. However, correlation is given importance in such cases because a high correlation with ΔΔG always supports majority of mutations to be correctly predicted as stabilising or destabilising, but it’s not vice versa.

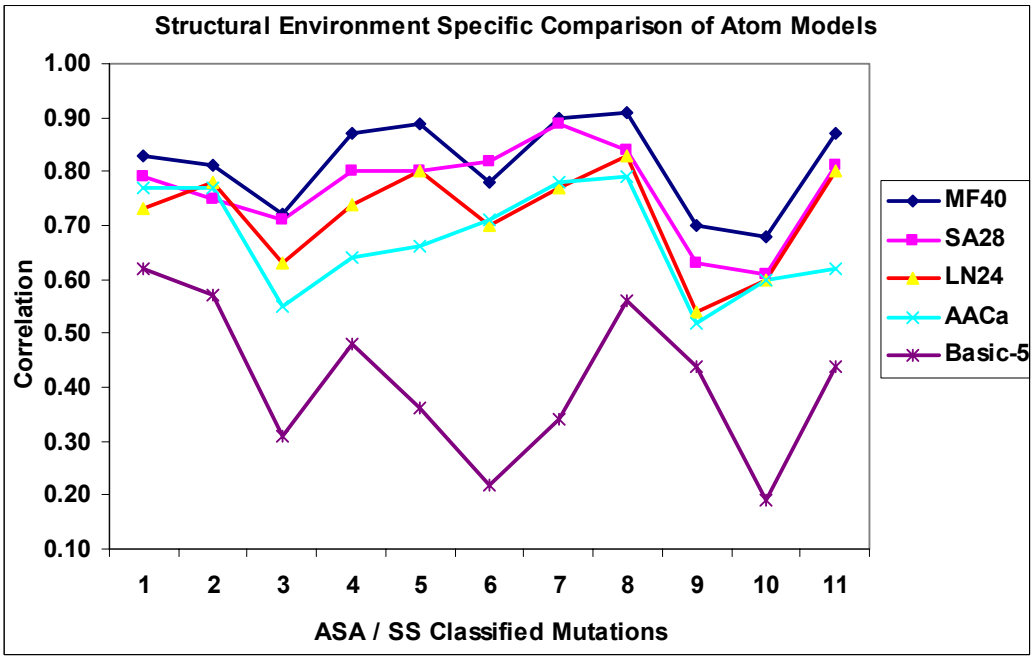


(a)

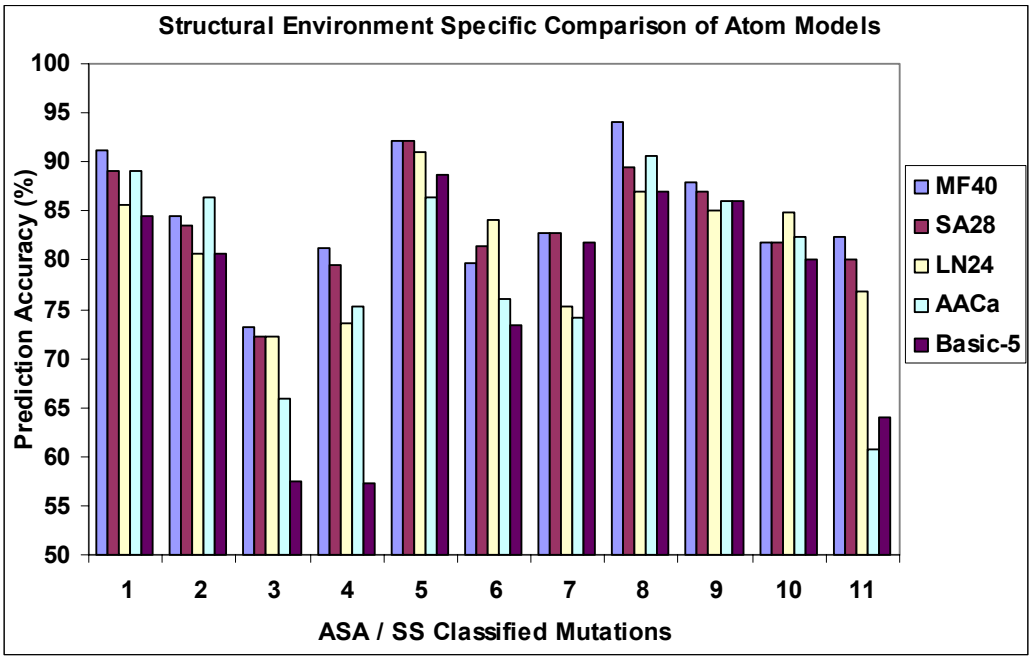


(b)

Fig. 13: Comparison of 5 different atom classification models used in the prediction of changes in protein stability. Overall (a) Correlation and (b) prediction accuracy for predicting thermal ΔΔG values.



(a)



(b)

Fig. 14: Structural environment specific comparison of 5 different atom classification models used in the prediction of changes in protein stability. (a) Correlation and (b) prediction accuracy for predicting thermal ΔΔG values. Prediction efficiencies in 11 different structural regions were compared using CL11 classification method (Table 15).

4.3.3 Effect of Torsion Angle Potentials

Torsion angle potentials were considered as one of the effects to construct the prediction equation with other effects being the 40 different atom types. Torsion angles of amino acids from the structural training dataset and mutations were not classified following the same classification mechanism implemented for radial distribution because the torsion angle distributions for different secondary structure regions are quite different from each other. It also does not make a huge difference for the residues that have variable compactness.

Improvement of correlation (Fig. 10a) with experimental ΔΔG and prediction accuracy (Fig. 10b) show a slightly increased efficiency of including torsion potentials. For the unclassified mutations in Raw dataset, correlation increased from 0.49 to 0.52. In the CL12A classified dataset, the correlation was observed to be 0.85. Though the increment in correlation is low, the predicted

ΔΔG values are slightly adjusted close to the real ΔΔG values which eventually results in minor increase in prediction accuracy from 84.45% to 84.77%. To ensure the reliability of the prediction model against overfitting effect, CL11 dataset classification was used with torsion potentials where the correlation remained 0.85 (Fig. 3a) for all 1538 mutations with a slightly decreased prediction accuracy of 84.06% (Fig. 3b). CL11 was then selected for further comparison of statistical models and validation tests. The minor difference in prediction accuracy might have resulted due to overfitting effect of the initial multiple regression with all the 40 atoms used for the prediction model.

Since the torsion angle potential was not colinear with the atom distribution, it was always selected as one of the effects to be included in the final prediction model. However, torsion was not needed for the thermal stability of few structural regions (partially buried alpha helices, buried and partially buried turns/coils), since the atom potentials provided the maximum information and torsion potential was removed by the automated selection of stepwise regression (stepwise regression and selection in Table 12).

4.3.4 Gaussian Apodisation

The torsion angle potential derived from the PISCES dataset was analysed with and without Gaussian apodisation. Atom potentials were maintained without any change for this comparison. The overall correlation with experimental energy values for 1538 mutations showed that Gaussian apodisation around the peaks of the torsion potential showed only a narrow improvement (Fig. 15a). Torsion angle potential with and without the implementation of apodisation function showed correlation coefficient of 0.82 and 0.81 respectively (Fig. 15a). But, the prediction accuracy increased from 81.51% to 83.28% (Fig. 15b) after using the apodisation function. Due to this increment, it was evident that mutations in a certain environment in protein retain stable conformation with altered torsion angles observed for any specific amino acid.

However, it was then decided to check the effect of Gaussian apodisation on the mutations after the ASA and secondary structure based classification. It

performs clearly better compared to the model without Gaussian apodisation. The mutations in sheets with more than 35% solvent accessibility are predicted with a correlation of 0.92 (Fig. 16a) using the PI-7 dataset with apodisation. All other datasets gave a correlation ranging from 0.84 to 0.85 in this protein region (Fig. 16a).

Thus, it can be proved that the mutations in specific protein region adapt altered torsion angle conformation, if there is significant difference in correlation coefficient observed from any of those classified bins after the Gaussian apodisation.

4.3.5 Evaluating Structural Training Datasets for Torsion Potentials

Prediction efficiency of thermal ΔΔG was observed between the structural training datasets to assess their ability. PI-7 dataset (PISCES with 50% maximum sequence identity) performs better on all regions of proteins with a good correlation coefficient comparing to Top500 and SCOP-ASTRAL datasets. The mutations in sheets with more than 35% solvent accessibility are predicted with a correlation of 0.92 (Fig. 16a) using the PI-7 dataset with apodisation. All other datasets, including PI-7 without Gaussian apodisation, gave a correlation ranging from 0.84 to 0.85 in this protein region (Fig. 16a).

On the other hand, it suffers slightly in certain regions [solvent exposed helices and buried region in others (non helices/non sheets)] due to edge effect (section 3.5.2) discussed previously (Fig. 6) during the optimisation of torsion potentials. While increasing the maximum tapering angle of the Gaussian apodisation favours more amount of altered torsion angle conformations to be predicted efficiently, it also disfavours by producing edge effect (Fig. 6). In spite of this problem, a maximum tapering angle of 7° was used with a balance between these positive and negative effects. Thus, Gaussian apodisation exhibits increased efficiency in the prediction model.

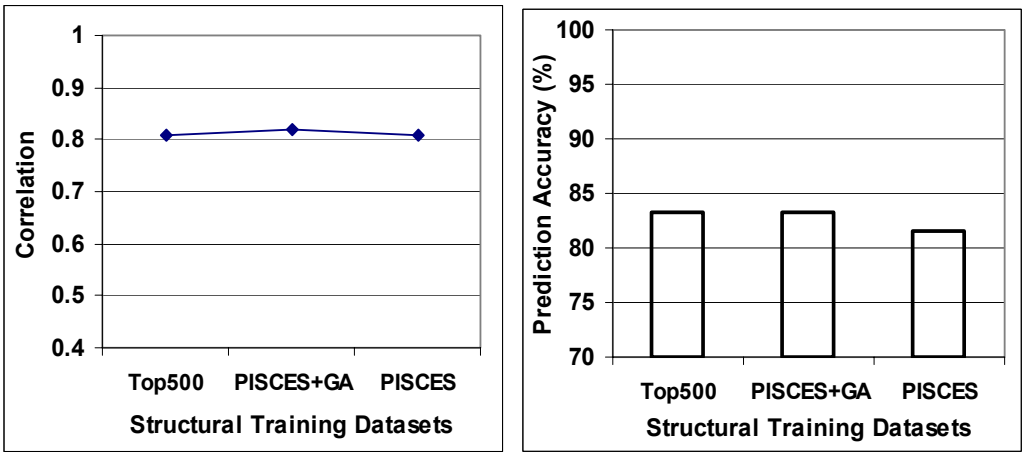


Fig. 15: Comparison of structural training datasets for their efficiency to render torsion angle potentials. Atom potentials are maintained constant for all these validation.

