Self Complementarity: its Applications in Probing Protein Internal Architecture, Fold Recognition and Structure Validation

Thesis submitted for the Degree of
Doctor of Philosophy (Science)
in
Biochemistry

by

Sankar Chandra Basu

Department of Biochemistry
University of Calcutta

2013

Self Complementarity: its Applications in Probing Protein Internal Architecture, Fold Recognition and Structure Validation

Thesis submitted for the Degree of
Doctor of Philosophy (Science)
in
Biochemistry

by

Sankar Chandra Basu

Department of Biochemistry
University of Calcutta

2013

"*Come, Lord, Thou Great Teacher, who has taught us that the soldier is only to obey and speak not.*"

- *Swami Vivekananda*

Dedicated to my guide

Acknowledgments

First of all, I would like to express my respect, love and heartfelt gratitude to Prof. Rahul Banerjee for giving me the opportunity to work under his supervision in this exciting area of science. He has been a friend, philosopher and guide all throughout this period of five years for me. His constant care, constructive criticism and cautious guidance accompanied me all the way through this expedition. I must acknowledge that he has exceptional intuitive prowess and insightful novel ideas which showed me how to think creatively and originally in the premise of a given scientific problem. The original programs for surface complementarity calculations were actually written by him and already available in the laboratory at the time of my joining. I have used them extensively and modified for different applications as per requirement. I shall always remain indebted to him. People rarely have the good fortune to come across such a person not only as a visionary but more importantly as a human being.

I express my sincere gratitude to Prof. Dhananjay Bhattacharyya (Computational Science Division, Saha Institute of Nuclear Physics, Kolkata) who has been a principal collaborator in the research project and has supported me constantly from the very beginning till the very end. He has generously given computer space and availability to software packages whenever they were required and has taught me the fundamentals of both the theoretical and applied aspects of Molecular Dynamics Simulations.

I am also grateful to Prof. Dipak Dasgupta (Biophysics Division, Saha Institute of Nuclear Physics, Kolkata) who had kindly granted me the primary workstation that I used in the course of the project and provided important funding supports when required. I convey my gratitude also to Prof. Pradeep Kumar Mohanty (Theoretical Condensed Matter Physics Division, Saha Institute of Nuclear Physics, Kolkata) for many fruitful discussions, especially related to network analysis and graph theory at different stages of the thesis.

I would like to acknowledge all senior professors as well as non-academic members of the Crystallography and Molecular Biology division for their support. In particular, I would like to mention Prof. Nitai Bhattacharyya (Head of the Division) for being so kind to me regarding many academic and administrative issues time and again.

I feel blessed to have a friend like Abhirup Bandyapadhyay (NIT, Durgapur) who always showed serious interests in interdisciplinary science and has helped me enormously in developing the basic understanding in different branches of algebra, geometry and statistics.

I would like to acknowledge all of my Post MSc friends, Soma, Eashita, Anita, Nandini, Arunabha, Swadesh, Binita, Kasturi, Sreeja, Moushumi, Samir and Manas. I feel very fortunate to be a part of this highly meritorious batch with immense scientific aptitude and humble attitude. I had the opportunity to share some of the finest moments of my Ph.D. tenure with Eashita, Samir, Kasturi, Jayeeta di, Soumajit da who all were very helpful and cooperative in the division. Ramanuj and Seema have been two special friends for me to whom I could talk with so much ease and share different situations.

I should definitely mention Sukanya di, Sanchita di, Manas, Angana and Sangeeta di for their help and support in sharing and adjusting with computer runtime and disc-space.

I must specifically mention about Venu da, who was the only senior when I joined lab. He treated me with extreme care and affection. Alaka di was so very caring and enthusiastic and remains to be a refreshing memory. I feel fortunate to have a friend like Sourav who always stood beside me even at the toughest of situations. I would also mention specifically about Barnali who had been particularly helpful especially at the initial years. Ambarnil da also made life simpler as a fellow bioinformatician in the last year.

I acknowledge financial support from the intramural grant of Department of Atomic Energy, Government of India, (CBAUNP project, Saha Institute of Nuclear Physics).

Last but not the least, the support of my family and friends from other than SINP (debarchana, avik da, arundhuti, kunal, debalina, shyamoleena, subhadeepa and dhriti), their love and endurance to me are something which cannot be described in words.

- Sankar

Contents

Preface

The protein folding problem remains a major unsolved problem in structural biology. In one aspect it deals with the mapping of protein primary sequences to their three dimensional folds, referred to as the second genetic code. Two factors which condition the isomorphism between sequence and fold, are (1) the pattern of hydrophobicities embedded in the polypeptide chain and (2) the packing of amino acid side chains to give densely packed protein interiors. Several groups have attempted to represent the internal architecture in proteins as some kind of network. Previous studies on protein contact networks based on the proximity of interacting point-atoms have elucidated some basic properties of such networks. The current study is based on shape complementarity and overlap of interacting side-chain surfaces which essentially extends the 'Jigsaw Puzzle' model in the domain of protein contact networks. The study has been effective in characterizing and classifying the topological patterns found within the protein interiors along with the emergence of special packing motifs like closed triplets. ***Chapter 2*** *deals with these insights.*

One of the key concepts of bimolecular recognition is complementarity between interacting surfaces which is said to have a dual aspect (1) surface complementarity arising due to the steric fit of closely packed atoms in van der Waals contact and (2) electrostatic complementarity mediated by long range electric fields due to charged or partially charged atoms. Although the term 'complementarity' naturally lends itself to inter-protein association, the current study attempts to extend the concept into protein interiors. In spite of the differences in physicochemical features of interfaces and interior atoms, the concept has been found to be fruitful in bridging the gap between binding and folding. ***Chapters 3*** *to* ***5*** *deals in great detail to the above research proposal. In this regard, a novel graphical validation tool for protein structures has been developed and made avaialeble as a standalone suite of programs in the public domain (*http://www.saha.ac.in/biop/www/sarama.html*).*

Designing novel folds and hydrophobic cores is another well-posed research problem in structural biology. There are examples of successful full sequence as well as core design for small proteins like ubiquitin. The current study proposes a computational method to redesign the hydrophobic core of a beta-barrel protein, cyclophilin based on the concept of complementarity discussed in ***Chapter 6.***

The Appendix is divided into two parts. The first part ***(Appendix I)*** *deals with a molecular dynamic simulation of cyclophilin and a novel analysis scheme for dynamic contact networks which can elucidate evolutionary relationships amongst members of the same fold. The second part* ***(Appendix II)*** *analyzes the geometry and electrostatics of salt bridges within native protein interiors.*

A part of the results described in this thesis have already been reported in the following publications.

1. Mapping the distribution of packing topologies within protein interiors shows predominant preferences for specific packing motifs. **Basu, S.**, Bhattacharyya, D., Banerjee, R. *BMC Bioinformatics.* (2011) 12, 195.

2. Self-complementarity within proteins : bridging the gap between binding and folding. **Basu, S.,** Bhattacharyya, D., Banerjee, R. *Biophysical Journal.* (2012) 102, 2605-2614.

3. SARAMA: a standalone suite of programs for the Complementarity Plot – a graphical structure validation tool for proteins. **Basu, S.,** Bhattacharyya, D., Banerjee, R. *Journal of Bioinformatics and Intelligent Control.* (2013). In Press.

Abbreviations

APCN	:	All atom Point Contact Network
ASCN	:	All atom Surface Contact Network
Bur	:	Burial / Burial ratio
CP	:	Complementarity Plot
CS	:	Complementarity Score
d_{net}	:	Network Distance
$\boldsymbol{E_m}$	:	Electrostatic Complementarity
μ	:	Mean
Ov	:	Overlap
Pd	:	Packing Density
Pen	:	Penalty
P_{grid}	:	Grid probability
φ_s	:	Swivel Angle
Res	:	Amino Acid Residue Identity
rGb	:	Residue given burial (Accessibility Score)
RMSD	:	Root Mean Square Deviation
SAA	:	Solvent Accessible Area
S_e	:	Sequence Entropy
σ	:	Standard Deviation
$\boldsymbol{S_m}$	:	Shape Complementarity
s_{net}	:	Network Similarity
θ_t	:	Tilt Angle
Vdw	:	Van der Walls
Vor	:	Voronoi
w.r.t.	:	With respect to

Introduction

1. Background

One of the objectives of the 'protein folding problem' is to decipher the 'second genetic code' which is to correlate the linear amino acid primary sequence to the three dimensional structure of the protein molecule. Generally, the protein folding problem can be divided into three inter-related though distinct sub-problems: i) the thermodynamic problem of how the three dimensional structure of a folded polypeptide chain remains thermally stable as a consequence of inter-atomic forces in both the protein molecule and the surrounding solvent ii) the kinetic problem of how a protein can converge to its native fold in characteristic time scales and iii) the computational problem of how to predict the native three dimensional structure from its amino acid sequence **(Dill et al, 2007)**.

With the advent of fast computers, computational protein structure prediction has received a great deal of attention for several decades, also driven by its many industrial and biomedical applications. If successful, it is by far much simpler and less expensive to predict protein structures computationally (if the procedure can be completed in reasonable computer time) compared to cumbersome experiments in structural biology involving X-ray crystallography, NMR or cryo-electron microscopy. Recent statistics show that there are currently ~28 million protein sequences deposited in the UniProtKB database **(Bairoch et al., 2005)** (http://www.ebi.ac.uk/swissprot) whereas the corresponding number of protein structures in the Protein Data Bank (PDB) **(Berman et al., 2000)** (http://www.rcsb.org/pdb) is only 85602 (dated 30.07.2013), which is about 0.3% of the total number of available protein sequences. In the post genomic era **(Figure 1)** this mismatch between the number of available sequences and the number of experimentally derived three dimensional structures appears to be increasing. Thus, there is an urgent need for robust computer based algorithms to accurately predict three dimensional protein structures.

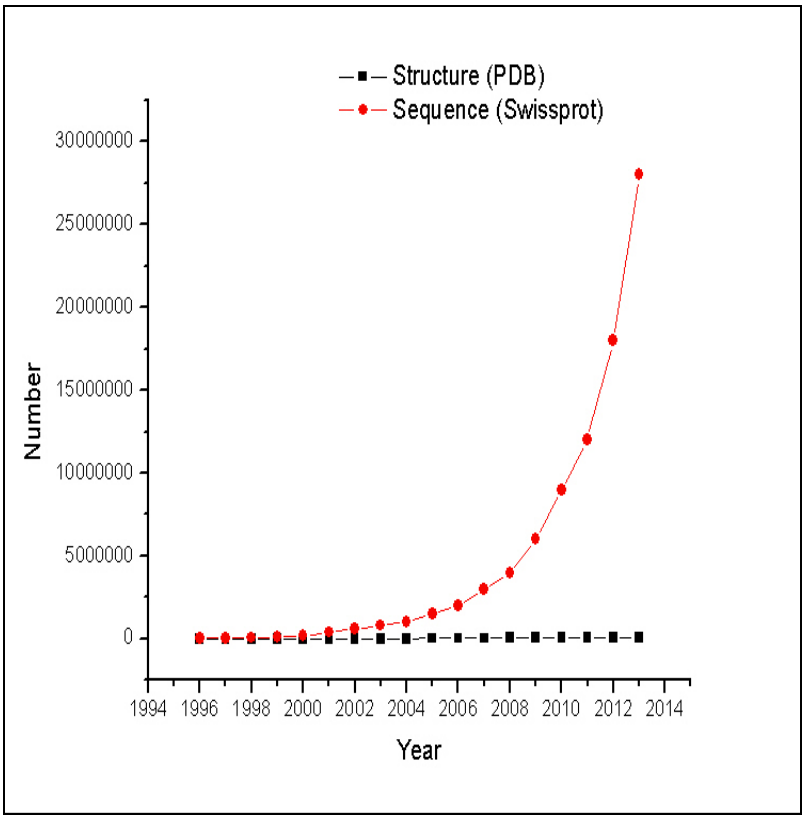


Figure 1. Growth in the number of protein structures (**black** squares) and sequences (**red** circles) in time.

1.1. Protein structure prediction

Broadly speaking, structure prediction methods fall into two classes: physics-based **(Oldziej et al., 2005)** and bioinformatics- or knowledge-based, **(Simons et al., 1997; Zhang and Skolnick, 2004a)** where the latter depends on the analysis of databases in order to design pseudo energy functions and derive probability distributions or propensities with regard to biophysical, geometrical features of natively folded structures. On the other hand, purely physics-based approaches rely on the accurate understanding of the physical mechanisms underlying the protein folding process. For a pair of homologous proteins with greater than 35% sequence identity, one (with an available experimental atomic structure) can be used as a template to derive an atomic model of the other. Many web-servers have been developed using different template-based modeling (TBM) techniques such as, GenTHREADER **(Jones, 1999),** FUGUE **(Shi et al., 2001),**

ORFeus **(Ginalski et al., 2003b),** PROSPECTOR **(Skolnick et al., 2004)** and MUSTER **(Wu and Zhang, 2008)**. However, in the absence of appropriate templates one has to resort to *ab initio, de novo* or free modeling methods. In contrast to *ab initio* modeling, *de novo* methods are allowed to incorporate prior statistical information, prediction of secondary structures, fragment assembly etc. in their algorithms.

Critical Assessment of Techniques for Protein Structure Prediction (CASP) is a biannual gathering of experts in the field of protein structure prediction taking place since 1994 **(Moult et al., 1995)**. The ranking results show that TBM and bioinformatics-based methods are far superior in both speed and prediction accuracy than *ab-initio* or physics-based methods **(Moult et al., 2011)**. To date, the most efficient and accurate knowledge-based method appears to be homology modeling **(Sali and Blundell., 1993)**, which generally predicts unknown structures to within 3 Å (C^{α}-RMSD) of the experimentally determined target, for homologous pairs with greater than 30% sequence identity. On the other end of the spectrum, *ab initio* methods for small single domain proteins of less than 90 amino acids appear to be moderately successful with algorithms predicting structures within ~2 to 6 Å of the native **(Bradley et al., 2005; Dill et al., 2007)**. Thus, there is considerable scope for the improvement of *ab initio* methods. Although TBM methods can generate reasonably accurate models, they generally do not give additional insight into the protein folding process, which could probably come from the successful implementation of *ab initio* methods.

1.2. ab-inito and knowledge-based methods

In *ab initio* modeling, a suitable energy function is first designed and a conformational search is conducted guided by this energy function, leading to the generation of a number of possible conformations (decoys) from which the final models are eventually selected. The power and validity of all *ab-initio* methods depend on (i) the intricate coupling of the energy function and the search procedure, (ii) the ability of the energy function to identify the structure closest to its most thermodynamically stable state compared to the decoys, (iii) the speed and efficiency of the conformational search

protocol to identify the low-energy states and (iv) the accurate selection of native-like models from a pool of decoys. Energy functions are again classified into either physics- or knowledge-based functions. In a physics-based energy function, atomic interactions are taken into account either by means of quantum mechanics or the use of a compromised force field with a large number of selected atom types **(Hagler et al., 1974, Weiner et al., 1984)**. Molecular dynamics simulations based on all atom physics-based force fields (e.g., AMBER **(Weiner et al., 1984; Cornell et al., 1995)**, OPLS **(Jorgensen and Tirado-Rives, 1988),** CHARMM **(Brooks et al., 1983)** etc.) is generally the method of choice for conducting a conformational search. However, MD-based *ab-initio* protein structure predictions usually take months or years to complete even for small proteins. An instructive example was the success of Pande and coworkers **(Zagrovic et al., 2002)** (http://folding.stanford.edu/) to fold the villin headpiece (consisting of 36 residues) to 1.7 Å (C^{α} RMSD w.r.t experimental target) with a total simulation time of 300 ms (~ 1000 CPU years). However, these methods are far from being routinely used for the structure prediction of typical-size proteins (~100-300 residues) for reasons of speed and accuracy.

On the other hand, knowledge-based potentials are empirical in nature, statistically derived from database analysis and can be divided into two types **(Skolnick et al., 2006)** : those containing sequence-independent terms e.g., hydrogen bonding, local backbone stiffness etc. **(Zhang et al., 2003)** and those with sequence dependent terms, e.g. pair-wise residue contact potential **(Skolnick et al., 1997)**, distance dependent atomic contact potential **(Samudrala and Moult., 1998; Shen and Sali., 2006)**, and secondary structure propensities **(Zhang et al., 2003; Zhang and Skolnick., 2005)**. Most knowledge-based energy functions are coupled to Monte Carlo (MC) conformational search procedures e.g., TASSER **(Zhang and Skolnick., 2004b)**, I-TASSER **(Wu et al., 2007)** and have much faster runtime (hours) compared to physics-based potentials. One of the most successful and popularly used *ab initio* method (in the knowledge-based / MC category) has been the 'Fragment Assembly' **(Bowie and Eisenberg., 1994)** approach which assembles small fragments (mainly 9-mers) taken from the PDB. An analogous approach was implemented in ROSETTA **(Simons et al., 1997)** by Baker and coworkers,

which was remarkably successful for the free modeling (FM) targets in CASP experiments. Latter, substantial improvements were made in the ROSETTA method **(Bradley et al., 2005; Das et al., 2007)** which also involved several physics-based energy terms (van der Waals interactions, pair wise solvation free energy, and an orientation-dependent hydrogen-bonding potential). The methodology initially generates reduced C_β conformations specified with heavy backbone and thereafter, an all-atom refinement is performed on a set of selected low-resolution models using the above mentioned physics-based energy terms. Multiple rounds of MC minimization are then carried out for the conformational search **(Li and Scheraga., 1987)**. One of the most notable successes of ROSETTA has been the blind prediction of an *ab initio* target (T0281 from CASP6, 70 residues) whose C_α-RMSD from its crystal structure was 1.6 Å **(Bradley et al., 2005)**. In recent years, extensive samplings are being carried out using the worldwide distributed computing network of Rosetta@home (http://boinc.bakerlab.org/rosetta/) allowing about 500,000 CPU hours for each target domain. However, it should be noted that comparing the performance of different prediction methods is made particularly difficult as different algorithms are tested on different proteins of choice (by different research groups) rather than a standard protein test set **(Helles, 2008)**.

2. Fold Recognition

Fold recognition is a structure prediction of comparable lesser complexity wherein the main-chain coordinates are given and the problem is to select side-chain sequences (which could include side–chain conformations) consistent with and supportive of a given native fold. One of the prime concerns of the 'fold recognition' problem is to correctly identify the fold, which happens to be the same, for a pair of sequences with low identity upon alignment (<30%, falling in the twilight zone) amidst a pool of random sequences. This is especially important for folds with a large diversity in sequences and functions among their members (e.g., immunoglobulin, Rossman-like, Tim barrel, globin fold etc.). A further refinement of the fold recognition problem could be to solve for the correct set of χ angles determining side-chain conformations.

Threading techniques **(Brayant and Lawrence., 1993; Jones, 1999)** in fold recognition are especially useful in selecting possible tertiary structures, for a rapidly expanding pool of genomic sequences with no identifiable evolutionary relationships. Scoring functions discriminating the correct sequence – structure match from decoys, also finds widespread use in protein folding simulations **(Park and Levitt., 1996)**, *ab-initio* structure predictions **(Samudrala and Moult., 1998; Kinch et al., 2011)** and the selection of the best model from a repertoire of NMR structures for molecular replacement calculations **(Huang et al., 1996)**. Such scoring functions are again physics-based that is based on atomic interactions modeled by appropriate force fields or knowledge-based, formulated by including parameters which have been extracted from a database of experimental structures. Knowledge-based scoring functions are probably preferred due to the ease with which they lend themselves to efficient computation over large decoy sets and the last decade has witnessed considerable improvements in both their performance and variety. Several knowledge based discriminative scoring functions are now available, based on the analysis of pair-wise amino acid interactions by techniques of statistical thermodynamics **(Sippl, 1995; Arab et al., 2010)**, weighted matching of sequence profiles generated from multiple sequence alignment **(Yang et al., 2011)**, use of torsion angle profile and profile based gap penalties **(Zhang et al., 2008)**, average solvent accessible surface areas of residues in correctly folded proteins **(Bahadur and Chakrabarti., 2009),** extraction of correlated mutations **(Sadowski et al., 2011),** use of fold-specific position-specific scoring matrices **(Hong et al., 2011),** incorporation of local structural preference potential **(Hu et al., 2011)** and on the structural features of hydrophobic cores **(Huang et al., 1996)**. Most scoring functions easily distinguish the native structure from decoys composed of random sequences and the current challenge is to identify the native structure from a pool of decoys which are native-like both in terms of sequence and/or certain three dimensional features. Several such datasets are now available composed of numerous decoy models generated by highly diverse computational methods, most of which take special care to optimize the rotamer arrangement and minimize steric clashes. Of these, PROSTAR **(Holm and Sander, 1992)**, Park and Levitt decoy sets **(Park and Levitt., 1996)**, Rosetta **(Tsai et al., 2003)**

and several others effectively challenge scoring functions to prove their mettle. However, the most challenging test for these scoring functions is perhaps to identify the native structure among its best-predicted near-native models submitted by different groups in CASP. On the other hand, PREFAB (http://www.drive5.com/muscle/prefab.htm) **(Edgar, 2004)** is a widely used database to identify pairs of proteins with low (<30%) sequence identities (upon alignment) although belonging to the same fold.

There are both single as well as multiple decoy sets. In single-decoys, a successful hit refers to the correct discrimination of the native from its decoy counterpart whereas in case of multiple decoys, assessment of the discriminatory ability of a scoring function is given by a 'Z-score' (associated with a corresponding rank of the native structure). Missfold **(Holm and Sander, 1992)**, Pdberr **(Branden and Jones, 1990)**, sgpa **(Avbelj et al., 1990)** are common examples of single decoys generated with different strategies. 'Missfold' consists of 26 pairs of proteins with identical chain length but different sequences and conformations. The 'Pdberr' decoy set consists of three correctly solved X-ray crystal structures along with their erroneous decoy counterparts, whereas 'sgpa' contains the experimental structure of *Streptomyces griseus* Protease A (2SGA) and its two corresponding decoys, generated by molecular dynamics simulations. A relative success rate is compared amongst different knowledge-based scoring functions in the following table **(Table 1)**. As can be seen, most functions perform equally well in these single decoy sets.

Table 1: Comparison in the performances of different knowledge-based scoring functions on single decoy sets: The functions include R_s, R_p **(Bahadur and Chakrabarti., 2009)**, RAPD, CDF **(Samudrala and Moult, 1998)**, Surfield **(Arab et al., 2010),** atomic knowledge based potential (AKBP) **(Lu and Skolnick, 2001)**, Residue Contact Potential (RCP) **(Skolnick et al., 2000)**. The number of successful hits / total number of trials are tabulated. Data obtained from Table 3 of **Bahadur and Chakrabarti., 2009.**

Scoring Functions	Misfold	Pdberr and sgpa
R_s	24/24	5/5
R_p	20/24	5/5
RAPD	24/24	5/5
CDF	19/24	5/5
Surfield	23/23	-
AKBP	24/24	5/5
RCP	24/24	4/5

Among different multiple decoy sets, '4-state reduced' **(Park and Levitt, 1996)** and Fisa **(Simons et al., 1997)** are the two most common examples. The former consists of 7 sequences (medium in chain length : 54-75 residues), each with nearly 600-700 decoys (with sequences identical to the native) that include structures with RMSD (C^{α} atoms) ranging from 0.8 to 9.4 Å from the native whereas 'Fisa' **(Simons et al., 1997)** contains 4 small (43-76 residues) all-α proteins with 500 decoys for each set. Different knowledge-based scoring functions are tested and compared in these multiple decoys. One of the most successful approaches appears to be the design of functions based on non-interacting ideal gas reference state assuming that atoms can be modeled as ideal gas molecules. Distance Scaled Finite Ideal gas Reference State or DFIRE **(Zhang et al., 2004)** is one such scoring function based on the subsequent assumption that the distribution of pairwise interaction follows the uniform distribution in the whole volume of the protein. Discrete Optimized Protein Energy function or DOPE **(Shen and Sali, 2006)** is another where no interacting atoms are present in a homogeneous sphere as reference state. Both these functions seem to perform better than most other functions in multiple decoys **(Table 2)**.

Table 2: Comparison in the performances of different knowledge based scoring functions on multiple decoy sets: The functions include DFIRE **(Zhang et al., 2004)**, Rosetta **(Misura et al., 2006)**, ModPipe-Pair (MPP), ModPipe-Surf (MPS) **(Melo et al., 2002)**, TE13, LHL **(Li et al., 2003)**, Force Model (FM) **(Mirzaie et al., 2009)**, DOPE **(Shen and Sali, 2006)**, MJ **(Miyazawa and Jernigan, 1996),** R_s, R_p **(Bahadur and Chakrabarti., 2009).** All entries in the table refer to the rank of the native structure as detected by the corresponding method. Part of the table has been reproduced from **Arab et al., 2011.**

Decoy Set	PDB ID	DFIRE	Rosetta	MPP	MPS	TE13	LHL	FM	DOPE	MJ	R_s	R_p
4state reduced	1CTF	1	1	1	1	1	1	1	1	1	1	1
	1R69	1	2	1	17	1	1	8	1	1	1	19
	1SN3	1	1	1	7	6	1	23	1	2	5	23
	2CRO	1	5	1	103	1	1	4	1	1	1	1
	3ICB	4	6	15	33	-	5	2	1	-	1	6
	4PTI	1	1	1	71	7	1	13	1	3	1	1
	4RXN	1	1	1	18	16	51	85	1	1	1	1
Fisa	1FC2	254	158	491	1	-	-	1	357	-	-	-
	1HDD-C	1	90	293	18	-	-	1	1	-	-	-
	2CRO	1	26	11	146	-	-	1	1	-	-	-
	4ICB	1	1	196	2	-	-	1	1	-	-	-

As has been mentioned, probably the most challenging test for a fold-recognition function is to be able to rank the native targets amongst the CASP models. A thorough comparison **(Table 3)** using CASP7 models shows that functions based on average solvent accessibilities **(Bahadur and Chakrabarti., 2009)** performs the best, though, as can be seen from the percentage of native structures correctly identified, there is substantial room for improvement.

Table 3. $\mathbf{Z_{nat}}$ corresponds to the average Z-score of the native structure. Percentage of the native structure with rank 1 or within rank 10 from among all the solutions submitted in CASP7 are tabulated. The functions include RAPDF **(Samudrala and Moult, 1998)**, DFIRE **(Zhang et al., 2004)**, QMEAN3 **(Benkert, 2008)** and R_p , R_s **(Bahadur and Chakrabarti., 2009)**. Data obtained from Table 6 of **Bahadur and Chakrabarti., 2009.**

	% of native structure		
Method	$\mathbf{Z_{nat}}$	**Rank1**	**Rank10**
RAPDF	-2.09	57.89	81.05
DFIRE	-1.25	62.11	75.79
QMEAN3	-2.27	62.11	78.95
R_p	1.69	53.52	91.55
R_s	2.17	71.83	98.59

A part of this thesis describes the design and utility of new scoring functions for fold recognition (discussed in detail in Chapter 3) tested on several state-of-the-art decoy sets and compared with the best knowledge-based functions currently available in the literature. These functions were built utilizing the concept of 'complementarity' in bimolecular recognition which is briefly reviewed in the following section.

3. Complementarity

In molecular recognition the term 'complementarity' is used to describe the match between two interacting molecular surfaces and is supposed to have a dual aspect, 1) shape complementarity arising out of the steric fit between closely packed interfacial atoms in van der Waal's contact and 2) electrostatic complementarity mediated by long range electric fields due to charged or partially charged atoms. Within the domain of biomolecular recognition, the concept appears to be particularly appealing for protein-protein interfaces due to their large interfacial surface area (~1600 $Å^2$) buried upon complexation **(Lo Conte et al., 1999)** which is possible due to the match between the interacting surfaces in terms of both shape and chemical properties. However, in case of small molecule ligands binding to proteins, significant diversity in conformation,

variation in shape and physicochemical environments experienced by the identical ligand in binding pockets of unrelated proteins have been demonstrated **(Stockwell and Thronton, 2006; Kahraman et al., 2007; Kahraman et al., 2010).**

3.1. Shape Complementarity

Early methods for computing shape complementarity at protein-protein interfaces include the estimation of buried surface area **(Chothia, 1974)**, paucity of buried water molecules **(Chothia and Janin, 1975)** and packing density of interfacial atoms **(Richards, 1974)**. In 1983, an analytical method to calculate smooth 3D contours for proteins was developed by Connolly **(Connolly, 1983a, Connolly, 1983b)** describing a protein surface as critical points and surface normals which has formed the basis of curvature-dependent shape complementarity for protein-protein interfaces. Subsequently, Lawrence and Colman **(Lawrence and Colman, 1993)**, defined a shape correlation statistic, S_c, to probe shape complementarity in protein quaternary association, protein-inhibitor and antigen-antibody complexes, a modified version of which has been adopted in this study. A brief description of the shape correlation statistic, S_c is as follows. To start with, the entire Connolly surface **(Connolly, 1983a)** of both the interacting protein molecules is sampled as discrete area elements at a sufficiently high sampling density (15 dots / $Å^2$). Then, the interface atoms buried upon association of the two interacting protein molecules are identified. Nearest neighboring dot surface points of each buried area element from each of the two surfaces are subsequently identified within a distance cutoff of 1.5 Å and the following expression **(Lawrence and Colman, 1993)** calculated.

$$S^{A\to B} = (n_A \cdot n`_A)\exp[-w(|x_A - x`_A|^2)]$$
$$S^{B\to A} = (n_B \cdot n`_B)\exp[-w(|x_B - x`_B|^2)]$$
$$S_c = (\{S^{A\to B}\} + \{S^{B\to A}\})/2$$

where $S^{A\to B}$ and $S^{B\to A}$ may be defined at every point on P_A and P_B **(Figure 2)** and w is a scaling constant set to 0.5 $Å^{-2}$ and the braces define the median (50^{th} percentile) of the distribution of $S^{A\to B}$ and $S^{B\to A}$ values over the surfaces P_A and P_B respectively. The

median was chosen as a better measure of central tendency since the distribution of S_c values was found to be negatively skewed **(Lawrence and Colman, 1993)**.