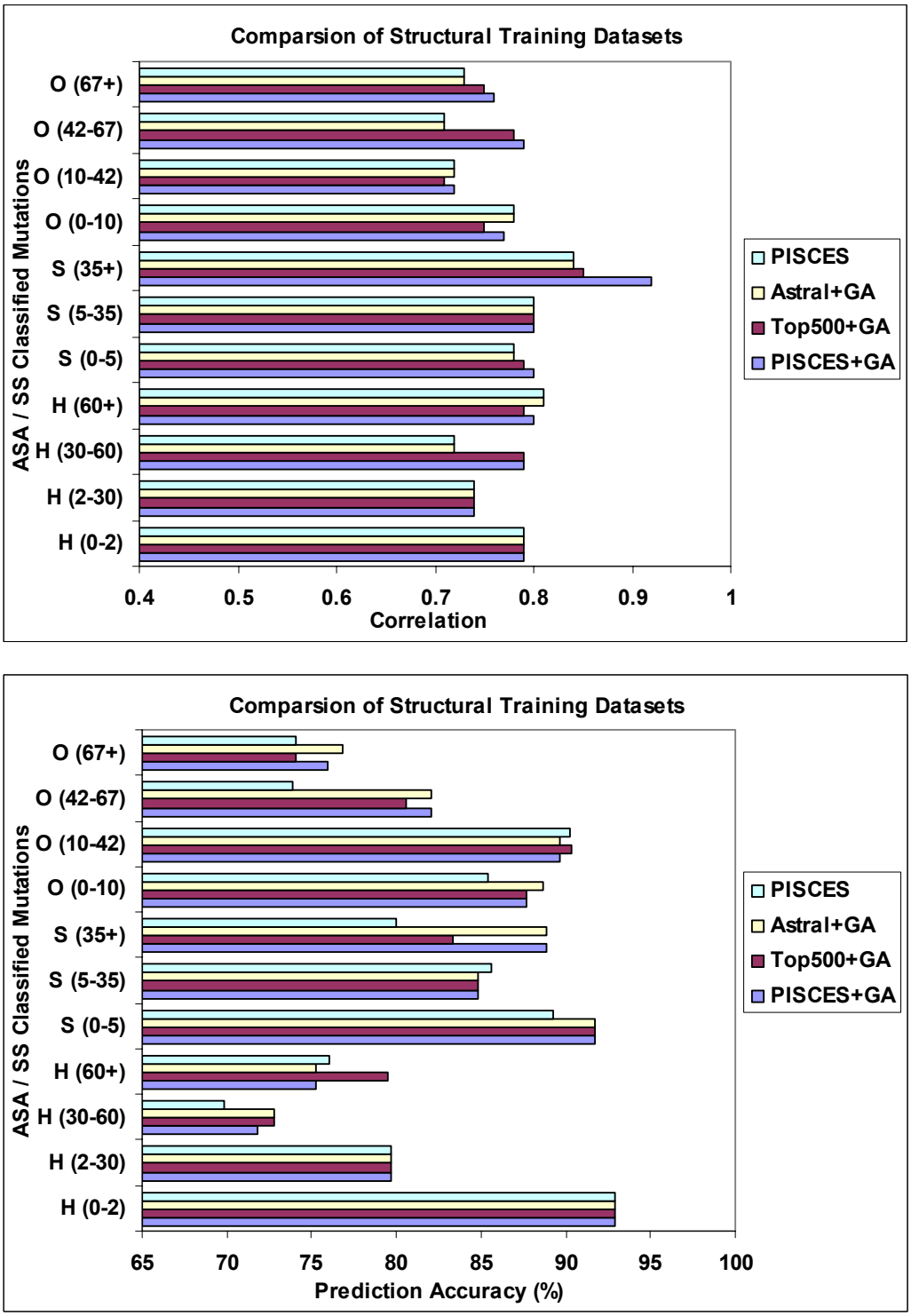


Fig. 16: Comparison of structural training datasets to render torsion angle potentials. For the atom potentials, PISCES (Pi-7) was maintained as constant. For the torsion potentials, 3 datasets were compared PISCES, SCOP-ASTRAL and Top500 were compared. (a) Correlation and (b) prediction accuracy for predicting mutations with thermal $\Delta\Delta G$ values. PISCES-GA and PISCES denote the PI-7 datasets used with and without Gaussian apodisation.

4.3.6 Distinguishing the Structural Regions

The secondary structure and accessible surface area based classifier was optimised for the number of mutations in each structural regions so that they are at least three times more than the number of atoms and torsion potentials together (Tables 14, 15 and 16 in prediction model construction). For this, several ranges of ASA have been tested for mutations in different secondary structure and the range that provides enough mutations (three times more than the number of predictors) in each bin has been finalised. Initially, a mutation dataset of 1236 mutations was used for the analysis. Later, more mutations were added from literature and increased the number to 1543 mutations totally. But, 5 mutations were then removed from this, since they had extreme values of $\Delta\Delta G$, which may disturb the regression during validation tests.

The statistical potentials were then derived subsequently using the same secondary structure and ASA parameters. By using these flexible values for maintaining the minimal number of mutations, the ability of getting high correlation with experimental values and prediction efficiency of stabilising or destabilising mutations were never altered. When the overfitting problem of statistical models was reduced by using the stepwise regression of pair potentials and torsion angle potentials, the performance of the classifier remained stable for all the mutations in all the bins. Thus, it was concluded that the values of relative ASA can be flexibly altered to accommodate the increasing number of mutations in future, while the prediction specificity of the mutations from subsequent potentials is directly proportional to shrinking the bin size of relative ASA. Since none of the experimental conditions were used for distinguishing the mutations, the classifier exhibits high availability and flexibility for the analysis of predicting protein stability changes upon point mutations.

A significant improvement of the correlation (0.81) and prediction accuracy (83.35%) for CL9 was noticed, when compared to the ASA and SS unclassified system (Table 19). This classification proved to be very efficient and comparable, because the Boltzmann's energy values were derived for the bins

separately and used to predict the stability values (ΔΔG) from the mutations of the respective structural regions.

Comparing to CL9, CL12A_2 and CL12B_2 (Table 19) showed better correlation and prediction accuracy (%). Structural regions of CL12B_2 classification showed slightly high correlation because the turns were considered separately in the secondary structure implementation, when compared to CL12A_2. But, the correlation obtained from CL12B_1 dropped down, since there were very few mutations in sheets with a particular ASA range in CL12B_1 (CL12B in table 15 and CL12B_1 in table 19 are same).

However, out of all these methods, CL12A_2 showed good overall prediction accuracy, sensitivity and specificity. Though CL12B_2 showed maximum correlation, its prediction accuracy was slightly low. These two methods were compared (CL12A and CL12B) with 1538 mutations in the previous chapter (section 4.2.4). Since the mutations in sheets were less in number, the regions based on the ASA range (in the sheets) were divided into 3 instead of 4 ranges. Thus, the finalised method (CL11) defines 11 structural regions for the amino acids of mutations and structural training datasets. Besides, this method was implemented on 1543 mutations. It showed consistently good correlation and prediction accuracy (Fig. 17). So, this classification was used for the further prediction of mutant stability.

Classification Systems	**Correlation**	**PA (%)**	**TP**	**TN**	**FP**	**FN**	**Sens**	**Spec**
	1236 mutations							
Raw	0.53	72.13	701	161	247	86	0.74	0.89
CL9	0.81	83.35	692	304	104	95	0.87	0.88
CL12A_1	0.84	84.85	698	316	92	89	0.88	0.89
CL12A_2	0.86	86.28	706	325	83	81	0.89	0.9
CL12B_1	0.83	85.19	704	314	94	83	0.88	0.89
CL12B_2	0.88	84.59	572	246	75	74	0.88	0.89
	1543 mutations							
CL11	0.85	84.06	311	986	97	149	0.76	0.68

Table 19: Optimisation of the classification of different structural regions.

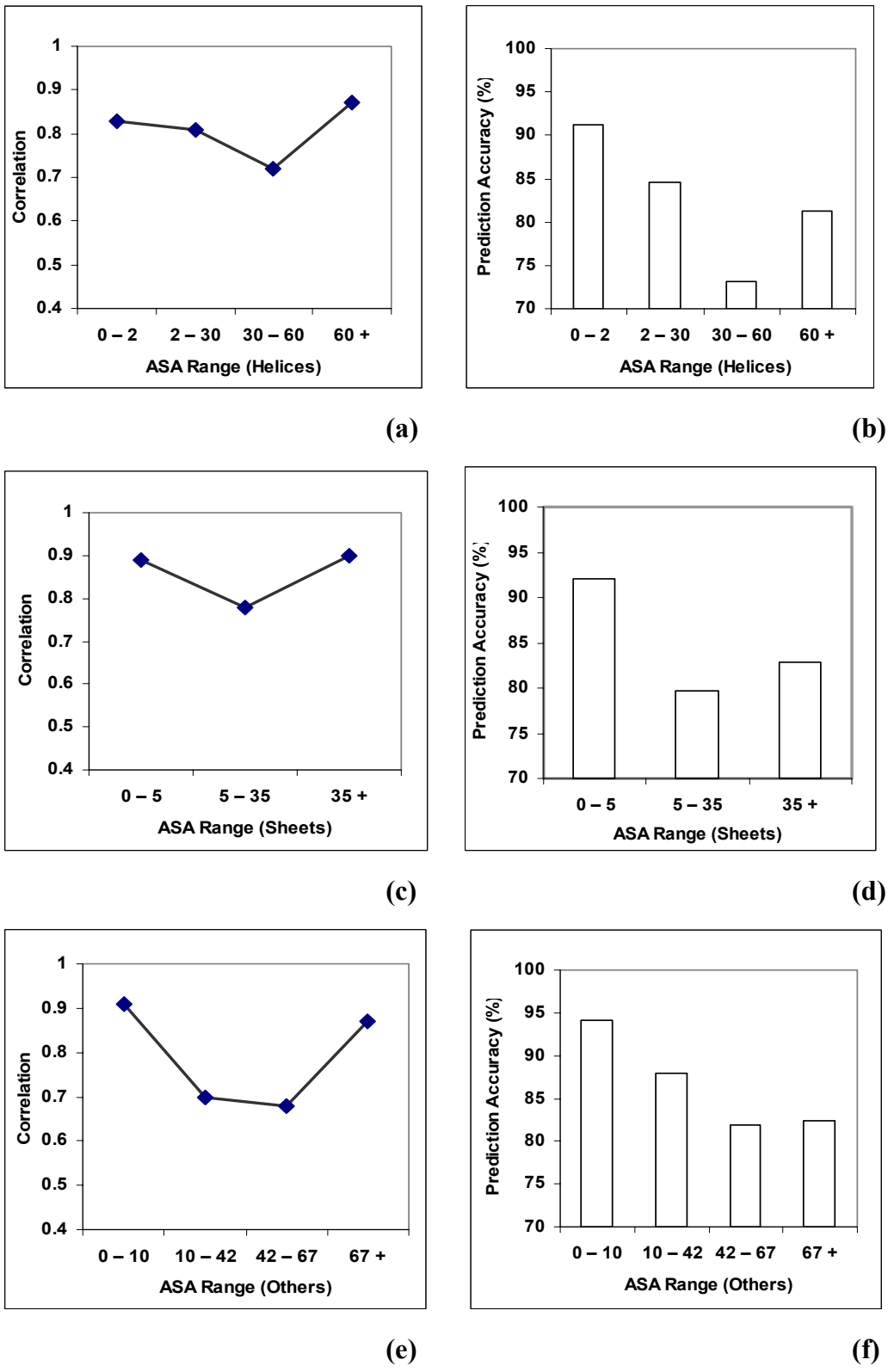


Fig. 17: Correlation and prediction accuracy of 1518 mutations with thermal ΔΔG classified into 11structural regions. (a)(b) helices, (c)(d) sheets and (e)(f) others.