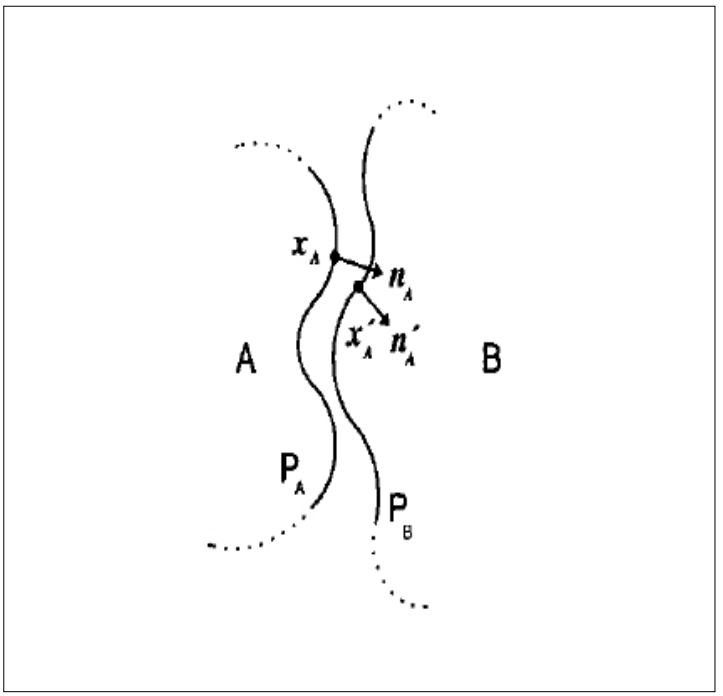


Figure 2. The shape correlation statistic, S_c. A and B are two interacting proteins. P_A and P_B are the portions of the molecular surface of A and B which are buried from solvent by their mutual interaction; x_A and x_B are two points on P_A and P_B, with n_A and n_B being two unit vector normals (outwardly and inwardly oriented respectively) to P_A and P_B at x_A and $x`_A$ which is the point on P_B nearest to x_A. Figure reproduced from **Lawrence and Colman, 1993.**

The shape correlation statistic has a theoretical range of -1.0 to 1.0. It is evident from the definition of S_c that interfaces with S_c = 1.0 will fit together precisely (e.g., identical surfaces), those with S_c = 0 will have uncorrelated topography and those with S_c = -1.0 will have perfect mismatch or anti-correlation among their protrusions and crevices. Thus, good surface fit for naturally occurring biomolecular interfaces should approach the value of 1.0. For oligomeric assemblies and protein / protein inhibitor

interfaces, S_c values were found to be in the range: 0.70 to 0.76 whereas for antigen / antibody interfaces, S_c ranges from 0.64 to 0.68.

The shape correlation statistic, S_c, is advantageous over other earlier quantitative measures of shape complementarity for more than one reason. Firstly, S_c measures correlation of directions and is therefore relatively insensitive to the precise values of atomic radii used to generate the molecular surface. $S^{A \rightarrow B}$ and $S^{B \rightarrow A}$ are designed in such a manner that the scalar product term dominates at close surface proximity (the exponential weighting term drops to 0.9 at a surface element separation of 0.45 Å). Thus, S_c measures complementarity rather than surface separation whenever the surfaces are very close to each other **(Lawrence and Colman, 1993)**. Another measure to estimate molecular goodness of fit (specifically designed for protein interiors) is the 'small–probe contact dot' algorithm **(Word et al., 1999a)**. Contact dot surfaces are somewhat related to the concept of configuration dependent exposed surfaces **(Lee and Richards, 1971)**. However, unlike the Lee & Richards algorithm, where a probe sphere of 1.4 Å (water) is rolled around the van der Waals surface of each atom to compute its solvent accessible surface, here, a smaller probe sphere (typical radius of 0.25 Å) is used. This small probe is placed on a set of predefined points and leaves a dot when it touches another atom located at least three covalent bonds away from the atom on whose surface point the probe sphere has been placed **(Figure 3)**. One of the major conclusions drawn from the 'small probe contact dot' algorithm is the importance of explicit hydrogens **(Word et al., 1999b)** in the analysis and their contacts in specific interactions between and within molecules. Also, based on this algorithm, ('clashes' of the contact dots inclusive of hydrogen contacts) a structure validation technique has been incorporated in the software 'Molprobity' **(Davis et al., 2007)**.

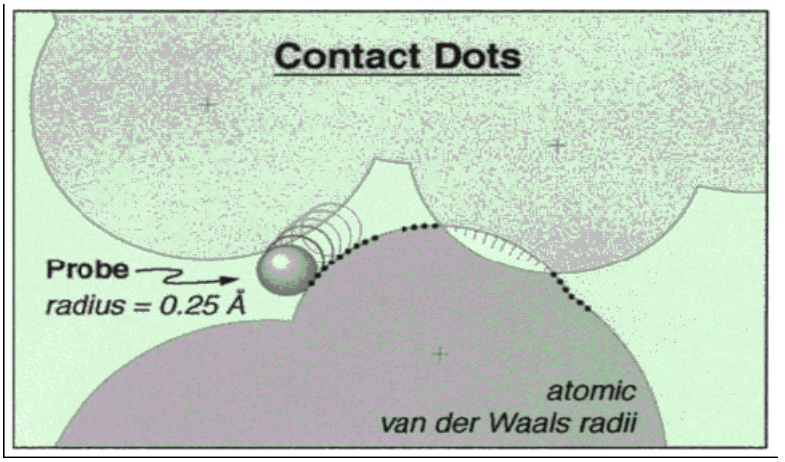


Figure 3. A schematic representation of the small-probe contact dot algorithm. The small probe (0.25 Å) sphere rolls over the van der Waals surface of each atom, leaving a dot sequentially wherever it also touches another atom that is not within three covalent bonds. The unfavorable contact at overlap sites of non-Hydrogen bonding atoms is emphasized by spikes instead of dots. Figure reproduced from **Word et al., 1999b.**

3.2. Electrostatic Complementarity

Another aspect of complementarity at the protein-protein interface is a consequence of the electrostatic interaction between proteins which has been studied extensively. Previously, elementary analysis was performed by simply counting the number of charged residues and salt bridges at protein-protein interfaces **(Janin and Chothia, 1990; Jones and Thornton, 1995)**. Later, when the solvent continuum electrostatics model became available for proteins, electrostatic potential of protein structures were frequently determined by iteratively solving the 2^{nd} order partial differential Poisson-Boltzmann equation for the protein-solvent system as implemented in the software DelPhi **(Gilson et al., 1988; Nicholls and Honig, 1991)**. The Poisson-Boltzmann equation describes the variation of electrostatic potential in space due to a distribution of charges in a multi-dielectric environment (Poisson equation) coupled to the distribution of partial charges on the protein and counter-ion (the latter assumed to be a Boltzmaan distribution). The Poisson-Boltzmann equation thus is an efficient way to

solve for the electrostatic potential as a function of the dielectric constant and the charge-density throughout space **(Mandell et al., 2001)**.

Generally, the electrostatic potential is visually represented by color coding regions (red for negative and blue for positive potential) on the molecular surface by drawing equipotential contours on and around the protein. Such visualizations have been insightful in protein-protein and protein-ligand docking **(Getzoff et al., 1983)** and predicting protein-protein association sites **(McDonald et al., 1991)**. Apart from visualization of the electrostatic potential at the surface buried in the interface, binding free energy calculations contributed by the electrostatic union of the proteins are also possible by continuum electrostatic models, given the availability of coordinates of a protein-protein complex **(Gilson et al., 1988)**.

Analyses of the electrostatic nature of protein-protein interfaces using continuum electrostatic calculations have been extensively used to determine whether surfaces involved in protein–protein interactions have either "charge complementarity" **(Novotny and Sharp, 1992; Roberts et al., 1991)** or "electrostatic complementarity" **(Braden and Poljak, 1995; Demchuk et al., 1994; Hendsch and Tidor, 1994).** In the work of McCoy et al., **(McCoy et al., 1997)** the complementarity appeared to be more in terms of electrostatic potential rather than charge. They further showed that over and above interfacial salt bridges which do make an important contribution to the electrostatic potential the rest of the atoms from the polypeptide chains also contribute significantly to the overall electrostatic complementarity at the interface. The delineation of the dielectric boundary is an essential aspect for continuum electrostatic calculations. In the same work it was also demonstrated that a partially desolvated model (protein being partially desolvated by the volume of the other protein in the complex, thus, leaving a low dielectric region in the close vicinity of the interacting molecule) was most appropriate for such continumm electrostatic calculations rather than a fully solvated model (where the dielectric surrounding the protein is considered to be high, that is of the solvent: 80).

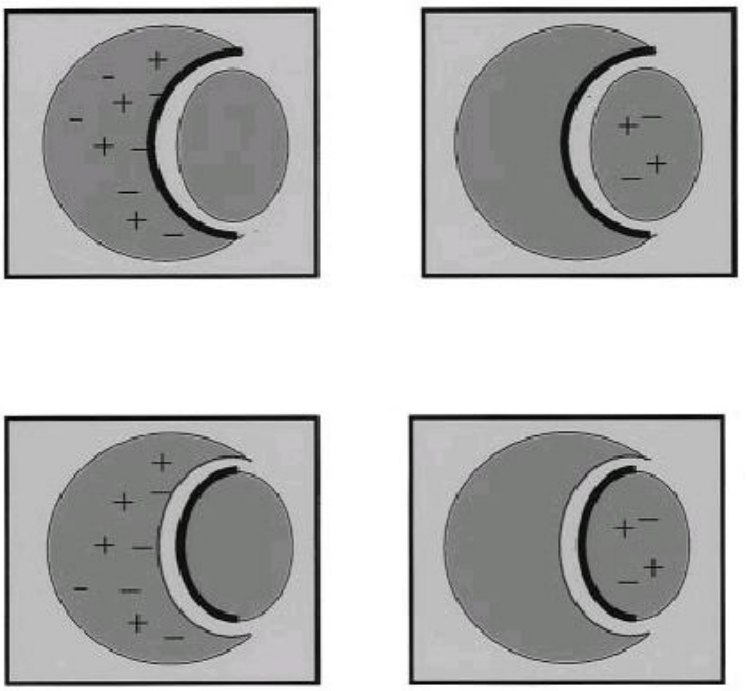


Figure 4. A schematic representation of the partially desolvated model for continuum electrostatic calculation at protein-protein interfaces. The thick black lines denote the buried molecular surfaces on the two protein molecules. Electrostatic potential calculated twice on each buried molecular surface. Each time the charged atoms of one of the two interacting proteins contribute to the potential. The atoms of the other protein molecule are only assigned their van der Waals radii with zero charges (dummy atoms) to maintain the scaling and orientation of the molecule on the grid and also to correctly delineate the dielectric boundary. Therefore the region occupied by the partner molecule has a low dielectric constant. Figure reproduced from **McCoy et al, 1997**.

To compute the electrostatic (potential) complementarity, first, the molecular surfaces **(Connolly, 1983a)** of the two interacting proteins were generated and the surface elements corresponding to the interfacial atoms determined. Then, the potential at each interface were enumerated (by solving numerically the linearized Poisson-Boltzmaan equation as implemented in DelPhi **(Nichollos and Honig, 1991)**) twice, once each due to the partially charged atoms of the two protein molecules. Thus, each surface point was tagged with two values of electrostatic potential. The negative of the correlation coefficient (Pearsons / Spearman) of these two sets of potential values (for

each of the two interfaces) were then averaged to give the electrostatic complementarity (EC) at the interface. But for a few exceptions (e.g., 3HFM), the electrostatic complementarity values were mostly found to be within the range of 0.55 to 0.7 for most protein-protein complexes **(McCoy et al, 1997)**.

3.3. Complementarity in Docking

The elevated values attained for surface and electrostatic complementarity for intra-protein association has served the basis for the design of scoring functions in protein-protein docking which eventually aims to derive a model for the bound structure starting from the 3D coordinates of two independently crystallized proteins which are known to interact in solution. A variety of geometric and electrostatic complementarity functions have been extensively used in the design of scoring functions for different protein-protein docking algorithms (e.g., ClusPro, ZDOCK, DOT) **(Chen and Weng, 2003; Mandell et al., 2001; Comeau et al., 2004; Tovchigrechko et al., 2002)**. These discriminatory scoring functions incorporate binding energy components into the process, based on the assumption that the native structure is at a global free energy minimum. Other discrimination methods try to refine the interface, as the surface side-chains of the independently crystallized proteins are frequently found in wrong orientations **(Kimura et al., 2001)**. This interface-refinement substantially improves the van der Waals contact energy, leading to an increase in surface complementarity. Also, many electrostatic interactions and hydrogen bonds can be recognized as energetically favorable, facilitating a more successful discrimination **(Comeau et al., 2004)**. Many docking algorithms especially those employing convolution techniques or computer vision techniques have used geometric complementarity as their primary scoring function **(Fischer et al., 1993; Norel et al., 1994)**. In other functions, a somewhat higher weightage is given to geometric descriptors than electrostatic components **(Comeau et al., 2004)**. Some shape complementarity (e.g., in ZDOCK) functions are not explicitly based upon protein surface curvature or surface area, rather rewards continuous surface patches at the binding site and penalizes clashes **(Chen and Weng, 2003)**. Most common geometric

(complementarity) descriptors are based on van der Waals contact energy **(Mandell et al., 2001)** and steric scoring schemes based upon 'soft' potentials **(Walls and Sternberg, 1992)**. Since it is not feasible to explore all possible conformations, the 'softness' involved in the design of these geometric descriptors is particularly important to tolerate structural imperfections without leading to an increased number of false positives **(Chen and Weng, 2003)**. On the other hand, both coulombic potential (with a distance dependent dielectric) and Poisson-Boltzmaan electrostatic energies have been used as different electrostatic descriptors. A good composite scoring function attempts to model hydrogen bonds electrostatically and hydrophobic interactions through van der Waals contacts (e.g., in DOT) **(Mandell et al., 2001)**. Electrostatic terms are often associated with a desolvation free energy term using atomic contact potentials **(Comeau et al., 2004)**. Atomic contact potential is a smooth potential and thus varies little for small distance perturbations (e.g., coordinate errors), in sharp contrast to coulombic potential terms which are particularly sensitive **(Comeau et al., 2004)** especially to coordinate errors of solvent accessible side-chains. Hence, an increased number of structures are allowed to pass through the electrostatic filter compared to the desolvation filter, in an attempt to retain more near-native models. Some docking algorithms (e.g., DOT) use surface and electrostatic complementarity terms in a two-step manner for the initial screening and subsequent clustering of putative complexes starting from the order of billions **(Mandell et al., 2001)**. In particular, coulombic electrostatic energies have been effectively used as a secondary filter to discard predictions with unfavorable charge interactions although having high geometric-fit **(Gabb et al., 1997)**. However, a composite energy function has been shown to perform better than either van der Waals energy (geometric fit) or electrostatic energy alone **(Mandell et al., 2001)**. In line with this view, the current study also probes both geometric and electrostatic 'complementarity' within protein interiors and combines them to be used in several ensuing applications. One such application of surface complementarity was the study of side-chain packing within the protein interior which is briefly reviewed in the next section.

4. Side-chain packing within protein interiors and contact networks

Stereo-specific packing of side-chain atoms within native protein interiors has been considered to be one of the crucial factors determining the isomorphism between sequence and fold **(Crick et al., 1953)**. Interior packing within proteins is generally very dense (packing density: 0.7 to 0.8) resembling crystalline solids **(Richards, 1977)**. Francis Crick **(Crick et al., 1953)** generalizing from the packing of helices in proteins (knobs into holes) postulated that side-chain packing most probably resembled a three dimensional *jigsaw* puzzle (the '*jigsaw* puzzle' model). In contrast the 'nuts and bolts' model lies at the other extreme, which believes that dense packing in the protein interior does not require the geometric specificity between interacting side chains and can be achieved merely by the compaction of atoms within a constrained volume **(Bromberg and Dill, 1994)**. One way to distinguish the two models is the degree of conformational freedom acquired by buried side chains upon systematic volume expansion of the polypeptide chain. In the *jigsaw* puzzle model, the interlocking of the side-chains remains intact upon a systematic expansion of the polypeptide chain till a critical point of disjuncture (~25% increase in volume) beyond which there is expected to be an abrupt increase in conformational entropy **(Shakhnovitch et al., 1989)**. On the other hand, a gradual and continuous increase in side-chain entropy can be expected in case of the nuts and bolts model, as substantial conformational freedom could be gained from the very beginning of the volume expansion **(Dill et al., 1995)**. Over the years however, some doubt has been cast on the *jigsaw* puzzle model as most proteins appear to be resilient to core mutations provided the hydrophobic composition of the core remains unaltered. A dramatic example was in the case of phage T4 lysozyme **(Gassner et al., 1996)** where the protein could retain its overall fold (though with reduced thermal stability and activity), despite several core mutations (seven residues to methionine).

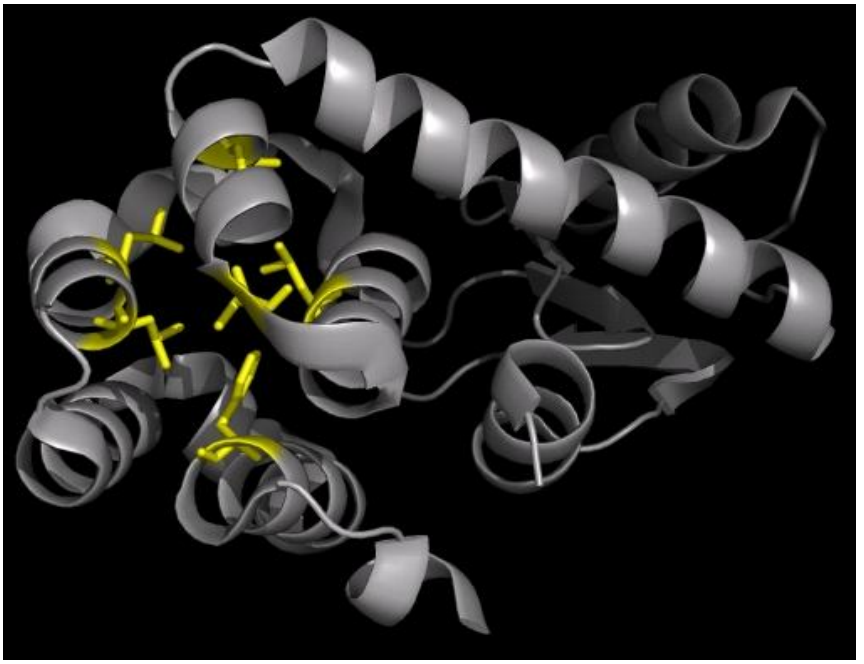


Figure 5. Multiple methionine substitutions within the core of T4 Lysozyme. The core residues which have been mutated to methionine are colored in yellow (PDB ID: 1KS3) **(Gassner et al., 1996)**.

Structural studies of these and similar mutants have shown how conformational relaxation of both main- and side-chain atoms could compensate the deleterious effects of such mutations, thereby preserving the overall fold **(Eriksson et al., 1992; Buckle et al., 1996)**. The random mutation of the 12 out of 13 core residues of ribonuclease barnase was another example where 23% of the mutants retained their enzymatic activity in vivo **(Axe et al., 1996).** Design of novel hydrophobic protein cores show that modes of packing other than the native could also sustain a stable fold **(Lim and Sauer, 1989; Hurley et al., 1992)**. It followed that the pattern of hydrophobicities embedded in the polypeptide chain could probably play a more decisive role in determining the fold, than the details of side chain packing, and a binary (hydrophobic H – polar P) representation of the polypeptide chain should be sufficient to encode for the three dimensional fold of the protein molecule **(Beasley and Hecht, 1997)**. This hypothesis probably also found support from the assumption that the hydrophobic effect plays a predominant role

(relative to van der Waals forces and hydrogen bonding etc.) in protein folding **(Dill et al., 1990)**. The more recent 'fuzzy-oil-drop' model also assumes the hydrophobic collapse to be the driving force in the folding process and thereby takes into account the dynamic fluctuations in native protein cores **(Brylinski et al., 2006; Brylinski et al., 2007)**. However, studies on aromatic side-chain interactions in proteins have shown that aromatic pairing occurs after, rather than before, the formation of secondary structures **(Thomas et al., 2002)**. Both experimental and computational studies have been carried out to design proteins based on the binary H-P code. Design of a 4-helix bundle with arbitrarily chosen polar and non-polar residues periodically placed in the sequence like that of an alpha-helix led to a soluble fraction of 60% of the designed population **(Kamtekar et al., 1993)**. *In silico* lattice simulation studies in two and three dimensions could actually fold binary H-P sequences into compact structures **(Sikorski and Skolnick, 1989; Lau and Dill, 1990)**. All such studies indicated that the pattern of hydrophobicities in the primary sequence appears to play a more crucial role in determining the overall fold than the geometrical constraints in packing. Nevertheless, stereospecificity of the interacting interior residues within proteins does contribute to fold stability as randomly redesigned cores (without any specific attempt to achieve optimal packing) led to either a disordered collapsed globule (overpacking) or complete unravelling of the structure (underpacking) **(Dahiyat and Mayo, 1997; Lazer et al., 1997)**. It thus becomes imperative to elucidate the geometrical constraints imposed on side chain packing which could also contribute to the *de novo* design of stable protein structures. In other words, attainment of dense, well-packed protein cores does not appear to arise automatically in the design process nor is it acquired simply by chance **(Desjarlais et al., 1995; Dahiyat et al., 1997)**. An instructive example was the repeated failure to design parallel $(\alpha/\beta)_8$ – TIM barrel **(Goraj et al., 1990; Tanaka et al., 1994)**, finally resolved successfully by Offredi *et al*. **(Offredi et al., 2003)**, where a term optimizing for side chain packing specificity was deliberately included in the computational process. Over the years, theories with regard to side-chain packing within proteins have not only provided insight into protein structures **(Brocchieri and Karlin, 1994; Mitchell et al., 1997; Banerjee et al., 2003; Misura et al., 2004)** but also have

facilitated their modelling **(Cooper et al. 2010; Miao et al., 2011)** validation **(Hooft et al., 1996; Davis et al., 2007; Sheffler et al., 2009)**, prediction **(Bradley et al. 2005; Raman et al., 2009)** and design **(Benjamin and Havranekb, 2011; Li et al., 2013)**.

Initial attempts to find preferred modes of packing within the native proteins, however, were largly unsuccessful **(Behe et al., 1991)**, especially for binary association between aromatic side-chains **(Singh and Thornton, 1985)**. However, significant deviations from a random distribution for interplanar and polar angles specifying the geometry for interacting pairs of side chains were also found **(Samanta et al., 1999; Brocchieri and Karlin, 1994; Mitchel et al, 1997)**.

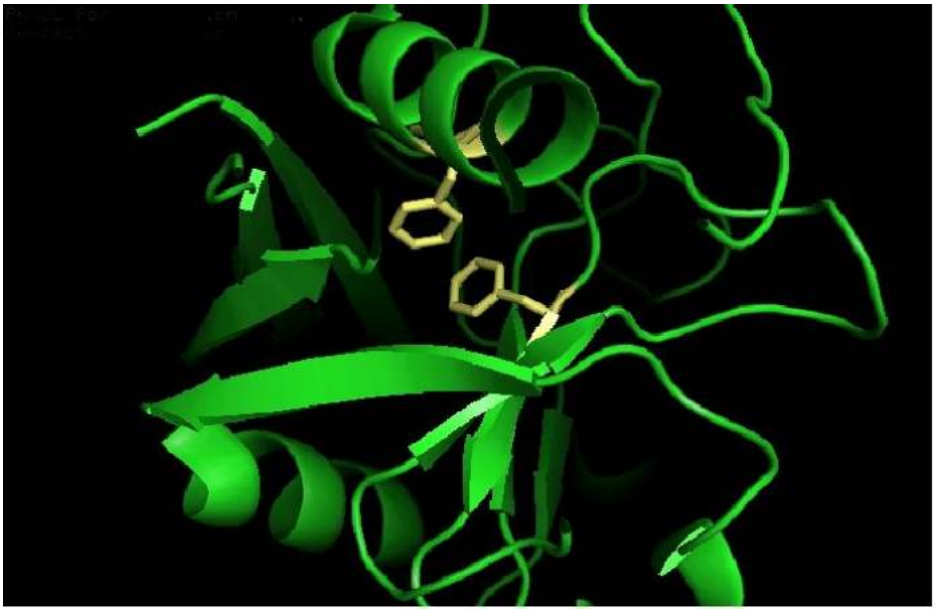


Figure 6. Packing of two interacting phenyl alanine side-chains within the protein interior (PDB ID: 2HAQ)**.**

Most of these studies indentified the binary pairs of interacting residues based on the proximity of their side-chain atoms which may not always reflect specific inter-residue interactions. This interaction criterion was later replaced by a more sophisticated criterion based on surface complementarity **(Lawrence and Colman, 1993, Banerjee et**

al., 2003). The study demonstrated that binary association between two hydrophobic side chains (Leu-Leu, Leu-Phe etc), with high surface fit and maximal overlap between their corresponding residue surfaces, did indeed exhibit specific inter-residue geometry (significant deviation from a random distribution in at least one of the inter-residue orientational angles). It was thus clear that at least for a subset of contacts (with high fit and overlap) predictions of the *jigsaw* puzzle model were indeed valid. In effect, the study also established quantitative measures (in terms of surface complementarity and overlap) to identify interacting residue pairs with specific geometry. Later studies strove to develop accurate models of the forces (including π-π, cation-π, van der Waals and hydrophobic interactions) sustaining these preferred modes of side-chain packing **(Misura et al., 2004)**. This led to the development of orientation dependent pair-wise potential for identification of the native structure from decoys **(Misura et al., 2004)** and their eventual use in structure prediction **(Bradley et al. 2005)**. Binary interacting pairs were classified according to their functional group (e.g, aliphatic pairs, arometic pairs etc.) and distinct specific modes of packing were identified and categorised (T-shaped, parallel stacked) according to their preferred orientation.

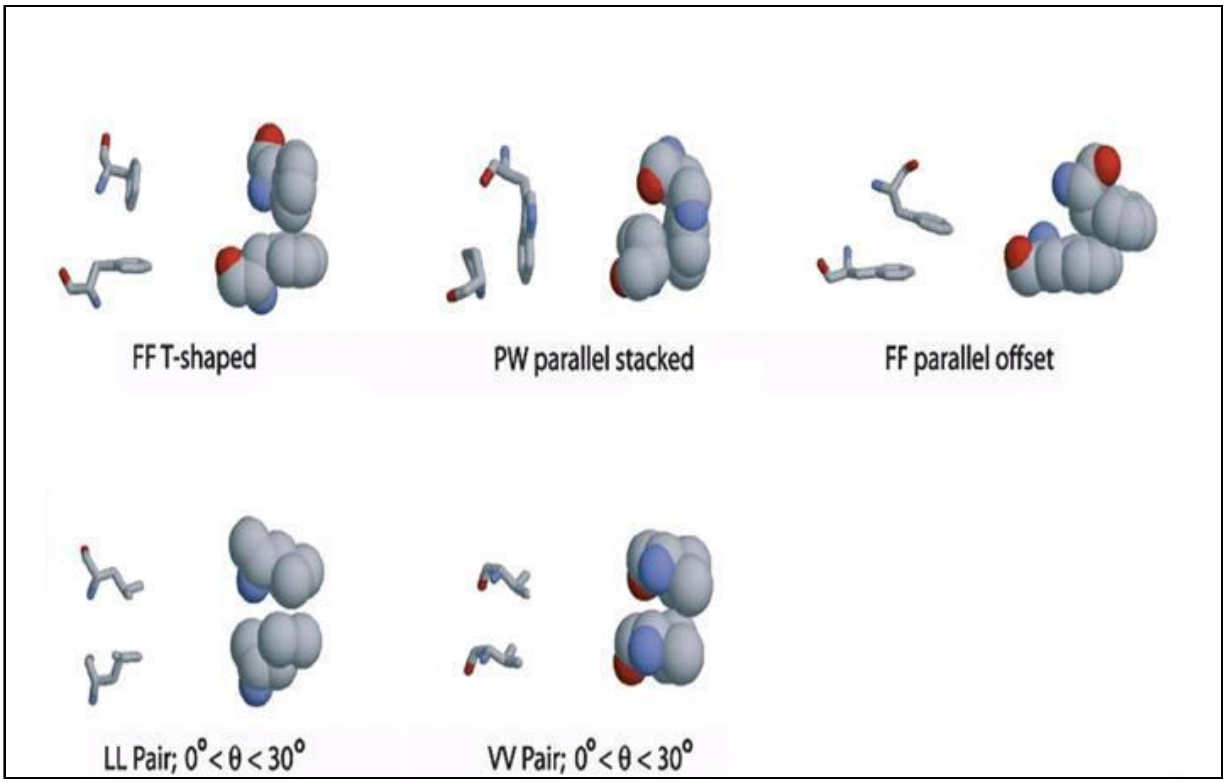


Figure 6. Prefered modes of binary side-chain packing within proteins. Ball-and-stick and space-fill representations of favorable side-chain pair orientations for Proline – Tryptophan (PW), Phenyl Alanine – Phenyl Alanine (FF), Valine – Valine (VV) and Leucine – Leucine (LL) pairs. Figure reproduced from **Misura et al., 2004.**

The inter-residue interactions sustaining a native fold could also be viewed as a network, rather than a discrete assortment of the binary interacting pairs. Several groups have viewed protein structures as contact networks **(Vendruscolo et al., 2001; Greene and Higman, 2003; Punta and Rost, 2005; Brinda and Vishveshwara, 2005; Li et al., 2007; Bagler and Sinha, 2007)** with variety of contact criteria used to define the inter-residue interactions. An elaborate review of such studies can be found in the introduction of Chapter 2.