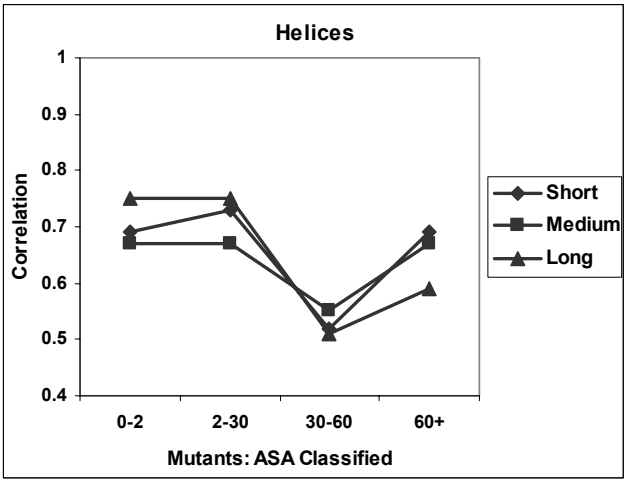
The prediction model with CL12A_2 classification showed more true positives and true negatives compared to the other classification. Out of 789 positive ΔΔG values, it predicted 706 positive ΔΔG values and out of 406 negative ΔΔG values, it predicted 325 negative ΔΔG values correctly, because the experimental energy value with positive values were more than the negative values in the 1236 mutation dataset. But the opposite was observed in the CL11 classification: there were more negative ΔΔG values than positive ΔΔG values, but the prediction algorithm predicted a consistently good correlation and prediction accuracy for 1543 mutations.

From these results, it can be concluded that the program predicts with 84.06% of accuracy (CL11 in table 19) for stabilising as well as destabilising ΔΔG. Therefore, optimising the number of structural regions improved the prediction model by providing a reliable and accurate prediction model. It also helped in analysing the behaviour of the interactions in the different secondary structures with different ASA ranges. Thus, the stability change strongly depends on the location of mutation with respect to secondary structures and solvent accessibility.

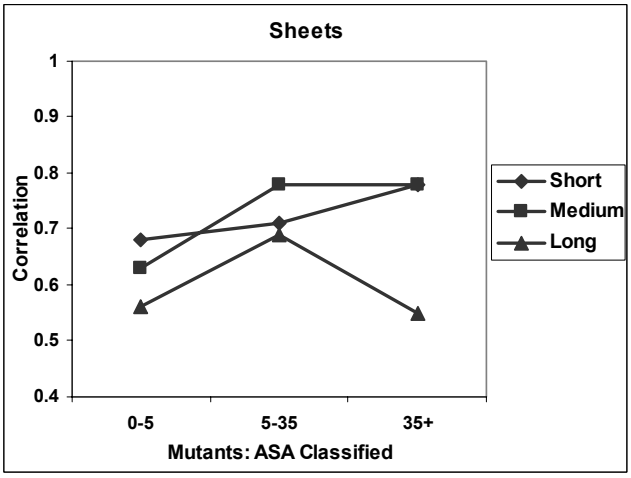
4.3.7 Short, Medium and Long Distance Ranges

The radial pair distribution function was dissected into short, medium and long range interactions. The maximum efficiency of these forces is given when they were used to develop the predict equation separately. Correlation and other statistics were then analysed. Effects of all these interactions were shown in fig. 18 for all the ASA and SS classified bins.

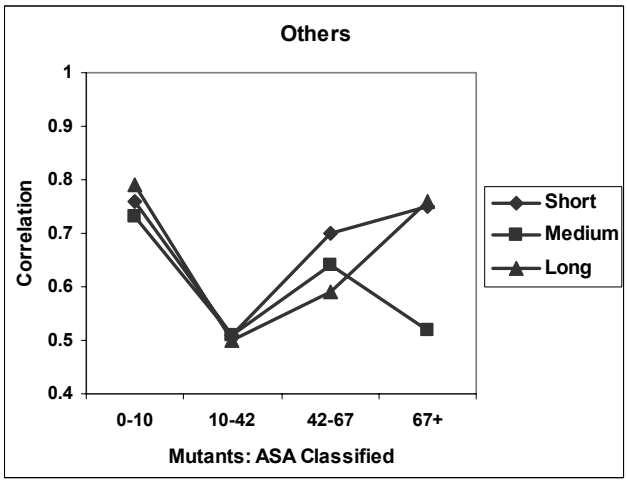
Since the 40 atom model did not provide enough population at selected distance r_d after dissecting the total radial distribution into short, medium and long ranges, the 20 atom model with amino acid C_α atoms was used. Though it showed a slightly reduced correlation of 0.81 with experimental ΔΔG of all 1518 mutations due to minimised atom definition of the protein environment, it was selected for observing short, medium and long range interactions.



(a)



(b)



(c)

Fig. 18: Effect of short, medium and long range interactions: (a) helices (b) sheets and (c) others indicate the effect of short, medium and long range interactions. Here, 20 atoms were used, since the stepwise selection (reduced 40 atoms) did not provide enough population in each bin. Mutations were classified into 11 bins as in CL11 model.

Long range interactions were observed to have relatively high influence on the stability of exposed mutations (Fig. 18c) in structural elements other than helices or sheets. This is due to the fact that turns and coils mostly have high solvent accessibility and exist in outer region of the protein. It was also observed that all interactions were quite low and deviates much in partially buried ASA region of helices (Fig. 18a, 18c), but beta sheet residues in the same region were influenced by numerous medium range interactions (Fig. 18b). Mutations in the buried region were predicted more efficiently than others due to the predictive power of short range atom potentials in assessing hydrophobic interactions in the region.

4.4 Prediction Model Validation

All the validation tests listed below were carried out for the $\Delta\Delta G/\Delta\Delta GH_2O$ datasets of mutations. Previous work in this area (Guerois et al. 2002; Bordner and Abagyan 2004) demonstrated successful split-sample validation in most of the cases. But, the training and test datasets were split randomly. In theory, the training and test datasets can be split in different ways to get good correlation and prediction accuracy with experimental $\Delta\Delta G/\Delta\Delta GH_2O$. In this case, different ways that were used to break the prediction models end up providing different prediction models with variable regression coefficients.

But, the validation tests in this work implement a specific way of breaking the mutation dataset. As described in methods, ASA and secondary structure information has been used for the mutations. A similar work done by others (Guerois et al. 2002; Hoppe and Schomburg 2002; Bordner and Abagyan 2004) reported that mutations of a specific secondary structure region could be applied to other structural regions. This ends up providing a model with wrong validation conditions. In the current work, the validation tests for the mutations were done within the specific structural regions (within specific secondary structures and ASA). This provides a reliable and accurate model that can be used in future for the prediction of changes in thermal and chemical stability of point mutations.

4.4.1 Split-sample validation

As described in the methods (section 3.7.7), the mutation dataset was spilt into training and test dataset. The training dataset acts as a representative dataset for the all the mutation dataset. Selection of this representative dataset uses the ASA and secondary structure specific information of mutated amino acid. For thermal $\Delta\Delta G$ values, the training and test datasets contain 822 and 696 mutations respectively. Split sample validation gave a correlation of 0.87 for the training dataset, and 0.77 for the representative test dataset (Table 20). Correctly predicted mutations for the training and test datasets were observed to be 84.6% and 81.6% with a standard error of 0.707 and 0.945 kcal/mol respectively. Sensitivity (specificity) was observed to be 0.77 (0.69) and 0.72 (0.65) for training and test datasets respectively. Thus, the model can be transferred to predict new mutations efficiently.

Prediction efficiency for the mutations from $\Delta\Delta GH_2O$ was also tested separately. The training dataset showed a correlation of 0.82 for 801 mutations with 86.39% (Table 21) of mutations correctly predicted. Regression coefficients calculated from the training dataset were then applied to the test dataset. It showed a correlation of 0.64 with 82.84% of mutations correctly predicted. Since transferring regression coefficients may change the magnitude of $\Delta\Delta G$, prediction accuracy must be critical in test dataset's validation. This supports the fact that the prediction model can be transferred to new mutations in future.

4.4.2 k-fold cross-validation

For the k-fold cross validation, several parts of original dataset can be tested for validating the reliability of transferring the potentials so that prediction model can be trusted in real time conditions. The results obtained from this validation are always highly supportive to the split-sample validation. For this cross validation, the mutation dataset was divided into 3, 4 and 5 subsets for 3-fold, 4-fold and 5-fold cross validation tests. As described in methods, one subset was used as test and the remaining subsets were used for training in any

specific cross validation. This procedure was followed separately for all k-fold cross validation tests. The ASA and secondary structure specific information was also used for these validation tests. For the 3-fold cross validation, the test dataset using thermal ΔΔG showed a correlation and prediction accuracy of 0.73 and 80.17% respectively (table 20). For the 4-fold and 5-fold cross validation tests, the correlation increased slightly (0.75 and 0.77 respectively). Prediction accuracy was observed to be 82.15% and 81.36% respectively. In all these validation tests, the prediction efficiency of the model remained comparable to the split-sample validation test. In all the above cases, more than 80% of the mutations are predicted correctly to be stabilising or destabilising (prediction accuracy). Thus, the k-fold cross validation supports the transferability of the prediction model and acts as an additional confirmation of results obtained from the split-sample validation.

Prediction efficiency with k-fold cross validation for $\Delta\Delta GH_2O$ values was also calculated. For the 3-fold and 4-fold cross validation tests, the test dataset showed a correlation of 0.68 (table 21). For the 5-fold cross validation, the correlation increased slightly (0.70). Prediction accuracy was observed to be more than 80% on all these tests (Table 21, Fig. 20).

Mutation Datasets	CC	PA (%)	Total mut.	TP	TN	FP	FN	Sens.	Spec.	Std. Err.
MLR1	0.86	84.85	1538	314	991	87	146	0.68	0.92	0.747
MLR1-OL	0.87	85.7	1518	330	971	88	129	0.72	0.92	0.712
SRM1	0.83	84.14	1538	306	988	90	154	0.67	0.92	0.787
SRM1-OL	0.86	85.31	1518	321	974	85	138	0.70	0.92	0.728
3-Fold – test	0.73	80.17	1518	298	919	140	161	0.65	0.87	1.052
4-Fold – test	0.75	82.15	1518	297	950	109	162	0.65	0.90	0.979
5-Fold – test	0.77	81.36	1518	297	938	121	162	0.65	0.89	0.95
Jack-knife	0.7	77.4	1518	344	831	115	228	0.60	0.88	1.17
Split-sample (train)	0.87	84.67	822	167	529	50	76	0.69	0.91	0.707
Split-sample (test)	0.77	81.32	696	141	425	55	75	0.65	0.89	0.945

Table 20: Results observed for the prediction of thermal stability using ΔΔG values. MLR and SRM indicate predictions using Multiple Linear Regression and Stepwise Regression Models respectively. –OL indicates the removal of 20 outliers. k-fold, jack-knife and split-sample show the results of validation tests.