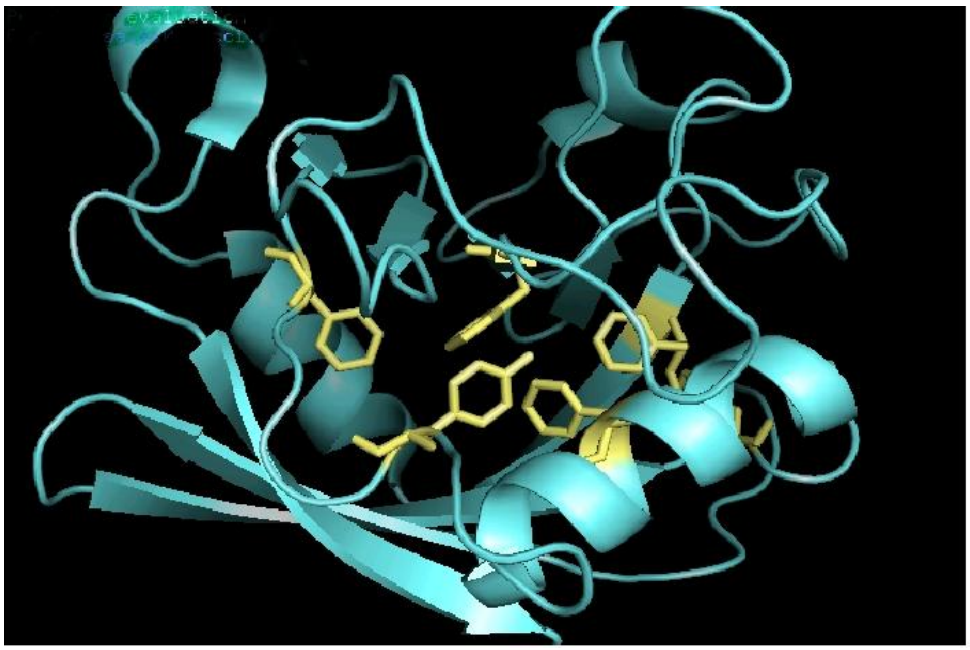


Figure 6. Multi-residue cross-talks within the protein interior. The internal archetechture viewed as a network (PDB ID: 2HAQ).

As suggested by the title, complementarity acts as a constant theme throughout the whole thesis. The first part of the thesis uses shape complementarity between interacting side-chain surfaces to extract networks (with specific geometrical constraints) and exhaustively maps their distribution in a database of high resolution native protein crystal structures (**Chapter 2**). The objective of the study was to identify recurrent packing modes (in terms of network topologies) within native protein interiors and analyze the geometrical constraints imposed on them. The second part of the thesis probes the electrostatic complementarity of residues buried within native protein interiors and compares the two (shape and electrostatic) complementarity measures (**Chapter 3**). The third part of the study designs scoring functions based on the combined use of shape and electrostatic complementarity and applies them to correctly identify the native fold amidst a set of decoys (**Chapter 4**). A novel graphical method was also developed (available as a standalone suite of programs at: http://www.saha.ac.in/biop/www/sarama.html) in order to detect local / global structural errors in experimentally or computationally derived atomic models (**Chapter 5**). Finally,

these complementarity measures have been used as filters in the computational design of the hydrophobic core of a beta-barrel protein (**Chapter 6**). The network formalism of a protein structure was also applied to study the dynamic persistence of critical inter-residue interactions and their evolutionary relationship in a given fold (**Appendix I**). Lastly, networks of ionic bonds have also been characterized and their essential geometrical and electrostatic features analyzed (**Appendix II**).

References

Arab S, Sadeghi M, Eslahchi C, Pezeshk H, Sheari A (2010). **A pairwise residue contact area-based mean force potential for discrimination of native protein structure.** *BMC Bioinformatics,* **11**: 16.

Avbelj F, Moult J, Kitson DH, James MN, Hagler AT (1990). **Molecular dynamics study of the structure and dynamics of a protein molecule in a crystalline ionic environment, Streptomyces griseus protease A.** *Biochemistry,* **29**: 8658-8676.

Axe DD, Foster NW, Fersht AR (1996). **Active barnase variants with completely random hydrophobic cores.** *Proc. Natl Acad. Sci. USA*, **93**: 5590–5594.

Bagler G, Sinha S (2007). **Assortative mixing in protein contact networks and protein folding kinetics.** *Bioinformatics.* **23**: 1760–1767.

Bahadur RP, Chakrabarti P (2009). **Discriminating the native structure from decoys using scoring functions based on the residue packing in globular proteins.** *BMC Structural Biology*, **9**: 76.

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LSL (2005). **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* **33**: D154–159.

Banerjee R, Sen M, Bhattacharyya D, Saha P (2003). **The jigsaw puzzle model:search for conformational specificity in protein interiors.** *J. Mol. Biol.* **333**: 211–226.

Beasley JR, Hecht MH (1997). **Protein design: the choice of de novo sequences.** *J. Biol. Chem.* **272**: 2031–2034.

Behe MJ, Lattman EE, Rose GD (1991). **The protein folding problem: the native fold determines packing, but does packing determine the native fold?** *Proc. Natl Acad. Sci. USA*, **88**: 4195–4199.

Benjamin B, Havranekb JJ (2011). **Automated selection of stabilizing mutations in designed and natural proteins.** *Proc Nat Acad Sci USA*, **109**: 1494–1499.

Benkert P, Tosatto SC, Schomburg D (2008). **QMEAN: A comprehensive scoring function for model quality assessment.** *Proteins*, **71**: 261-77.

Berman HM, Westbrook J, Feng Z, et al. (2000). **The protein data bank.** Nucleic Acids Res, **28**: 235–242.

Bowie JU, Eisenberg D (1994). **An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function.** *Proc Natl Acad Sci USA,* **91**: 4436–4440.

Braden BC, Poljak RJ (1995). **Structural features of the reactions between antibodies and protein antigens**. *FASEB J.* **9**: 9-16.

Bradley P, Misura KM, Baker D (2005). **Toward high-resolution de novo structure prediction for small proteins.** *Science* **309**: 1868–1871

Branden CI, Jones TA (1990). **Between objectivity and subjectivity.** *Nature* **343**: 687-689.

Brinda KV, Vishveshwara S (2005). **A network representation of the protein structures: implications for protein stability.** *Biophys. J.,* **89**: 4159-4170.

Brocchieri L, Karlin S (1994). **Geometry of interplanar residue contacts in protein structures.** *Proc. Natl Acad. Sci. USA*, **91**: 9297–9301.

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

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983). **CHARMM: a program for macromolecular energy, minimization, and dynamics calculations.** *J Comput Chem,* **4**: 187–217.

Bryant S, Lawrence C (1993). **An empirical energy function for threading protein sequence through folding motif.** *Proteins.* **16**: 92-112.

Brylinski M, Konieczny L, Roterman I (2006). **Fuzzy-oil-drop hydrophobic force field--a model to represent late-stage folding (in silico) of lysozyme.** *J. Biomol. Struct. Dyn,* **23**: 519-528.

Brylinski M, Prymula K, Jurkowski W, Kochanczyk M, Stawowczyk E, Konieczny L, Roterman I (2007). **Prediction of functional sites based on the fuzzy oil drop model.** *PLoS. Comput. Biol,* **3**: 909-923.

Buckle AM, Cramer P, Fersht AR (1996). **Structural and energetic responses to cavity creating mutations in hydrophobic cores: observation of a buried water molecule and the hydrophilic nature of such hydrophobic cavities.** *Biochemistry*, 35, 4298–4305.

Chen R, Weng Z (2003). **A novel shape complementarity scoring function for protein–protein docking.** *Proteins,* **51**: 397–408.

Chothia C (1974). **Hydrophobic bonding and the accessible surface areas in proteins.** *Nature,* **254**: 338-339.

Chothia C, Janin J (1975). **Principles of protein-protein recognition.** *Nature,* **256**: 705-708.

Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004). **ClusPro: a fully automated algorithm for protein–protein docking.** *Nucleic Acids Res,* **32**: W96-W99.

Connolly ML (1983a). **Analytical molecular surface calculation.** *J. Appl. Cryst,* **16**: 548–558.

Connolly ML (1983b). **Solvent-accessible surfaces of proteins and nucleic acids.** *Science,* **221**: 709 –713.

Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popović Z and Foldit players (2010). **Predicting protein structures with a multiplayer online game**, *Nature,* **466**: 756.

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995). **A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.** *J Am Chem Soc,* **117**: 5179–5197.

Crick FHC (1953). **The packing of a-helices: simple coiled coils.** *Acta Crystallog*. **6**: 689–697.

Dahiyat BI, Mayo SL (1997) **Probing the role of packing specificity in protein design.** *Proc. Natl. Acad. Sci. USA,* **94**: 10172-10177.

Dahiyat BI, Sarisky CA, Mayo SL (1997). **De novo protein design: towards fully automated sequence selection.** *J Mol Biol.* **273**: 789-796.

Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Michael DT, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmstrom L, Wollacott AM, Wang C, Andre I, Baker D (2007). **Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home.** *Proteins* **69**: 118–128.

Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, III, Snoeyink J, Richardson JS, Richardson DC, (2007). **MolProbity: all-atom contacts and structure validation for proteins and nucleic acids.** *Nucl. Acids. Res.* **35:** W375–W383.

Demchuck E, Mueller T, Oschkinat H, Sebald W, Wade RC (1994). **Receptor binding properties of four-helix bundle growth factors deduced from electrostatic analysis.** *ProteinSci.* **3**: 920-935.

Desjarlais JR, Handel TM (1995). **New strategies in protein design.** *Curr. Opin. Biotechnol.* **6**: 460-466.

Dill KA, Bromberg S, Yue K, Feibig KM, Yee DP, Thomas PD, Chan HS (1995). **Principles of protein folding—a perspective from simple exact models.** *Protein Sci.* **4**: 561–602.

Dill KA (1990). **Dominant forces in protein folding.** *Biochemistry*, **29**: 7133–7155.

Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA (2007). **The protein folding problem: when will it be solved?** *Curr Opin Struct Biol*, **17**: 342-346.

Edgar RC (2004). **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucl Acid Res,* **32**: 1792-1797.

Eriksson AE, Baase WA, Zhang XJ, Hienz DW, Blaber M, Baldwin EP, Matthews BW (1992). **Response of a protein structure to cavity creating mutations and its relation to the hydrophobic effect.** *Science*, **255**: 178–183.

Fisher D, Norel R, Wolfson II, Nussinov R (1993). **Surface Motifs by a Computer Vision Technique: Searches, Detection and Implications for Protein-Ligand Recognition.** *Proteins: struct. Funct. Genet,* **16**: 278-292.

Gabb HA, Jackson RM, Sternberg MJ (1997). **Modelling protein docking using shape complementarity, electrostatics and biochemical information.** *J. Mol. Biol,* **272**: 106–120.

Gassner NC, Baase WA, Matthews BW (1996). **A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme.** *Proc. Natl Acad. Sci. USA,* **93**: 12155–12158.

Getzoff ED, Tainer JA, Weiner PK, Kollman PA, Richardson JS, Richardson DC (1983). **Electrostatic recognition between superoxide and copper, zinc superoxide dismutase.** *Nature*, **306**: 287-290.

Gilson M, Sharp K, Honig B (1988). **Calculation of the Total Electrostatic Energy of a Macromolecular System: Solvation Energies, Binding Energies, and Conformational Analysis.** *Proteins: Struct. Func. Genet,* **4**: 7-18.

Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L (2003b). **ORFeus: detection of distant homology using sequence profiles and predicted secondary structure.** *Nucleic Acids Res,* **31**: 3804–3807.

Goraj K, Renard A, Martial J (1990). **Synthesis, purification and initial structural characterization of octarellin, a *de novo* polypeptide modeled on the alpha / beta barrel packing.** *Protein Eng,* **4**: 745-749.

Greene LH, Higman VA (2003). **Uncovering Networks within protein structures.** *J. Mol. Biol,* **334:**781-791.

Hagler A, Euler E, Lifson S (1974). **Energy functions for peptides and proteins I. Derivation of a consistent force field including the hydrogen bond from amide crystals.** *J Am Chem Soc* **96**: 5319–5327.

Helles G (2008). **A comparative study of the reported performance of ab initio protein structure prediction algorithms.** *J R Soc Interface* **5**: 387–396.

Hendsch ZS, Tidor B (1994). Do salt-bridges stabilize proteins? A continuum electrostatics analysis. *Protein Sci.* **3**: 211-226.

Holm L, Sander CJ (1992). **Evaluation of protein models by atomic solvation preference.** *J. Mol. Biol***, 225**: 93-105.

Hong Y, Chintapalli SV, Ko KD, Bhardwaj G, Zhang Z, Rossum Dv, Patterson RL (2011). **Predicting protein folds with fold-specific PSSM libraries.** *PLoS One,* **6**: e20557.

Hooft RWW, Sander C,, Vriend G, (1996). **Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures.** *Proteins,* **26**: 363-376.

Hu Y, Dong X, Wu A, Cao Y, Tian L, Jiang T (2011). **Incorporation of local structural preference potential improves fold recognition.** *PLoS One,* **6**: e17215.

Huang ES, Subbiah S, Tsai J, Levitt M (1996). **Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations.** *J Mol Biol,* **257**: 716-725.

Hurley JH, Baase WA, Matthews BW (1992). **Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme.** *J Mol Biol,* **224**: 1143-1159.

Janin J, Chothia C (1990). **The structure of protein-protein recognition sites.** *J. Biol. Chem.* **265**: 16027-16030.