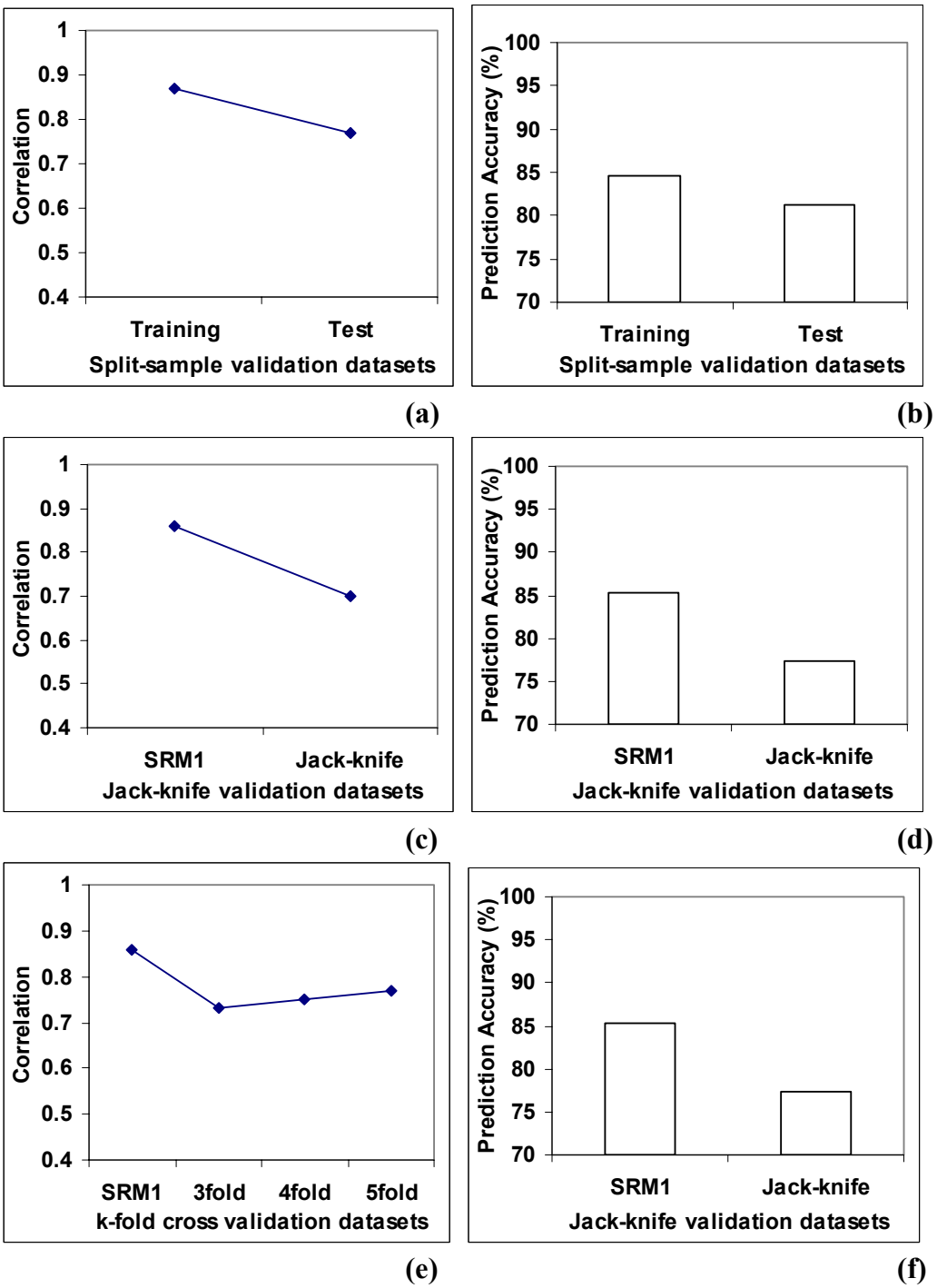


Fig. 19: Prediction model validation for thermal ΔΔG values: (a) Correlation and (b) prediction accuracy for thermal ΔΔG calculated from training and test datasets for spilt-sample validation. (c) Correlation and prediction accuracy for jack-knife validation. (e) Correlation and prediction accuracy for k-fold (3-fold, 4-fold and 5-fold) cross validation.

4.4.3 Jack-knife Test and Outliers

Jack-knife validation test is always considered as a stringent validation test. The dataset with 1518 mutations with thermal ΔΔG values was used for this validation test. As indicated in the previous validation tests, ASA and secondary structure based classifier was used initially to distinguish the mutations from the different structural regions. Later, one mutation was left out from the classified dataset, whereas the remaining mutations were used as training dataset. The same procedure was repeated for all the 1518 mutations to calculate ΔΔG values for the test dataset. Finally, the correlation and prediction accuracy was observed to be 0.7 and 77.4% respectively. No other attempt has been made to remove further outliers because the prediction efficiency observed in this validation is enough to construct the prediction model. Removing few outliers (1 or 2) may surely improve the correlation coefficient. But, the structural and stability information furnished by a specific mutation might be lost from the prediction model due to the removal of outliers. Furthermore, the standard error remained at 1.17 kcal/mol, which was very close to the values observed in the previous tests (Table 20). Prediction accuracy was also close to 80% in this validation. Thus, the prediction model was validated for each mutation specifically in the dataset. Results of Jack-knife support the accuracy of predicted ΔΔG values for the new mutations.

Jack-knife test with $\Delta\Delta G_{H_2O}$ values gave a correlation 0.66 with more than 70% of prediction accuracy (Fig. 20, Table 21). No attempt was made to remove the outliers. However, the reduction in prediction efficiency is due to the fact that mutations are not evenly located throughout all structural regions. This makes difficult for classification method to be more specific for a structural region. But, the number of mutations deposited in Protherm is increasing continually. If minimal amount of mutations for any specific structural region are available, it will then be possible to increase the prediction efficiency of the model significantly. As described already, the classification method using relative ASA is flexible to accommodate any amount of mutations in future.

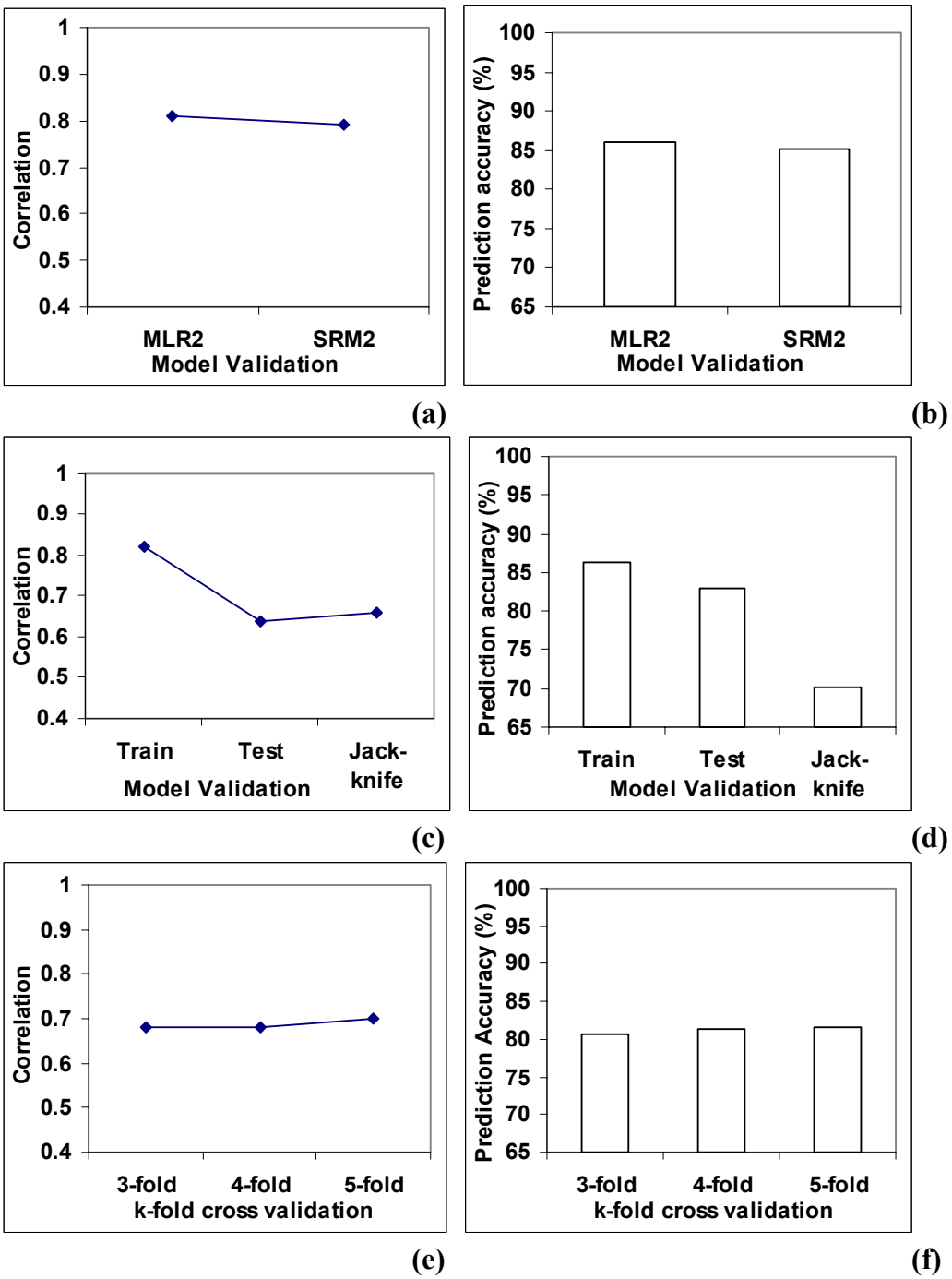


Fig. 20: Prediction model validation for $\Delta\Delta GH_2O$ values from chemical denaturation: (a) Correlation and (b) prediction accuracy calculated from multiple (MLR2) and stepwise (SRM2) regression models (c) Correlation and prediction accuracy from spilt-sample (training and test datasets) and jack-knife validation. (e) Correlation and prediction accuracy for k-fold (3-fold, 4-fold and 5-fold) cross validation.

Mutation Datasets	CC	PA (%)	Total mut.	TP	TN	FP	FN	Sens.	Spec.	Std. Err.
MLR2	0.81	86.02	1581	1176	184	140	81	0.94	0.57	0.918
SRM2	0.79	85.07	1581	1168	177	147	89	0.93	0.55	0.933
3-Fold – test	0.68	80.71	1581	1122	154	170	135	0.89	0.48	1.17
4-Fold – test	0.68	81.28	1581	1126	159	165	131	0.90	0.49	1.17
5-Fold – test	0.7	81.53	1581	1135	154	170	122	0.90	0.48	1.15
Jack-knife	0.66	70.15	1581	854	255	403	69	0.93	0.39	1.455
Split-sample (train)	0.82	86.39	609	83	76	33	0.89	0.99	0.70	0.82
Split-sample (test)	0.64	82.84	781	562	85	80	54	0.91	0.52	1.282

Table 21: Results observed for the prediction of stability using $\Delta\Delta GH_2O$ values from chemical denaturation. MLR2 and SRM2 indicate predictions using multiple linear regression and stepwise regression models respectively. k-fold (3-fold, 4-fold and 5-fold), jack-knife and split-sample show the results of validation tests.

4.5 Comparison with Other Models

Since the other prediction methods differ from using a dataset of different sizes, it's not possible to compare the results directly. In addition to that, other prediction methods use limited evaluation of their prediction to test the transferability, accuracy and reliability of their method.