Jones DT (1999) **GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences.** *J Mol Biol*, **287**:797–81.

Jones SJ, Thornton JM (1995). **Protein-protein interactions: a review of protein dimer structures.** *Prog. Biophys. Mol. Biol.* **63**: 31- 65.

Jorgensen WL, Tirado-Rives J (1988) **The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin.** *J Am Chem Soc,* **110**:1657–1666.

Kahraman A, Morris RJ, Laskowski RA, Thronton JM (2007). **Shape variation in protein pockets and their ligands.** *J. Mol. Biol.* **368**: 283-301.

Kahraman A, Morris RJ, Laskowski RA, Favia AD, Thronton JM (2010). **On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins.** *Proteins*. **78**: 1120-1136.

Kamtekar S, Schiffer HX, Babik JM, Hecht MH (1993). **Protein design by binary patterning of polar and non-polar amino acids.** *Science*, **262**: 1680–1685.

Kimura SR, Brower RC, Vajda S, Camacho CJ (2001). **Dynamical view of the positions of key side chains in protein–protein recognition.** *Biophys. J.*, **80**: 635–642.

Kinch L, Shi SY, Cong Q, Cheng H, Liao Y, Grishin NV (2011). **CASP9 assessment of free modeling target predictions.** *Proteins.* **79**: 59-73.

Lau KF, Dill KA (1990). **Theory of protein mutability and biogenesis.** *Proc. Natl Acad. Sci. USA,* **87**: 638–642.

Lawrence MC, Colman PM (1993). **Shape complementarity at protein/protein interfaces.** *J. Mol. Biol.* **234**: 946–950.

Lazar GA, Desjarlais JR, Handel TM (1997). ***De novo*** **design of the hydrophobic core of ubiquitin.** *Protein Sci.* **6**: 1167-1178.

Li Z, Scheraga HA (1987) **Monte Carlo-minimization approach to the multiple-minima problem in protein folding.** *Proc Natl Acad Sci USA,* **84**: 6611–6615.

Li X, Hu C, Liang J (2003). **Simplicial edge representation of protein structures and alpha contact potential with confidence measure.** *Proteins.* **53**: 792-805.

Li Z, Yang Y, Zhan J, Dai L, Zhou Y (2013). **Energy Functions in De Novo Protein Design: Current Challenges and Future Prospects.** *Annual Review of Biophysics.* **42**: 315-335.

Li J, Wang J, Wang W (2007). **Identifying folding nucleus based on residue contact networks of proteins.** *Proteins.* **71**: 1899-1907.

Lim WA, Sauer RT (1989). **Alternative packing arrangements in the hydrophobic core of lambda repressor.** *Nature*. **339**: 31-36.

Lo Conte L, Chothia C, Janin J (1999). **The atomic structure of protein-protein recognition sites.** *J Mol Biol.* **285**: 2177-2198.

Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF (2001). **Protein docking using continuum electrostatics and geometric fit.** *Protein Eng.* **14**: 105–113.

McCoy AJ, Epa CV, Colman PM. (1997). **Electrostatic complementarity at protein/protein interfaces.** *J. Mol. Biol.* **268**: 570–584.

McDonald NQ, Lapatto R, Murray-Rust J, Gunning J, Wlodawer A, Blundell TL (1991). **Newprotein fold revealed by a 2.3 Å resolution crystal structure of nerve growth factor.** *Nature*, **354**: 411-414.

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

Miao Z, Cao Y, Jiang T (2011). **RASP: Rapid modeling of protein side-chain conformations.** *Bioinformatics* **23**: 1-6.

Mirzaie M, Eslahchi C, Pezeshk H, Sadeghi M (2009). **A distance-dependent atomic knowledge-based potential and force for discrimination of native structures from decoys.** *Proteins.* **77**: 454-463.

Misura KM, Chivian D, Rohl CA, Kim DE, Baker D (2006). **Physically realistic homology models built with ROSETTA can be more accurate than their templates.** *Proc. Natl. Acad. Sci. USA*. **103**: 5361-5366.

Misura KM, Morozov AV, Baker D (2004), **Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction.** *J. Mol. Biol.* **342**: 651–664.

Mitchell, JBO, Laskowski RA, Thornton JM (1997). **Non randomness in side-chain packing : the distribution of interplanar angles.** *Proteins: Struct. Funct. Genet.* **29**: 359–376.

Miyazawa S, Jernigan RL (1996). **Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading.** *J. Mol. Biol.* **256**: 623-644.

Moult J, Fidelis K, Kryshtafovych A, Tramontano A (2011). **Critical assessment of methods of protein structure prediction (CASP)—round IX.** *Proteins.* **79**: 1–5.

Nichollos A, Honig B (1991). **A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation.** *J Comput Chem,* **12**: 435-445.

Norel R, Fischer D, Wolfson HJ, Nussinov R (1994). **Molecular Surface Recognition by a Computer Vision based Technique.** *Protein Eng,* **7**: 39-46.

Offredi F, Dubail F, Kischel P, Sarinski K, Stern AS, Van de Weerdt C, Hoch JC, Prosperi C, Francois JM, Mayo SL, Martial JA (2003). ***De novo*** **backbone and sequence design of an idealized α/β–barrel protein: evidence of stable tertiary structure.** *J. Mol. Biol.* **325**: 163-174.

Oldziej S, Czaplewski C, Liwo A, et al. (2005). **Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests.** *Proc Natl Acad Sci USA,* **102**: 7547–7552.

Park B, Levitt M (1996). **Energy functions that discriminate X-ray and near-native folds from well-constructed decoys.** *J Mol Biol,* **258**: 367-392.

Punta M, Rost B (2005). **Protein folding rates estimated from contact predictions.** *J. Mol. Biol.* **348**: 507-512.

Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D (2009). **Structure prediction for CASP8 with all-atom refinement using Rosetta**. *Proteins.* **77**: S89-99.

Richards, F. M. (1974). **The interpretation of protein structures total volume, group volume distributions and packing density.** *J. Mol. Biol.* **82**: 1–14.

Richards FM (1977). **Areas, volumes, packing and protein structures.** *Annu. Rev. Biophys. Bioeng.* **6**: 151-176.

Roberts VA, Freeman HC, Olson AJ, Tainer JA, Getzoff ED (1991). **Electrostatic orientation of the electron-transfer complex between plastocyanin and cytochrome C.** *J. Biol.Chem.* **266**: 13431-13441.

Sadowski MI, Maksimiak K, Taylor WR (2011). **Direct correlation analysis improves fold recognition**. *Comput Biol Chem* **35**: 323-332.

Sali A, Blundell TL (1993). **Comparative protein modelling by satisfaction of spatial restraints.** *J Mol Biol*, **234**: 779-815.

Samanta U, Pal D, Chakrabarti P (1999). **Packing of aromatic rings against tryptophan residues in proteins.** *Acta Crystallog. D,* **55**: 1421–1427.

Samudrala R, Moult J (1998). **An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction.** *J Mol Biol* **275**: 895–916.

Shakhnovitch EI, Finkelstein AV (1989). **Theory of cooperative transitions in protein molecules. Why denaturation of protein molecules is a first-order phase transition.** *Biopolymers*, **28**: 561–602.

Sheffler W, Baker D (2009). **RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation.** *Protein Sci.* **18**: 229–239.

Shen MY, Sali A (2006). **Statistical potential for assessment and prediction of protein structures.** *Protein Sci* **15**: 2507–2524.

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

Sikorski A, Skolnick J (1989). **Monte carlo simulation of equilibrium globular protein folding: a-helical bundles with long loops.** *Proc. Natl Acad. Sci. USA*, **86**: 2668–2672.

Simons KT, Kooperberg C, Huang E, et al. (1997). **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.** *J Mol Biol,* **268**: 209–225.

Singh J, Thornton JM (1985). **The interaction between phenylalanine rings in proteins.** FEBS Letters, **191**: 1–6.

Sippl M (1995). **Knowledge based potentials for proteins.** *Curr Opin Struct Biol*, **5**: 229-235.

Skolnick J, Kihara D, Zhang Y (2004). **Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm.** *Protein* **56**: 502–518.

Skolnick J (2006). **In quest of an empirical potential for protein structure prediction.** Curr Opin Struct Biol **16**: 166–171.

Skolnick J, Jaroszewski L, Kolinski A, et al. (1997). **Derivation, testing of pair potentials for protein folding. When is the quasichemical approximation correct?** *Protein Sci.* **6**: 676–688.

Skolnick J, Kolinski A, Ortiz A (2000). **Derivation of protein-specific pair potentials based on weak sequence fragment similarity.** *Proteins.* **38**: 3-16.

Stockwell GR, Thronton JM (2006). **Conformational diversity of ligands bound to proteins**. *J Mol Biol.* **356**: 928-944.

Tanaka T, Kuroda Y, Kimura H, Kidokoro S, Nakamura H (1994). **Cooperative deformation of a *de novo* designed protein.** *Protein Eng.* 1994, **7**: 969-976.

Thomas A, Meurisse R, Charloteaux B, Brasseur R (2002). **Aromatic side-chain interactions in proteins. I. Main structural features.** *Proteins.* 2002, **48**: 628-634.

Tovchigrechko A, Wells CA, Vakser IA (2002). **Docking of protein models.** *Protein Sci.* **11**: 1888–1896.

Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D (2003). **An improved protein decoy set for testing energy functions for protein structure prediction.** *Proteins*, **53**: 76-87.

Vendruscolo M, Pacl E, Dobson CM, Karplus M (2001). **Three key residues from a critical contact network in a protein folding transition state.** *Nature.* **409**: 642-645.

Yang Y, Faraggi E, Zhao H, Zhou Y (2011). **Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates.** *Bioinformatics.* **27**: 2076 - 2082.

Walls PH, Sternberg MJE (1992). **New algorithm to model protein-protein recognition based on surface complementarity.** *J. Mol. Biol.* **228**: 277-297.

Weiner SJ, Kollman PA, Case DA, et al. (1984). **A new force field for molecular mechanical simulation of nucleic acids and proteins.** *J Am Chem Soc.* **106**: 765–784.

Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC (1999a). **Visualizing and Quantifying Molecular Goodness-of-Fit: Small-probe Contact Dots with Explicit Hydrogen Atoms.** *J. Mol. Biol.* **285**: 1711-1733.

Word JM, Lovell SC, Richardson JS, Richardson DC (1999b). **Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation.** *J. Mol. Biol.* **285**: 1735-1747.

Wu S, Zhang Y (2008). **MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information.** *Proteins* **72**: 547–556.

Wu S, Skolnick J, Zhang Y (2007). **Ab initio modeling of small proteins by iterative TASSER simulations.** *BMC Biol,* **5**:17

Zagrovic B, Snow CD, Shirts MR, et al. (2002). **Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing.** *J Mol Biol,* **323**: 927–937.

Zhang Y, Kolinski A, Skolnick J (2003). **TOUCHSTONE II: a new approach to ab initio protein structure prediction.** Biophys J., **85**: 1145–1164.

Zhang Y, Skolnick J (2004a). **SPICKER: a clustering approach to identify near-native protein folds.** *J Comput Chem* **25**: 865–871.

Zhang Y, Skolnick J (2004b). **Automated structure prediction of weakly homologous proteins on a genomic scale.** *Proc Natl Acad Sci USA,* **101**: 7594–7599.

Zhang C, Liu S, Zhou H, Zhou Y (2004). **An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state.** *Protein Sci.* **13**: 400-411.

Zhang Y, Skolnick J (2005). **The protein structure prediction problem could be solved using the current PDB library.** *Proc Natl Acad Sci USA.* **102**: 1029–1034.

Zhang W, Liu S, Zhou Y (2008). **SP5: Improving Protein Fold Recognition by Using Torsion Angle Profiles and Profile-Based Gap Penalty Model**. *PLoS One.* 3: e2325.

Probing internal architecture of proteins using shape complementarity: exploring packing motifs and triplet cliques

1. Introduction

Dense packing of amino acid side-chains within the protein interior a signature of correctly folded proteins. Traditionally, there have been two models of protein packing: (1) the 'jigsaw puzzle' and (2) the 'nuts and bolts' model which lie on the opposite ends of the spectrum. The jigsaw puzzle model **(Crick, 1953)** postulates that the stereo specific interdigitation of amino acid side chains gives rise to densely packed protein interiors. On the other hand, the nuts and bolts model **(Bromberg and Dill, 1994)** does not require the association of side chains with specific geometry and asserts that the internal architecture of proteins arises simply due to the compaction of side chain atoms within a constrained volume. Despite dense packing of side chains, packing defects are also known to exist within molecular interiors, so most probably a single universal model might not account for all aspects internal molecular architecture. However, using a surface complementarity function, a previous report from this laboratory **(Banerjee et al., 2003)** demonstrated that binary association between two hydrophobic side chains (Leu-Leu, Leu-Phe etc), with high surface fit and maximal overlap between their corresponding residue surfaces, did indeed exhibit specific inter-residue geometry. It was thus clear that at least for a subset of contacts (with high fit and overlap) predictions of the jigsaw puzzle model were indeed valid.

One drawback of all such studies was that they treated the inter-residue interactions (sustaining a native fold) as a discrete assortment of binary interacting pairs whereas they are more accurately modeled as a network. Several attempts have been made to view protein structures as contact networks **(Plaxco et al., 1998; Vendruscolo et al., 2001; Greene and Higman, 2003; Amitai et al., 2004; Punta and Rost., 2005; Brinda and Vishveshwara., 2005; Aftabuddin and Kundu, 2007; Li et al., 2007; Bagler and Sinha., 2007)** wherein the amino acids have been designated as nodes and their mutual non-covalent interactions as edges. The character of these networks (in terms of degree distribution, clustering coefficients, characteristic pathlength etc.) exhibit variability depending on the cutoffs used to define inter-atomic contact. By and large,

most protein contact networks preserve 'small-world' character (local cohesiveness, global reach) **(Greene and Higman, 2003; Bagler and Sinha, 2007; Vendruscolo et al., 2002; Atilgan et al., 2004; Bagler and Sinha, 2005)** and display signatures of assortative mixing (preferential attachment of new nodes to pre-existing high degree nodes) **(Bagler and Sinha, 2007)**. However, degree distribution can be exponential, sigmoidal or dependent on a single exponent – as a function of the criteria used to define the atomic interactions **(Brinda and Vishveshwara, 2005)**. It has also been noted that in certain aspects protein contact networks differ significantly from other real world networks, for example in the restricted number of edges a node can have. Apart from providing insights into protein structures, these networks have been used to identify residues implicated in folding nuclei **(Li et al., 2007)** and transition states **(Vendruscolo et al., 2001)**, identifying functional residues involved in the active site **(Amitai et al., 2004)**, hubs stabilizing the packing of secondary structural elements **(Brinda and Vishveshwara, 2005)**, rationalization of the difference in protein stabilities from thermophilic / mesophilic organisms **(Brinda and Vishveshwara, 2005)** and estimation of folding rates **(Plaxco et al., 1998; Punta and Rost, 2005)**. The utility of the network view of the protein internal architecture is thus fairly well established.

The following section presents a detailed analysis of the distribution of protein contact networks including classification and characterization of different packing topologies found in the interior of globular proteins. Such an analysis led to the recognition that certain packing topologies defined as packing motifs were found preferably in proteins. A limited region of the topological space was exhaustively mapped in terms of frequently occurring packing motifs, combinations of which could lead to networks of larger sizes. It was found that indeed larger networks could be assembled out of a basis set of smaller ones. One such frequently occurring motif namely the three residue clique received special attention with regard to its composition and geometry of associating residues.

Central to pursuing the research objectives outlined above was the extension of the jigsaw puzzle model into protein contact networks. Thus protein contact networks have been defined primarily in terms of surfaces rather than distance between point atoms (although such networks have also been studied in parallel for the sake of comparison). As mentioned previously, earlier studies **(Banerjee et al., 2003)** had established quantitative measures (in terms of surface complementarity and overlap) to identify those residue pairs whose interacting side chains exhibit specific geometry. These measures have now been used to define 'surface contact networks' based only on those inter-residue interactions which severely constrain geometry and thus could play a predominant role in stabilizing a particular fold.

2. Materials and Methods

2.1. The Database:

Initially, 918 protein crystal structures were culled from the protein data bank (RCSB-PDB) **(Berman et al., 2000)** with a maximum R factor of 20%, resolution cutoff of 2.0 Å, polypeptide chain length of 75 to 500 residues and homologues were removed at 30% sequence identity or above. For oligomeric proteins the largest polypeptide chain was retained for the calculations. For atoms with multiple occupancies, those with the highest occupancy were selected and the first conformer for equal occupancies. Proteins with incomplete side chain atoms and those with missing stretches of amino acid residues were individually surveyed in RasMol **(Pembroke, 2000)**. If the missing stretch(s) or residue(s) involving incomplete side chain atoms was found to be either in the extremities (N / C terminal) of the chain or on completely exposed loop regions with no participation in interior packing, the protein was included in the database, otherwise rejected. The final database (**DB1**) consisted of 719 polypeptide chains **(**see **Supplementary Information in the CD enclosed)** of which 18.3% was all alpha, 19.8% - all beta, 32.3% - alpha/beta and 29.3% - alpha+beta. The protein class for each chain was decided by visual examination in Rasmol and a search in the SCOP database. 40 multidomain proteins were appropriately truncated and their domains allotted to the relevant class. The program

REDUCE **(Word et al., 1999)** was used to geometrically fix hydrogen atoms on the proteins prior to the calculations.

2.2. Burial ratio:

The exposure of residues to solvent (probe radius 1.4 Å) was estimated by the ratio (burial) of solvent accessible areas (SAA) **(Lee and Richards, 1971)** of the amino acid, X in the polypeptide chain to that of an identical residue located in a Gly–X–Gly peptide fragment with a fully extended conformation. Residues that were completely (0.00 <= burial ratio <= 0.05) or partially buried (0.05 < burial ratio <= 0.3)) were only considered in the analysis.

2.3. Networks construction:

As is well known every network can be represented as a graph, G = (V, E) which formally consists of a set of vertices (or nodes) V and a set of edges (or links) E between them. Trivially a graph can contain one or more standalone nodes (a node which is not connected to any other node in the graph) and a subgraph is called a component **(Harary, 2001)** of the graph provided each node is connected at least to one other node of the graph. Since, in the protein contact networks to be defined, no standalone node was considered, 'graph' and 'component' were treated synonymously. A node stands for the side chain of a particular residue, and two types of networks were defined based on surfaces and point atoms. For the case of point atoms, if any two atoms located on two different side chains were within 3.8 Å of each other, the two representative nodes were connected by a link. The number of atomic contacts between two side chains was considered to be the weight of the connecting edge. The network spanning the entire protein was constructed by exhaustively searching for contacts in the neighborhood of buried residues until no more nodes could be included in the network. Thus a protein could have more than one contact network embedded within it with no common nodes between them. The smallest networks considered had three nodes. With the exception of glycine all other residues were considered as nodes. Based on the interaction criteria defined above, 'point contact networks' are undirected.

2.4. Van der Waals surface generation:

The van der Waals surfaces for the proteins (including all hydrogen atoms) were sampled at 10 dots / $Å^2$, the atomic radii being assigned from the all atom molecular mechanics force field **(Cornell et al., 1995)**. The details of the surface generation have been discussed in an earlier study from this laboratory **(Banerjee et al., 2003)**. In case of disulphide bridges care was taken to remove the extra points due to the interpenetration of the van der Waals spheres of the covalently linked sulphur atoms. Thus, the entire surface of the polypeptide chain was sampled as an array of discrete area elements defined by their location (x, y, z) and the direction cosines (dl, dm, dn) of their normals.

2.5. Surface Complementarity:

Based on the van der Waals surface, surface complementarity (Sm) **(Lawrence and Colman, 1993)** and overlap (Ov) were defined as in a previous report from this laboratory **(Banerjee et al., 2003)**. Briefly, for a surface point (a) located on a buried side chain (referred to as a target), its nearest neighbor (b) was identified from the surface points of its surrounding residues, within a distance of 3.5 Å. Then the following expression was computed:

$$S(a,b) = \mathbf{n}_a \cdot \mathbf{n}_b \,.\exp\left(-w\,.\,{d_{ab}}^2\right) \qquad (1)$$

where $\mathbf{n}_a$ and $\mathbf{n}_b$ are two unit normal vectors corresponding to dot surface points a and b respectively, with d_{ab} the distance between them and w a scaling factor, set to 0.5. Thus for a target, a distribution of S values was obtained for all its side chain dot surface points. The surface complementarity (S_m) for a particular target was defined as the median of this distribution $\{S(a,b)\}$. The entire side chain surface of a target can be partitioned into patches based on the neighboring residues whose surface point(s) were identified as its nearest neighbors. For a specific target (A) and neighbor (B) the overlap ($Ov^{A\rightarrow B}$) between them was defined as

$$Ov^{(A \to B)} = \frac{N_{AB}}{N_A} \quad (2)$$

where N_{AB} is the number of points on the target (A) that have their nearest neighboring points on B and N_A is the total number of surface points for A. The surface complementarity for this patch involving A, B will henceforth be referred to as $S_m^{A \to B}$. Contact between any two residues (target and neighbor) can now be defined in terms of surfaces (based on S_m and Ov). Any two residues (target: A, neighbor: B) are said to 'interact' with each other when $S_m^{A \to B}$, $Ov^{A \to B}$ are greater than equal to 0.4 and 0.08 respectively. It will be noted that the measures of S_m and Ov are non–commutative, that is $S_m^{A \to B}$, $Ov^{A \to B}$ are not necessarily equal to $S_m^{B \to A}$ and $Ov^{B \to A}$. We formally define inter-residue surface 'contact' when their 'interactions' are mutually reciprocal, that is both $S_m^{A \to B}$, $S_m^{B \to A}$ and $Ov^{A \to B}$, $Ov^{B \to A}$ simultaneously satisfy the interaction criteria. For any contact $<S_m>$ and $<Ov>$ were taken to be the mean of ($S_m^{A \to B}$, $S_m^{B \to A}$) and ($Ov^{A \to B}$, $Ov^{B \to A}$) respectively. Similar to point atom contact networks a node in this case is also representative of the residue side chain (surface). Two nodes are connected by an edge when their corresponding residue surfaces are in 'contact'. Weight of such an edge was defined as $\sqrt{< S_m >^2 + < Ov >^2}$, analogous to calculating the magnitude of two mutually orthogonal vector components. Based on the definitions given above, such networks, henceforth referred to as 'surface contact networks', will also be undirected.

Two distinct types of networks have been defined and used in this study (1) All Residue Surface Contact Network (ASCN) and (2) All Residue Point Atom Contact Network (APCN). All contact networks were represented computationally in terms of one-zero adjacency matrices, (N×N, for a network of N nodes) where the matrix element a_{ij} = 1 denotes node *i* to be connected to node j and 0 otherwise. Since both types of networks were undirected, their corresponding adjacency matrices were essentially

symmetric. Based on these adjacency matrices, the following network parameters were estimated:

Degree: defined as the number of edges emanating from a node.

Strength of a node: defined as the sum of the weights of all edges of a node, *i* given by:

$$S_i = \sum_{j=1}^{N} a_{ij} . w_{ij} \tag{3}$$

where w_{ij} is the weight of the edge linking the i^{th} and the j^{th} node and the summation is over all nodes (N) of the network.

Weighted and unweighted clustering coefficients: Expressions for these coefficients are defined as follows:

Unweighted:

$$C(i) = \frac{|\{e_{jh}\}|}{{}^{k_i}C_2} \tag{4}$$

where k_i is the degree of the i^{th} node and $|\{e_{jh}\}|$ is the total number of actually existing connections among the set of nodes (taken pairwise, {j,h}) from the direct neighborhood of node *i* and ${}^{ki}C_2$ is the number of maximum possible connections within the same set **(Watts and Strogatz, 1998)**.

Weighted:

$$C_w(i) = \frac{1}{S_i\,(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2}\; a_{ij}\; a_{jh}\; a_{ih} \tag{5}$$

where the symbols have the same significance as given above and under identical conditions **(Barrat et al., 2004)**.

2.6. Cliquishness:

Clique is an induced subgraph where every node is connected to every other node. In case of an undirected graph containing a clique of n nodes, the embedded clique should contain ${}^{n}C_2$ edges. On the other hand, a complete graph will have any two nodes connected to each other. In this analysis the term 'isolated clique' refers to such complete graphs. Order (number of constituent nodes, n_c) of the maximal clique was searched progressively in all networks starting from triplets. Initially, a systematic search for all possible combinations of 3 nodes (from a network) was performed to identify the closed triplet cliques and on occurrence, n_c was set to 3. Then from the immediate neighborhood of a 3-clique, each node was sampled to test whether it satisfies the interaction criteria with all three nodes of the preexisting clique. A new node, on satisfaction of this criterion, was added to the previous clique and n_c was increased by one. The search was continued till convergence.

2.7. Deviation from random topology:

To estimate deviation from a random topology, unweighted and weighted clustering coefficients were individually averaged over all nodes in a network and were compared with the same measure obtained for random graphs of identical size. Following standard methods, first, the link density ($\boldsymbol{L_d}$) of a graph was estimated, defined as the ratio of the total number of actually existing edges in the graph and the number of maximum possible edges if it were a complete graph. Random graphs of identical size were generated by systematically calling each pair of nodes along with a random number seed and the pair was assigned a weighted connection if the random number was found to be lesser than the corresponding $\boldsymbol{L_d}$ value obtained from the original graph.

2.8. Relative Geometry of three-node packing motifs:

The methodology of Singh and Thornton **(Singh and Thornton, 1985)** was adopted to identify preferred modes of packing in terms of the specific geometry of interacting amino acid side chains. An internal right handed frame of reference was defined for all the hydrophobic residues based on their side chain atoms. Conventionally, the Z axis was taken to be normal to the principal plane defined by either the ring atoms (phenyl for Phe, Tyr and indole for Trp) for aromatic residues or a defined set of three side chain atoms for branched chain amino acids (Val, Leu, Ile) **(Figure 1).**

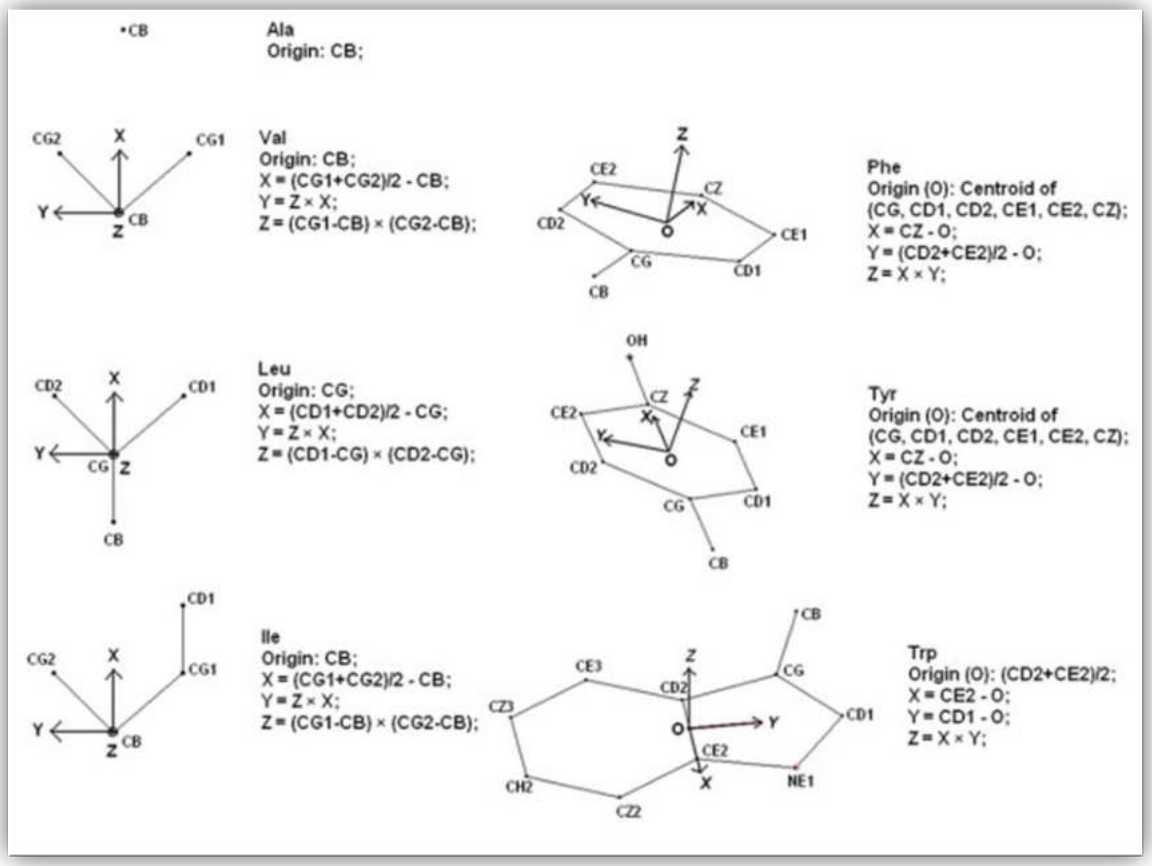


Figure 1. Internal frames defined on individual residues. Internal (right handed) frames of reference for the amino acid residues defined on the side chain atoms.