Gillis and Rooman derived distance and torsion potentials using 10 proteins with mutations at the buried and solvent accessible regions of protein (Gilis and Rooman 1997; 2000). The correlation coefficient between the predicted and experimental $\Delta\Delta G$ was observed to be 0.80 and 0.67 for 121 buried and 106 surface mutations respectively. Our method differs by classifying the amino acids of structural training dataset and point mutations using the solvent accessibility and secondary structure specificity. Torsion potentials are not added separately by weighting factors, but they are considered together with other atoms separately for the multiple regression. Furthermore, they are also normalised using Gaussian function to accommodate torsion angle perturbation in mutants.

Hoppe et al. used similar statistical potentials with a training dataset of 546 mutations with a correlation of 0.75 and applied the parameters to a test dataset

of 866 mutants with a correlation of 0.62 (Hoppe and Schomburg 2002). But, the $\Delta\Delta G$ and $\Delta\Delta GH_2O$ values were mixed in the prediction model.

Khatun et al. developed contact potentials and took 3 datasets of 2317 mutations totally from 13 proteins (Khatun et al. 2004). For a big dataset of 1356 mutations, the correlation was 0.66 and 0.46 during the training and testing of the split sample validation respectively. They suggested the use of an atomistic form of potentials for future improvement of protein stability prediction. Zhou et al. used a finite ideal gas reference state for the statistical potential and reported a correlation of 0.55 for 1023 mutants in 35 proteins. But, the mutations that have decreased number of atoms were only used to avoid strains associated small-to-large mutations.

Guerois et al. developed a set of empirical energy functions with known interactions and showed a correlation 0.70 between the experimental and predicted energy values for 1088 mutants (Guerois et al. 2002).

Bordner et al. used a dataset of 1816 mutants with $\Delta\Delta GH_2O$ (Bordner and Abagyan 2004). A split sample validation with 908 selected mutants is used as training with a correlation of 0.79 and the remaining mutants were for validation with a covariance of 0.68. However, no other validation tests were carried out to test the accuracy and reliability.

Caprioti et al. developed neural network methods to discriminate the stabilising and destabilizing mutations and had an accuracy of 80% (Capriotti et al. 2004). When coupled with empirical energy values with known experimental pH and temperature conditions, the prediction can raise up to 90%. Though the experimental conditions are present for many mutations selected for their analysis, it becomes difficult to correlate and predict the experimental conditions of new mutations in site-directed mutagenesis and other similar methods.

Caprioti et al. also developed SVM (support vector machine) based models for predicting protein thermostability upon point mutations. Two separate methods were employed to construct prediction models from protein structure and

sequence. But, the mutation dataset used by their method has several redundant mutations. i.e. the mutations with same ΔΔG and experimental values (pH) were repeated in many cases. So, the prediction efficiency showed is not directly comparable. In our method, mutation dataset from Protherm includes 1236 non-redundant mutations (January 2005). But, the dataset used by Caprioti et al. boasts 2046 mutations (December 2004). Their dataset is available in the internet from their website (Capriotti et al. 2005a; Capriotti et al. 2005b).

4.6 Public World Wide Web Access

A WWW application has been developed to predict changes in protein stability upon point mutations. The algorithm will be available from CUBIC bioinformatics toolbox at the following URL. Help materials and details are provided in the web application. Prediction algorithm requires 3D structure of the protein. If the 3D structure is not known, a structure which is highly homologous can be used. In this case, the prediction may not be accurate. The WWW link for CUBIC bioinformatics toolbox is given below:

http://www.biotool.uni-koeln.de/

Application input:

- PDB ID
- Residue ID
- Wild type amino acid
- Mutant amino acid
- Chain ID (If required)

Some PDB structures either don't have chain specification or have only one chain. In that case, the program doesn't require chain ID. If there are multiple chains and the supplied input for wild type amino acid and residue ID matches only one chain, the program assumes that the input belongs to a specific chain. All others cases require chain ID to be selected. The program gives comprehensive information about the structure and stability in the next screens.

APPENDICES

Appendix A: Abbreviations

ASA	Accessible Surface Area
SS	Secondary Structure
SSE	Secondary Structure Element
Tol	Tolerance
VIF	Variance Inflation Factor
SRM	Stepwise Regression Model
MLR	Multiple Linear Regression
SLR	Simple Linear Regression
PDB	Protein Data Bank
PISCES	Protein Sequence Culling Server
PEF	Physical Effective Energy Function
SEF	Statistical Effective Energy Function
EEF	Empirical Effective Energy Function
MFP	Mean Force Potentials
AA	Amino Acid
Sens	Sensitivity
Spec	Specificity
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
DSC	Differential Scanning Calorimetry
CD	Circular Dichroism
SPACI	Summary PDB ASTRAL Check Index
AEROSPACI	Aberrant Entry Re-Ordered SPACI scores
ASPs	Atomic Solvation Parameters

Appendix B: Symbols/Units

ψ	Phi (torsion angle)
φ	Psi (torsion angle)
σ	Standard deviation (Sigma)
p	Statistical Significance
R	Pearson's Correlation Coefficient
Å	Anstroms
$\Delta\Delta G$	Difference in free energy of unfolding (between mutant and native protein) obtained with Schellman eqation in the case of thermal denaturation method.
$\Delta\Delta GH_2O$	Difference in free energy of unfolding in water, determined by denaturant denaturation of proteins and extrapolation of the data to zero concentration of denaturant [kcal/mol]

Appendix C: Amino Acid Properties

Amino acid	Code 3 letter	Code 1 letter	Mass	Surface	Surface (A-X-A)	Volume	Solubility
Alanine	Ala	A	71.09	115	110.2	88.6	16.65
Arginine	Arg	R	156.19	225	229.0	173.4	15
Aspartate	Asp	D	114.11	150	144.1	111.1	0.778
Asparagine	Asn	N	115.09	160	146.4	114.1	3.53
Cysteine	Cys	C	103.15	135	140.4	108.5	very high
Glutamate	Glu	E	129.12	190	174.7	138.4	0.864
Glutamine	Gln	Q	128.14	180	178.6	143.8	2.5
Glycine	Gly	G	57.05	75	78.7	60.1	24.99
Histidine	His	H	137.14	195	181.9	153.2	4.19
Isoleucine	Ile	I	113.16	175	185.0	166.7	4.117
Leucine	Leu	L	113.16	170	183.1	166.7	2.426
Lysine	Lys	K	128.17	200	205.7	168.6	very high
Methionine	Met	M	131.19	185	200.1	162.9	3.381
Phenylalanine	Phe	F	147.18	210	200.7	189.9	2.965
Proline	Pro	P	97.12	145	141.9	112.7	162.3
Serine	Ser	S	87.08	115	117.2	89.0	5.023
Threonine	Thr	T	101.11	140	138.7	116.1	very high
Tryptophan	Trp	W	186.12	255	240.5	227.8	1.136
Tyrosine	Tyr	Y	163.18	230	213.7	193.6	0.0453
Valine	Val	V	99.14	155	153.7	140.0	8.85

Units:

Mass [dalton], surface [$Å^2$] (Chothia 1976), volume [$Å^3$] (Zamyatnin 1972), Accessible Surface in A-X-A (Ala-X-Ala) extended state [$Å^2$] (Gromiha et al. 1999a) and solubility [g/100g, 25°C] (Merck & Co. Inc. 1989).

TABLE INDEX

three models use VIF cutoff values of 50, 30 and 20 respectively. SRM1 and SRM2 indicate the selection of atoms with stepwise regression model for the $\Delta\Delta G$ and $\Delta\Delta GH_2O$ mutation datasets respectively.

Table 15: No. of mutations (with thermal $\Delta\Delta G$) allocated according to the classification system using accessible surface area and secondary structure specificity. (a) CL12A and CL11 classify the mean force potentials and mutations into 12 and 11 structural regions respectively. (b) CL12B classifies turns separately with reduced number of ASA ranges.

Table 16: No. of mutations (with $\Delta\Delta GH_2O$ from chemical denaturation) allocated according to the classification system using accessible surface area and secondary structure specificity. CL11D classifies the mean force potentials and mutations into 11 structural regions.

Table 17: Comparison of structural training datasets: (a) Correlation coefficient (CC) and overall prediction accuracy (PA%) were compared. (b) No. of times a specific structural training dataset is overtaken by other datasets in providing better a correlation with experimental $\Delta\Delta GH_2O$ values. Sens and Spec mean Sensitivity and Specificity; TP- True Positive; TN – True Negatives; FP – False Positives; FN – False Negatives (Appendix A).

Table 18: Comparison of atom classification models: Correlation and prediction efficiency of 5 different atom types.

Table 19: Optimisation of the classification of different structural regions.

Table 20: Results observed for the prediction of thermal stability using $\Delta\Delta G$ values. MLR and SRM indicate predictions using Multiple Linear Regression and Stepwise Regression Models respectively. –OL indicates the removal of 20 outliers. k-fold, jack-knife and split-sample show the results of validation tests.

Table 21: Results observed for the prediction of stability using $\Delta\Delta GH_2O$ values from chemical denaturation. MLR2 and SRM2 indicate predictions using multiple linear regression and stepwise regression models respectively. k-fold (3-fold, 4-fold and 5-fold), jack-knife and split-sample show the results of validation tests.

FIGURE INDEX

Fig. 9: Scatterplot explaining the experimental and predicted ΔΔG values for 1538 (a) and 1518 (b) mutations. These mutation datasets were used with and without 20 outliers respectively. Outliers were removed to improve the prediction efficiency of the multiple regression model.

Fig. 10: Prediction improvement after the inclusion of torsion potentials (TP) with atom potentials (AP): (a) Correlation and (b) prediction accuracy for 1538 mutations. CL11 [helices (0-2, 2-30, 30-60, 60+), sheets (0-5,5-35,35+) and others (0-10,10-42,42-67,67+)] indicates a new classification system into 11 structural regions in order to reduce over fitting of variables that may occur in the previous classification system.

Fig. 11: Optimisation of atoms types using various statistical regression models: (a) correlation and (b) prediction efficiency based on the analysis and reduction of atom types after multicolinearity diagnostics. 'All' indicates the usage of all atoms for the statistical model. VIF<20, VIF<30 and VIF<50 indicate the statistical models that use atoms with VIF values less than 20, 30 and 50 respectively. (c) Correlation and (d) prediction accuracy based on the reduction of atoms with stepwise regression selection methods. 'All' and 'All-OL' indicate the datasets of mutations before and after the removal of outliers using normal multiple regression. 'SRM1' indicates the stepwise regression selection model using ΔΔG after the removal of outliers for the prediction of protein mutant stability. 'SRM2' indicates stepwise regression model using $\Delta\Delta GH_2O$ for 1581 mutations.

Fig. 12: Comparison of maximum sequence identity cutoff for the structural training datasets. PI-1, PI-2, PI-3, PI-4, PI-5and PI-6 indicate datasets with 25%, 30%, 35%, 40%, 45% and 50% maximum sequence identity respectively.

Fig. 13: Comparison of 5 different atom classification models used in the prediction of changes in protein stability. Overall (a) Correlation and (b) prediction accuracy for predicting thermal ΔΔG values

Fig. 14: Structural environment specific comparison of 5 different atom classification models used in the prediction of changes in protein stability. (a) Correlation and (b) prediction accuracy for predicting thermal ΔΔG values. Prediction efficiencies in 11 different structural regions were compared using CL11 classification method (Table 15).

Fig. 15: Comparison of structural training datasets for their efficiency to render torsion angle potentials. Atom potentials are maintained constant for all these validation.

Fig. 16: Comparison of structural training datasets to render torsion angle potentials. For the atom potentials, PISCES was maintained as constant. For the torsion potentials, 3 datasets were compared PISCES, SCOP-Astral and Top500 were compared. (a) Correlation and (b) prediction accuracy for predicting mutations with thermal ΔΔG values.

Fig. 17: Correlation and prediction accuracy of 1518 mutations with thermal $\Delta\Delta G$ classified into 11structural regions. (a)(b) helices, (c)(d) sheets and (e)(f) others.

Fig. 18: Effect of short, medium and long range interactions: (a) helices (b) sheets and (c) others indicate the effect of short, medium and long range interactions. Here, 20 atoms were used, since the stepwise selection (reduced 40 atoms) did not provide enough population in each bin. Mutations were classified into 11 bins as in CL11 model.

Fig. 19: Prediction model validation for thermal $\Delta\Delta G$ values: (a) Correlation and (b) prediction accuracy for thermal $\Delta\Delta G$ calculated from training and test datasets for spilt-sample validation. (c) Correlation and prediction accuracy for jack-knife validation. (e) Correlation and prediction accuracy for k-fold (3-fold, 4-fold and 5-fold) cross validation.

Fig. 20: Prediction model validation for $\Delta\Delta GH_2O$ values from chemical denaturation: (a) Correlation and (b) prediction accuracy calculated from multiple (MLR2) and stepwise (SRM2) regression models (c) Correlation and prediction accuracy from spilt-sample (training and test datasets) and jack-knife validation. (e) Correlation and prediction accuracy for k-fold (3-fold, 4-fold and 5-fold) cross validation.

PUBLICATIONS

Journal Article

Title: Structural Analysis and Prediction of Protein Mutant Stability using Distance and Torsion Potentials: Role of Secondary Structure and Solvent Accessibility. (submitted).

Posters

ISMB / ECCB 2004, Glasgow: Poster Title: Prediction of Factors Determining Changes in Thermostability in Protein Mutants.

GCB 2004, Bielefeld: Poster Title: Prediction of Factors Determining Changes in Thermostability in Protein Mutants (with improved results).

ISMB 2005, Detroit: Two posters were presented. Poster Title 1: Computational Analysis of RNA Binding Proteins Based On Composition, Sequence And Structural Information. Poster Title 2: http://www.iscbsc.org - A New Professional Web-based ISCB Student Council Framework for Computational Biology Support

ECCB 2005, Madrid: Optimisation Of Atomic Interaction Models For An Effective Description Of Protein Structure Parameters.

ACKNOWLEDGEMENT

I am extremely thankful to my Ph.D. supervisor, Prof. Dr. Dietmar Schomburg for selecting me through International Max Planck Research School and providing me an opportunity to work in structural bioinformatics in Cologne University BioInformatics Center (CUBIC).

I would like to extend my sincere thanks to International Max Planck Research School (IMPRS) for providing me fellowship for 3 years to conduct my research. Besides, the course work organised by IMPRS and the partner institutions are added advantage for the doctoral research.

It is my pleasure to thank my second supervisor Prof. Dr. Heinz Saedler to have timely discussions during my Ph.D. that helped me eventually to solve key issues in this work.

Getting a fellowship in IMPRS was only possible, since Dr. Guntram Bauer, former coordinator of IMPRS carefully screened the applications and invited me for a personal interview in Cologne. I extend my sincere thanks to him for his critical efforts for the development of IMPRS students. Dr. Ralf Petri, the present coordinator of IMPRS and an additional advisor for my research project offered me an exceptional assistance thourghout the period of my project. It's my duty to thank him for his help in various aspects.

I express my gratitude to Dr. M. Michael Gromiha, Senior Scientist in CBRC (AIST), Japan for offering me guidance to solve important problems in my project. He also helped to review the manuscripts for publication.

I solicit my sincere thanks to Dr. Christian Hoppe in my institute. As an expert and early research in this field his guidance rendered me strong support during the initial periods of the project.

I wish to thank Mr. Madenhalli Abhinandan Suresh, one of the CUBIC students who helped me to develop a part of my project to optimise the various parts of the prediction model.

I would also like to thank Dr. Silke Schrader, Mr. Markus Leber, Dr. Mark Lohman (Coordinator, CUBIC) and Mrs. Beate Marx, in offering me assitance to discuss and solve personal and scienfic problems.

I owe my special thanks to Prof. Dr. Kevin Karplus for having an extensive discussion in my research project in ISMB conference (Detroit).

I deeply express my gratefulness to my IMPRS colleagues and my lab colleagues, Mr. Oliver Martin, Mr. Philip Heuser, Mr. Pascal Benkert and Dr. Daniel Wetzler for their cooperation and support.

REFERENCES

Adachi, J., and Hasegawa, M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J Mol Evol* **42:** 459-468.

Ahern, T.J., and Klibanov, A.M. 1988. Analysis of processes causing thermal inactivation of enzymes. *Methods Biochem Anal* **33:** 91-127.

Alber, T., Sun, D.P., Wilson, K., Wozniak, J.A., Cook, S.P., and Matthews, B.W. 1987. Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. *Nature* **330:** 41-46.

Baud, F., and Karlin, S. 1999. Measures of residue density in protein structures. *Proc Natl Acad Sci U S A* **96:** 12494-12499.

Bava, K.A., Gromiha, M.M., Uedaira, H., Kitajima, K., and Sarai, A. 2004. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res* **32:** D120-121.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res* **28:** 235-242.

Betancourt, M.R., and Thirumalai, D. 1999. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci* **8:** 361-369.

Bordner, A.J., and Abagyan, R.A. 2004. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins* **57:** 400-413.

Bromberg, S., and Dill, K.A. 1994. Side-chain entropy and packing in proteins. *Protein Sci* **3:** 997-1009.

Capriotti, E., Fariselli, P., Calabrese, R., and Casadio, R. 2005a. Predicting protein stability changes from sequences using support vector machines. *Bioinformatics* **21 Suppl 2:** ii54-ii58.

Capriotti, E., Fariselli, P., and Casadio, R. 2004. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* **20 Suppl 1:** I63-I68.

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

Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S.E. 2004. The ASTRAL Compendium in 2004. *Nucleic Acids Res* **32:** D189-192.

Chi, E.Y., Krishnan, S., Randolph, T.W., and Carpenter, J.F. 2003. Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharm Res* **20:** 1325-1336.

Chothia, C. 1974. Hydrophobic bonding and accessible surface area in proteins. *Nature* **248:** 338-339.

Chothia, C. 1976. The nature of the accessible and buried surfaces in proteins. *J Mol Biol* **105:** 1-12.

Cline, M.S., Karplus, K., Lathrop, R.H., Smith, T.F., Rogers, R.G., Jr., and Haussler, D. 2002. Information-theoretic dissection of pairwise contact potentials. *Proteins* **49:** 7-14.

Colovos, C., and Yeates, T.O. 1993. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* **2:** 1511-1519.

Covalt, J.C., Jr., Roy, M., and Jennings, P.A. 2001. Core and surface mutations affect folding kinetics, stability and cooperativity in IL-1 beta: does alteration in buried water play a role? *J Mol Biol* **307:** 657-669.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. 1978. A model of evolutionary change in proteins. In Atlas of Protein Sequence and Structure *National Biomedical Research Foundation* **5:** 345-352.

Delarue, M., and Koehl, P. 1995. Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *J Mol Biol* **249:** 675-690.

Dengler, U. 1998. Kristallstruktur der D-2-Hydroxyisocaproat-Dehydrogenase aus Lactobacillus casei. *Ph.D. thesis*.

Dengler, U., Niefind, K., Kiess, M., and Schomburg, D. 1997. Crystal structure of a ternary complex of D-2-hydroxyisocaproate dehydrogenase from Lactobacillus casei, NAD+ and 2-oxoisocaproate at 1.9 A resolution. *J Mol Biol* **267:** 640-660.

Dönitz, J. 2001. Statistische energiefunktionen als bewertungskriterium für die qualität der vorhersage von loop bereichen in proteinen. *Diploma thesis***:** 64 pages.

Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci U S A* **81:** 140-144.

Engelman, D.M., Steitz, T.A., and Goldman, A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem* **15:** 321-353.

Engh, R.A., and Huber, R. 1991. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica Section A* **47:** 392-400.

Finkelstein, A.V., Badretdinov, A., and Gutin, A.M. 1995a. Why do protein architectures have Boltzmann-like statistics? *Proteins* **23:** 142-150.

Finkelstein, A.V., Gutin, A.M., and Badretdinov, A. 1995b. Boltzmann-like statistics of protein architectures. Origins and consequences. *Subcell Biochem* **24:** 1-26.

Fleming, P.J., and Richards, F.M. 2000. Protein packing: dependence on protein size, secondary structure and amino acid composition. *J Mol Biol* **299:** 487-498.

Gallivan, J.P., and Dougherty, D.A. 1999. Cation-pi interactions in structural biology. *Proc Natl Acad Sci U S A* **96:** 9459-9464.

Gilis, D., and Rooman, M. 1997. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.* **272:** 276-290.

Gilis, D., and Rooman, M. 2000. PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng* **13:** 849-856.

Gohlke, H., Hendlich, M., and Klebe, G. 2000. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* **295:** 337-356.

Gonnet, G.H., Cohen, M.A., and Benner, S.A. 1992. Exhaustive matching of the entire protein sequence database. *Science* **256:** 1443-1445.

Grantcharova, V.P., Riddle, D.S., and Baker, D. 2000. Long-range order in the src SH3 folding transition state. *Proc Natl Acad Sci U S A* **97:** 7084-7089.

Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedaira, H., and Sarai, A. 1999a. ProTherm: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Res* **27:** 286-288.

Gromiha, M.M., Oobatake, M., Kono, H., Uedaira, H., and Sarai, A. 1999b. Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. *Protein Eng* **12:** 549-555.

Guerois, R., Nielsen, J.E., and 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.

Hendsch, Z.S., and Tidor, B. 1994. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci* **3:** 211-226.

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89:** 10915-10919.

Honig, B.H., and Hubbell, W.L. 1984. Stability of "salt bridges" in membrane proteins. *Proc Natl Acad Sci U S A* **81:** 5412-5416.

Hopp, T.P., and Woods, K.R. 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci U S A* **78:** 3824-3828.

Hoppe, C., and Schomburg, D. 2002. Entwicklung einer richtungs– und abstandsabhängigen wissensbasierten Bewertungsfunktion für die Vorhersage der Thermostabilität von Proteinen. *Ph.D. dissertation***:** 180 pages.

Inoue, H., and Timasheff, S.N. 1968. The interaction of beta-lactoglobulin with solvent components in mixed water-organic solvent systems. *J Am Chem Soc* **90:** 1890-1898.

Jacobs, D.J., Dallakyan, S., Wood, G.G., and Heckathorne, A. 2003. Network rigidity at finite temperature: relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *Phys Rev E Stat Nonlin Soft Matter Phys* **68:** 061109.

Jacobs, D.J., Rader, A.J., Kuhn, L.A., and Thorpe, M.F. 2001. Protein flexibility predictions using graph theory. *Proteins* **44:** 150-165.

Janin, J. 1979. Surface and inside volumes in globular proteins. *Nature* **277:** 491-492.

Jiang, F., and Kim, S.H. 1991. "Soft docking": matching of molecular surface cubes. *J Mol Biol* **219:** 79-102.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8:** 275-282.

Karlin, S., Zhu, Z.Y., and Baud, F. 1999. Atom density in protein structures. *Proc Natl Acad Sci U S A* **96:** 12500-12505.

Khatun, J., Khare, S.D., and Dokholyan, N.V. 2004. Can contact potentials reliably predict stability of proteins? *J Mol Biol* **336:** 1223-1238.

Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., et al. 2000. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* **33:** 889-897.

Kumar, S., and Nussinov, R. 1999. Salt bridge stability in monomeric proteins. *J Mol Biol* **293:** 1241-1255.

Kyte, J., and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157:** 105-132.

Lacroix, E., Viguera, A.R., and Serrano, L. 1998. Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J Mol Biol* **284:** 173-191.

Lazaridis, T., and Karplus, M. 2000. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol* **10:** 139-145.

Lee, B., and Richards, F.M. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55:** 379-400.

Levitt, M., Gerstein, M., Huang, E., Subbiah, S., and Tsai, J. 1997. Protein folding: the endgame. *Annu Rev Biochem* **66:** 549-579.

Li, A.J., and Nussinov, R. 1998. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins* **32:** 111-127.

Livesay, D.R., Dallakyan, S., Wood, G.G., and Jacobs, D.J. 2004. A flexible approach for understanding protein stability. *FEBS Lett* **576:** 468-476.

Lomize, A.L., Reibarkh, M.Y., and Pogozheva, I.D. 2002. Interatomic potentials and solvation parameters from protein engineering data for buried residues. *Protein Sci* **11:** 1984-2000.

Lovell, S.C., Davis, I.W., Arendall, W.B., 3rd, de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S., and Richardson, D.C. 2003. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* **50:** 437-450.

Melo, F., and Feytmans, E. 1997. Novel knowledge-based mean force potential at atomic level. *J Mol Biol* **267:** 207-222.

Melo, F., and Feytmans, E. 1998. Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol* **277:** 1141-1152.

Melo, F., Sanchez, R., and Sali, A. 2002. Statistical potentials for fold assessment. *Protein Sci* **11:** 430-448.

Merck & Co. Inc., N., N.J. 1989. *The Merck Index*. Merck & Co. Inc., New Jercy, pp. 201.

Micheletti, C., Carloni, P., and Maritan, A. 2004. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. *Proteins* **55:** 635-645.

Micheletti, C., Cecconi, F., Flammini, A., and Maritan, A. 2002. Crucial stages of protein folding through a solvable model: predicting target sites for enzyme-inhibiting drugs. *Protein Sci* **11:** 1878-1887.

Mintseris, J., and Weng, Z. 2004. Optimizing protein representations with information theory. *Genome Inform Ser Workshop Genome Inform* **15:** 160-169.

Mitchell, J.B.O., Alex, A., and Snarey, M. 1999. SATIS: Atom Typing from Chemical Connectivity. *J. Chem. Inf. Model.* **39:** 751-757.

Munoz, V., and Serrano, L. 1994. Elucidating the folding problem of helical peptides using empirical parameters. *Nat Struct Biol* **1:** 399-409.

Munoz, V., and Serrano, L. 1995. Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. *J Mol Biol* **245:** 275-296.

Munoz, V., and Serrano, L. 1997. Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers* **41:** 495-509.

Muyoung, H., Suhkmann, K., Eun-Joung, M., Mookyung, C., Kwanghoon, C., and Iksoo, C. 2005. Perceptron learning of pairwise contact energies for proteins incorporating the amino acid environment. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* **72:** 011906.

Nei, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.

Niefind, K., and Schomburg, D. 1991. Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles. *J Mol Biol* **219:** 481-497.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D.G., and Notredame, C. 2004. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol* **340:** 385-395.

Paci, E., Lindorff-Larsen, K., Dobson, C.M., Karplus, M., and Vendruscolo, M. 2005. Transition state contact orders correlate with protein folding rates. *J Mol Biol* **352:** 495-500.

Pandey, B.P., Zhang, C., Yuan, X., Zi, J., and Zhou, Y. 2005. Protein flexibility prediction by an all-atom mean-field statistical theory. *Protein Sci* **14:** 1772-1777.

Rader, A.J., Hespenheide, B.M., Kuhn, L.A., and Thorpe, M.F. 2002. Protein unfolding: rigidity lost. *Proc Natl Acad Sci U S A* **99:** 3540-3545.

Reva, B.A., Finkelstein, A.V., Sanner, M.F., and Olson, A.J. 1997. Residue-residue mean-force potentials for protein structure recognition. *Protein Eng* **10:** 865-876.

Robinson, A.B., and Rudd, C.J. 1974. Deamidation of glutaminyl and asparaginyl residues in peptides and proteins. *Curr Top Cell Regul* **8:** 247-295.

Ruvinsky, A.M., and Kozintsev, A.V. 2005. The key role of atom types, reference states, and interaction cutoff radii in the knowledge-based method: new variational approach. *Proteins* **58:** 845-851.

Serrano, L., Kellis, J.T., Jr., Cann, P., Matouschek, A., and Fersht, A.R. 1992. The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J Mol Biol* **224:** 783-804.

Shih, P., Holland, D.R., and Kirsch, J.F. 1995. Thermal stability determinants of chicken egg-white lysozyme core mutants: hydrophobicity, packing volume, and conserved buried water molecules. *Protein Sci* **4:** 2050-2062.

Shirley, B.A. 1995. *Protein Stability and Folding: Theory and Practice*. Humana Press, pp. 387.

Shoichet, B.K., Baase, W.A., Kuroki, R., and Matthews, B.W. 1995. A relationship between protein stability and protein function. *Proc Natl Acad Sci U S A* **92:** 452-456.

Shortle, D. 1996. The denatured state (the other half of the folding equation) and its role in protein stability. *Faseb J* **10:** 27-34.

Sippl, M.J. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213:** 859-883.

Sippl, M.J. 1993. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des* **7:** 473-501.

Stark, G.R. 1965. Reactions of cyanate with functional groups of proteins. 3. Reactions with amino and carboxyl groups. *Biochemistry* **4:** 1030-1036.

Swaim, M.W., and Pizzo, S.V. 1988. Methionine sulfoxide and the oxidative regulation of plasma proteinase inhibitors. *J Leukoc Biol* **43:** 365-379.

Takano, K., Funahashi, J., Yamagata, Y., Fujii, S., and Yutani, K. 1997. Contribution of water molecules in the interior of a protein to the conformational stability. *J Mol Biol* **274:** 132-142.

Takano, K., Ota, M., Ogasahara, K., Yamagata, Y., Nishikawa, K., and Yutani, K. 1999. Experimental verification of the 'stability profile of mutant protein' (SPMP) data using mutant human lysozymes. *Protein Eng* **12:** 663-672.

Topham, C.M., Srinivasan, N., and Blundell, T.L. 1997. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng* **10:** 7-21.

Tsai, C.J., and Nussinov, R. 1997. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci* **6:** 24-42.

Tyler-Cross, R., and Schirch, V. 1991. Effects of amino acid sequence, buffers, and ionic strength on the rate and mechanism of deamidation of asparagine residues in small peptides. *J Biol Chem* **266:** 22549-22556.

Wang, G., and Dunbrack, R.L., Jr. 2003. PISCES: a protein sequence culling server. *Bioinformatics* **19:** 1589-1591.

Whitaker, J.R., and Feeney, R.E. 1983. Chemical and physical modification of proteins by the hydroxide ion. *Crit Rev Food Sci Nutr* **19:** 173-212.

Wyman, J. 1964. Linked Functions and Reciprocal Effects in Hemoglobin - a 2Nd Look. *Advances in Protein Chemistry* **19:** 223-286.

Xu, J., Baase, W.A., Baldwin, E., and Matthews, B.W. 1998. The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci* **7:** 158-177.

Yutani, K., Ogasahara, K., Tsujita, T., and Sugino, Y. 1987. Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit. *Proc Natl Acad Sci U S A* **84:** 4441-4444.

Zamyatnin, A.A. 1972. Protein volume in solution. *Prog Biophys Mol Biol* **24:** 107-123.

Zhou, H., and Zhou, Y. 2002a. 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.

Zhou, H., and Zhou, Y. 2002b. Stability scale and atomic solvation parameters extracted from 1023 mutation experiments. *Proteins* **49:** 483-492.

Zhu, Z.Y., and Karlin, S. 1996. Clusters of charged residues in protein three-dimensional structures. *Proc Natl Acad Sci U S A* **93:** 8350-8355.

ZUSAMMENFASSUNG

Die Analyse der für die Stabilität von Proteinen zuständigen Faktoren stellt einen der Schlüsselbereiche der molekularbiologischen Forschung dar und bringt direkte Implikationen für die Strukturvorhersage von Proteinen und Protein-Protein Docking mit sich. Es wurde die Stabilität von Proteinen nach Punktmutationen mittels eines abstandsabhängigen paarweisen Potentials bestehend aus räumlichen Interaktionstermen sowie eines Torsionswinkelpotentials basierend auf Nachbareffekten als Grundlage der statistischen Mechanik analysiert. Der synergetische Effekt von lösungsmittelzugänglicher Oberfläche und präferierter Sekundärstruktur wurde verwendet, um die entstandenen Potentiale zu klassifizieren. Zusätzlich wurden kurz-, mittel- und langreichweitige Interaktionen der Proteinumgebung analysiert.

Verschiedene der Beschreibung einer Proteinstruktur zugrundeliegenden Prinzipien müssen ebenfalls sorgsam untersucht werden um die Relationen zwischen Sequenz, Struktur und Funktion zu verstehen. Wissensbasierte Potentiale basierend auf atomaren Interaktionen und den Haupttorsionswinkeln wurden von verschiedenen Forschergruppen bereits für die Erforschung von Proteinstrukturen, -stabilität und -interaktionen herangezogen. In neueren Arbeiten werden diese Potentiale ebenfalls zur Vorhersage von Proteinfunktionen und Enzymkatalyse verwendet. Es wurden 5 verschiedene Atommodelle selektiert mit jeweils verschiedenen Abstandsmaßen zur Berechnung der Interaktionen und diese hinsichtlich der Eignung zur Beschreibung der Proteinumgebung verglichen. Weiterhin wurden die Torsionspotentiale zusammen mit den Atom-Atom-Potentialen so optimiert, dass gerichtete Information über die einzelnen Aminosäuren in das Modell mit eingefügt werden können.

Folgende 5 atomare Klassifikationssysteme wurden benutzt: Ein einfaches 5-Atom (basic-5) Modell (C aliphatisch, C aromatisch, H, O, N), Aminosäure C_α-Atome (C_α20), Li-Nussinov Atommodell (LN24), SATIS Modell (SA28) und das Atommodell von Melo und Feytmanns (MF40). Kohlenstoffatome von aromatischer oder aliphatischer Natur zeigenen ein signifikant verschiedenes chemisches und funktionales Verhalten und wurden separat behandelt im basic5 Atommodell zusammen mit N, O und S. Li und Nussinov definierten 24 Atomtypen der Amonisäuren anhand von Polarität und Hydrophobizität der Atome, obwohl einige dieser Atome sicherlich partiell polaren und/oder hydrophoben Charakter aufweisen. SATIS (Simple Atom Type Information System) ist ein Protokoll für die Definition und automatisierte Zuweisung von Atomtypen und deren Klassifizierung anhand von Konnektivitäten. Ursprünglich wurden die Werte der freien Energie ($\Delta\Delta G$) des Entfaltens von Punktmutionen als experimentelles Mass für eine energetische Beschreibung des Proteins verwendet. Dies kann jetzt durch eine Evaluierung gegen artifizielle konformationen erweitert werden. Bereits zuvor wurden die gemessenen Unterschiede der freien Energie zwischen Wildtyp und Mutanten benutzt, um die Vorhersagegenauigkeit von wissensbasierten Potentialen zu bewerten. Ein Datensatz von 4024 nicht-redundanten Strukturen wurde hier benutzt, um die atomaren Interaktionsabstände sowie die Torsionswinkel ϕ und ψ abzuleiten. Zuvor wurde das Programm DSSP auf alle Strukturen des Datensatzes angewendet von die Zuordnung zu den Sekundärstrukturelementen zu ermöglichen. Bevor mit der Entwicklung des Torsionspotentials begonnen werden konnte, wurden gleichförmige sogenannte 'Bins' angelegt und diese mit einem konstanten Wert initialisiert, um Nullwerte bei der

Berechnung von Boltzmann-Energien zu vermeiden. Danach wurden diese Bins normalisiert über eine zirkuläre Gaussfunktion für ϕ und eine bivariate Normalverteilung für ψ. Da es zu Perturbationen der Torsionswinkel innerhalb bestimmter Aminosäuren in den Mutanten kommen kann, dient die Gauss-Funktion als Glättungsfunktion um so die Effizienz bei der Vorhersage solcher vom Ideal leicht abweicheden Konformationen zu erhöhen.

Die Ergebnisse wurden anhand der Korrelation zwischen den beobachteten experimentellen ΔΔG und den vorhergesagten ΔΔG validiert. Die Vorhersagegenauigkeit für korrekte Vorhersagen der Form 'stabilisierend' oder 'destabilisierend' wurde ebenfalls herangezogen. Die Resultate zeigen, dass das Atommodell von Melo und Feytmanns die strukturellen Parameter am besten wiedergibt, da ein Korrelationskoeffizient von 0,85 erreicht wird, wobei 85,31% von 1536 Mutanten korrekt vorhergesagt werden bezüglich ihrer stabilisierenden oder destabilisierenden Wirkung. SA28, LN24, C_α20 und basic5 Atommodelle ziegen Korrelationen von 0,82, 0,78, 0,76, und 0,55. In einem späteren Schritt wurden statistische schrittweise Regressionsmethoden benutzt, um die Anzahl der innerhalb eines Modells verwerdeten Atomtypen zu optimieren. Der Effekt der Torsionspotentiale mit und ohne die Apodisation über Gausssche Funktionen wurde verglichen. Dies zeigt, dass die Aminosäuren insbesondere in Beta-Faltblattstrukturen diesen gestörten Torsionswinkelbedingungen unterliegen im Vergleich zu anderen Sekundärstrukturelementen.

Für das finale Vorhersagemodell wurden zwei Datensätze von Punktmutationen zum Vergleich von theoretisch vorhergesagten stabilisierenden Energiewerten mit experimentell bestimmten ΔΔG und ΔΔGH2O aus thermalen und chemischen Denaturationsexperimenten herangezogen. Diese beinhalten 1538 und 1581 Mutationen und stammen von 101 Proteinen, die eine grosse Spanne von Sequenzidentitäten untereinander einnehmen. Die resultierenden Kraftfelder wurden genauestens evaluiert mittels einer grossen Anzahl von statistischen Tests. Die Resultate ergeben eine maximale Korrelation von 0,87 zwischen vorhergesagten und gemessenen ΔΔG Werten und eine Vorhersagegenauigkeit von 85,3% bezüglich der Klassifizierung einer Mutation als stabilisierend oder destabilisierend für den gesamten Datensatz. Ein Korrelationswert von 0,77 wurde sowohl für die Testdatensätze einer split-sample Validierung als eine k-fachen Crossvalidierung erreicht, während ein Jack-Knife Test eine Korrelation von 0,70 ergab. Obige Prozedur wurde ebenfalls für den Vergelich der theoretisch vorhergesagten Werte mit den experimentell bestimmten Werten für ΔΔGH2O durchgeführt. Es ergaben sich Korrelationswerte von 0.79 sowie eine Vorhersagegenauigkeit von 85.03%. Dieses Modell kann für die zukünftige Vorhersage von struktureller Stabilität in Proteinen in Ergänzung zu experimentellen Methoden verwendet werden. Ein neues Web-Tool befindet sich in der Entwicklung welches Teile des beschriebenen Algorithmus enthält. Dieses Werkzeug wird nach Publikation als Teil der CUBIC bioinformatics Toolbox zugänglich sein (unter www.biotool.uni-koeln.de).

SUMMARY

Analysing the factors behind protein stability is a key research topic in molecular biology and has direct implications on protein structure prediction and protein-protein docking solutions. Protein stability upon point mutations were analysed using a distance dependant pair potential representing mainly through-space interactions and torsion angle potential representing neighbouring effects as a basic statistical mechanical setup for the analysis. The synergetic effect of accessible surface area and secondary structure preferences was used as a classifier for the potentials. In addition, short, medium and long range interactions of the protein environment were also analysed.

Various principles underlying the protein structure description must also be studied carefully to efficiently understand the relationships between sequence, structure and function. Mean force potentials from atom interactions and main torsion angles were used by different investigators to evaluate the protein structure, stability and protein-protein interactions. In recent experiments, these were also used in the prediction of protein function and enzyme catalysis. Five different atom classification models with interactions in different distance ranges were selected to check their ability to effectively describe the protein environment. Furthermore, torsion angle potentials were also derived in addition to atom potentials so that orientational information of amino acids can be included to the model.

The five atom classification models that are used for atom potentials include the following: a basic five (basic5) atom model (C aliphatic, C aromatic, H, O, N), amino acid C_α atoms (C_α20), Li-Nussinov atom model (LN24), SATIS model (SA28) and Melo and Feytmans atom model (MF40). Carbon atoms with aromatic and aliphatic nature exhibit significantly different chemical and functional behaviour and they were considered separately in the basic5 atom model with N, O and S. Li and Nussinov defined 24 different amino acid atom types using the polarity and hydrophobicity of atoms, though some of the atoms may substantially have partial polar or apolar nature. SATIS (Simple Atom Type Information System) is a protocol for the definition and automatic assignment of atom types and the classification of atoms according to their covalent connectivity. The free energy values ($\Delta\Delta G$ and $\Delta\Delta GH_2O$ values from thermal and chemical denaturation) of unfolding from point mutation experiments were used as an experimental measure of protein stability. In future, this method can also be extended to evaluate other structure descriptors. It has already been reported that the measured free energy changes between wild type and mutant proteins can be predicted using statistical potentials. But, these models lack good prediction efficiency and reliability to predict protein mutant stability in future.

A dataset of 4024 non-redundant structures was used to derive the atom interactions and torsion angles ϕ and ψ, after running DSSP for the whole dataset. For torsion potentials, the bins were normalised with a standard procedure using the circular Gaussian function for ϕ and ψ having the bivariate normal distribution. Since the mutants may exhibit torsion angle perturbation in the selected amino acid position, the Gaussian function will increase the efficiency of predicting slightly altered amino acid conformations.

Results were validated based on the correlation observed between the experimental $\Delta\Delta G$ and predicted $\Delta\Delta G$ values. Prediction accuracy of being correctly predicted as stabilising or destabilising was also observed. Results show that the Melo and Feytmens atom model

predicts the protein stability to a maximum extent, since it showed a correlation coefficient of 0.85 with 85.31% of 1536 mutations correctly predicted to be either stabilising or destabilising. SA28, LN24, C_α20 and basic5 atom models showed a correlation coefficient of 0.82, 0.78, 0.76 and 0.55 respectively. Later, statistical stepwise regression methods were used to optimise the number of atoms used for the model. Effect of torsion angle potentials with and without the Gaussian apodisation was compared. This shows that the amino acids adapt perturbed torsion angle conformations in partially buried beta sheets than the other structural elements.

For the final prediction model, two datasets of point mutations were taken for the comparison of theoretically predicted stabilising energy values with experimental $\Delta\Delta G$ and $\Delta\Delta GH_2O$ from thermal and chemical denaturation experiments respectively. These include 1538 and 1581 mutations respectively and contain 101 proteins that share wide range of sequence identity. Results were carefully evaluated with a wide range of statistical tests. Results show a maximum correlation of 0.87 between predicted and experimental $\Delta\Delta G$ values and a prediction accuracy of 85.3% (stabilising or destabilising) for all mutations together. A correlation of 0.77 each for the test dataset of split-sample validation and k-fold cross validation tests was obtained and a correlation of 0.70 was shown by the jack-knife test. A similar model was implemented and the results were analysed for mutations with $\Delta\Delta GH_2O$. A correlation of 0.79 was observed with a prediction efficiency of 85.03%. This model can be used for the future prediction of protein structural stability together with various experimental techniques. A new web tool will be developed for this algorithm. This will be available as a part of the CUBIC bioinformatics toolbox (www.biotool.uni-koeln.de/).

ERKLÄRUNG

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. D. Schomburg betreut worden.

(Parthiban Vijayarangakannan)

LEBENSLAUF

Vijayarangakannan, Parthiban

Institut für Biochemie (AG Schomburg),
Zülpicher Str. 47,
50674 Köln
NRW, Germany.
Tel.: +49-221-4707427
Fax: +49-221-4707786
E-mail: parthi@uni-koeln.de

Persönliche Daten

Geburtsdatum	:	10. Mai 1978
Geburtsort	:	Dindigul, Indien
Sprachen	:	Tamil, Englisch
Familienstand	:	ledig
Reisepass Nr.	:	B0631480.
Staatsangehörigkeit	:	Indien

Ausbildung

Abschlußzeugnis	Schule/Universität	Ort	Studiendauer
M.Sc. (Life Sciences)	Bharathidasan University	Trichirappalli, Tamil Nadu, India.	1995-2000
Higher Secondary Course Certificate	St. Mary's Higher Secondary School	Dindigul, Tamil Nadu, India	1993-1995
Secondary School Leaving Certificate	St. Mary's Higher Secondary School	Dindigul, Tamil Nadu, India	1983-1993

Weitere Kurse

Ausbildung Zertifikat	Institution	Ort	Studiendauer
Office Automation (Lotus Smartsuite)	Zonix Computer Education	Trichirappalli, Tamil Nadu, India.	Mär 1998-Mär 1999
Java 2 with Servlets	Nilaa Institute of Excellence	Chennai, Tamil Nadu, India.	Mai 2000-Jun 2000
Summer Training	Indian Institute of Chemical Biology	Kolkata, West Bengal, India	Mai 1999-Jun1999

(Vijayarangakannan Parthiban)