To characterize the geometry of graphs or subgraphs consisting of three nodes, a plane, $P_{triangle}$ was defined passing through the origins of the three internal frames of reference **(Figure 2)**. The resulting triangle defined by connecting the three origins was

characterized by three internal angles Ω_1, Ω_2 and Ω_3 and the lengths of the three sides of the triangle r_{12}, r_{13}, and r_{23}. A preferred right handed frame was placed at the centroid of this triangle such that the X axis (X_{tr}) points towards the origin of a preferred residue chosen according to the composition of the triplet, the Z axis (Z_{tr}) taken normal to $P_{triangle}$ and $Y_{tr} = Z_{tr} \times X_{tr}$. Three inter-planar tilt angles namely $\boldsymbol{\theta 1_t}$, $\boldsymbol{\theta 2_t}$ and $\boldsymbol{\theta 3_t}$ were then defined as angles subtended between Z_{tr} and the Z axes of the three residue-internal frames. Three additional swivel angles $\boldsymbol{\varphi 1_s}$, $\boldsymbol{\varphi 2_s}$, $\boldsymbol{\varphi 3_s}$ were further defined as those subtended by Z_p (the component of Z_{tr}, projected on residue XY planes) and the X axes of the three residue-internal frames.

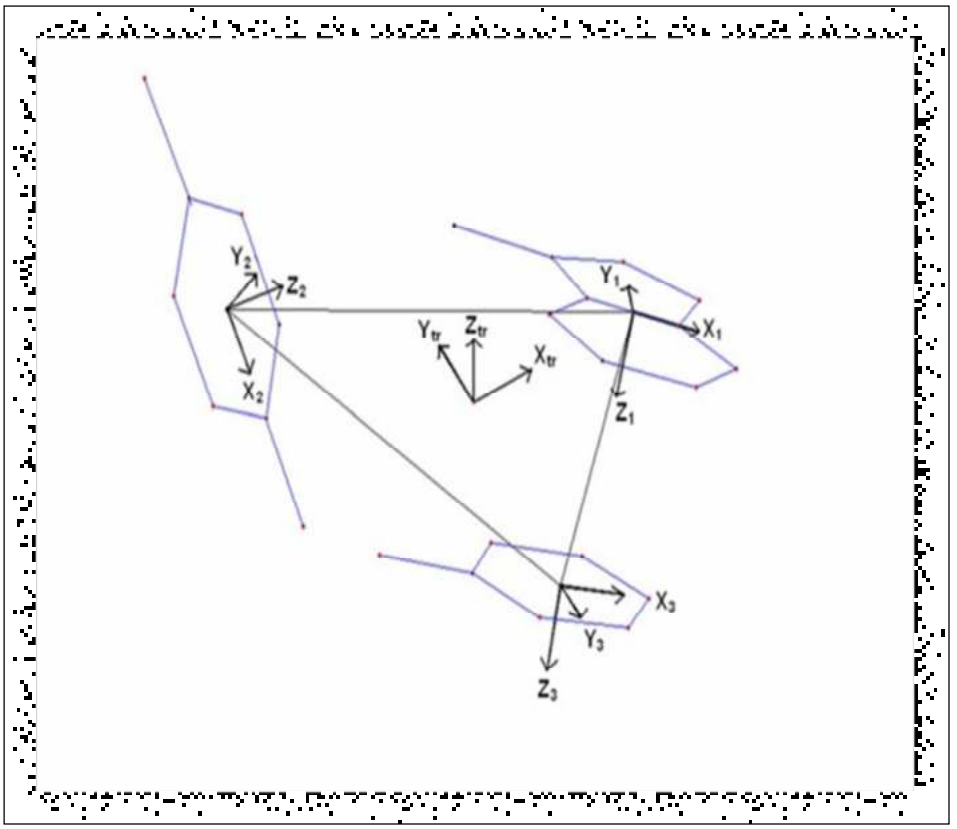


Figure 2. Global frame of reference defined on the triangle, based on a three residue clique. The triangle formed by joining the origins of the three internal frames of references ($\mathbf{X_1, Y_1, Z_1; X_2, Y_2, Z_2; X_3, Y_3, Z_3}$) defined on the residues, constituting the triplet clique. The global frame of reference ($\mathbf{X_{tr}, Y_{tr}, Z_{tr}}$) defined on the triangle, is also displayed.

The distributions of these angles in appropriate bins were analyzed for their deviation from a random distribution by means of χ^2. The distribution in the angle subtended by two randomly oriented vectors has probability density given by sin θ` dθ`/2, where θ` is the angle between the vectors **(Singh and Thornton, 1985)** whereas for two coplanar random vectors each bin should be equally populated. Thus, for a random distribution, the probability of $\theta 1_t$, $\theta 2_t$, $\theta 3_t$ falls as a function of sin θ` dθ`/2 (three-bin models for Phe and Tyr and six-bin models for Val, Leu, Ile, Trp : 30° bins) and each bin should be equally populated for $\varphi 1_s$, $\varphi 2_s$ and $\varphi 3_s$ (six-bin models for Phe, Tyr, Trp, Val, Leu, Ile : 60° bins).

2.9. Packing density:

Packing density is conventionally defined as the ratio of the volume enclosed by the van der Waals (VDW) envelope for an atom, atomic group or molecule to that of the actual volume occupied by it in space, conventionally taken to be its Voronoi volume **(Richards, 1974)** (which is the volume of a polyhedron, systematically extended around the atomic group until it comes into contact with similar polyhedra in its neighborhood). The program Voronoia.exe **(Rother et al., 2009)** was used to compute local packing densities around residues within a polypeptide chain, where solvent excluded (SE) volume **(Goede et al., 1997)** of the atomic group (defined as the space which is not accessible to any center of solvent spheres, calculated by rolling a solvent sphere of 1.4 Å probe radius over the protein surface) is calculated instead of conventional voronoi volume. Then packing density is then computed by the following ratio:

$$packing\ density = \frac{volume\,(VDW)}{volume\,(VDW) + volume\,(SE)} \qquad (6)$$

The method is considered an improvement over previous algorithms due to the fact that cavities are critically distinguished and eliminated from the actual spaces between two molecular entities and also the neighboring surfaces are cut about non planar boundaries.

3. Results and Discussion

3.1. Distribution of Networks on the basis of size

The primary object of this study was to find, characterize and classify recurring patterns in the packing of side chain atoms within a protein which sustains its native fold. In this task, those contacts were deliberately chosen which strongly and specifically condition the inter-residue geometry of association. Since the majority of atomic contacts inside a protein are contributed by side chains atoms, it is natural to represent such interior packing as a network, defined primarily in terms of fit and overlap between their corresponding van der Waals surfaces (ASCN). In addition, point atom contact networks (APCN) was also studied simultaneously (albeit with a fairly strong interaction cut off: 3.8 Å), by way of comparison.

Contact between any two surfaces can be characterized in terms of overlap (Ov) that is the extent to which two surfaces are conjoined and by their goodness-of-fit or surface complementarity (S_m) **(**see **Materials and Methods)**. A previous study from this laboratory demonstrated that when surface association between two amino acid side chains were greater than equal to 0.1 and 0.5 in Ov and S_m respectively (defined on a Connolly surface), angular distributions specifying inter-residue geometry exhibited significant deviations from a random distribution **(Banerjee et al., 2003)**. For a corresponding van der Waals surface, the values of S_m were found to be marginally lower for the same binary interactions. In contrast to point atoms, the definition of 'contact' **(**see **Materials and Methods)** between two surfaces is not necessarily mutually reciprocal (i.e. A contact B does not imply B contact A). In networks based on surface contact, nodes representing residues A and B were connected with an edge only when (1) the contact between A and B was mutually reciprocal and (2) their S_m's and Ov's both were greater than equal to 0.4 and 0.08 respectively. For strong association between two residue surfaces their contact is expected to mutually reciprocate, which also effectively simplifies the network to an undirected graph. For both point atom and surface contact networks, inter-atomic distance and surface-overlap bear a strong positive linear

correlation. S_m on the other hand appears to be an additional feature for the latter. Interestingly, the choice of 3.8 Å as the interaction cut off for point atoms appear to lead to maximum resemblance between the two categories of networks.

Distribution of networks in **DB1** on the basis of size was studied first. Networks of smaller size (3-10 nodes) dominated the distribution **(Figure 3)** with a rapid decay in frequency for larger networks (> 50 nodes). The distributions were however characterized by a long tail such that networks with greater than 200 nodes were also found, though with highly diminished frequency. The distributions for both point-atom and surface contact networks were very similar. The characteristic shape of the distribution could be adequately described by a power law ($f(x) = k.x^{-n}$, where x is the network size). The exponent, n was found to be 2.2 and 2.1 for ASCN and APCN respectively. Relaxation of the cutoffs on S_m and Ov did not appear to significantly alter the basic character of the distribution apart from decreasing the population for smaller networks thereby extending the tail for larger networks. On the other hand more stringent cut offs (Sm >= 0.5, Ov >= 0.1) led to the disintegration of the larger graphs, consequently increasing the frequency of small (3–10 nodes) and medium (11–20) sized networks with a drastic curtailment in the number of larger graphs (highest network size obtained was 49 in comparison to 223 for S_m>= 0.4, Ov >= 0.08) **(Table 1)**. Thus as has been previously observed **(Brinda and Vishveshwara, 2005)**, there appears to be a very narrow margin in terms of more stringent contact criteria which can abruptly change the spread and extent of contact networks within proteins.

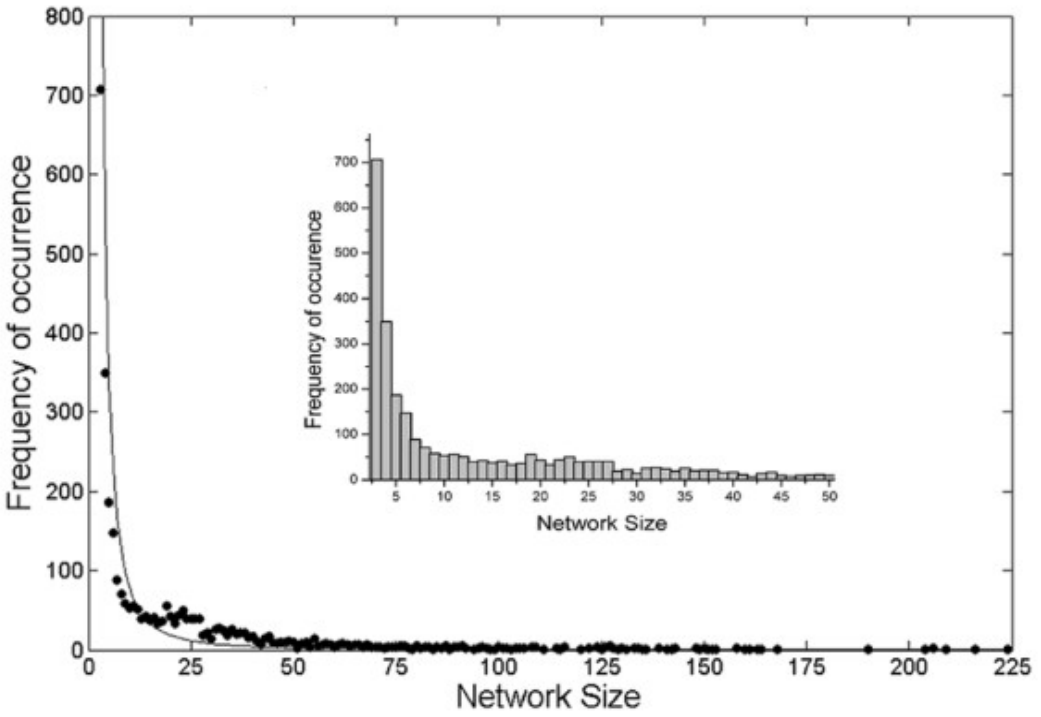


Figure 3. Distribution of surface contact networks according to size. Frequency distribution of networks of different sizes (n) for ASCN follows a power law decay (Corresponding histogram is displayed in the inset, the X axis being truncated at n = 50).

Table 1. Frequency distribution of contact networks according to size. Number of networks found in the database for different ranges of network size (i.e., number of constituent nodes). Cutoffs in surface complementarity (S_m) and overlap (Ov) respectively are given in bold within parentheses (for ASCN).

Network Size	Number of Networks		
	APCN	**ASCN (0.4, 0.08)**	**ASCN (0.5, 0.1)**
3	1168	707	1995
4	614	349	1016
5	433	187	641
6	273	147	452
7	198	90	314
8	148	71	230
9	125	58	195
10	99	53	134
11-20	564	435	476
21-30	236	341	47
31-40	130	217	8
41-50	72	105	4
51-100	165	203	-
101-150	60	63	-
151-200	33	11	-
201-250	10	6	-

The same calculations repeated for polypeptide chains distributed in bins with 75-150, 151-300, 301-500 residues gave similar curves, though for bins of larger chain length, networks of larger size appeared, thereby extending the long tail of the distribution. As expected, frequency distributions of polypeptide chains containing networks of a particular size gave a similar decaying trend with increasing network size; that is networks of smaller size were found embedded in polypeptide chains regardless of the chain length, whereas instances of larger graphs were progressively rare. These distributions tend to indicate that (in the subset of contacts where the geometry of association between residues are strongly and specifically conditioned) small (3-10 nodes) to medium (11-20 nodes) sized networks are found universally in all protein structures, whereas linkage and/or fusion of these smaller networks to form larger ones is

protein specific and is thus context dependent. Very large networks (> 150 nodes) were found only in 17 proteins almost of which had chain length exceeding 400 residues with closed packing between extended secondary structural elements (helices and sheets). Overall, the propensities for very large networks favored alpha/beta proteins.

In such protein contact networks, there is an obvious upper bound to the highest possible degree a node can have (dependent on the contact criteria) due to the limited volume of the residues involved in packing. For the present set of criteria, the highest degree of a node was found to be restricted to 8 and 9 for ASCN and APCN respectively.

It is highly likely in the context of a protein contact network, that local cohesiveness (or clustering) of side chains (or nodes) may lead to dense packing. In accordance with this idea, contact networks of all sizes were included in calculations of average unweighted (C) and weighted (Cw) clustering coefficients which gave rise to identical measures. In parallel, a statistically significant number of random graphs (of corresponding sizes) were generated **(**see **Materials and Methods)** for the direct calculation of their clustering coefficients. In a log-log plot **(Figure 4)**, average clustering coefficients of the contact networks decayed much less rapidly with increasing network size compared to corresponding random graphs. Since, this coefficient essentially determines the cliquishness of a typical neighborhood **(Watts et al., 1998)** in terms of clustering of local triplets **(Barrat et al., 2004)**, (closed) triplet cliques could be regarded as units of (non-zero) clustering. In other words, a graph of whatever size or connectivity will result in zero clustering (C=Cw=0) if there is no 'closed triplet' found embedded within it. It is to be noted that any higher order cliques could be considered as an association of nested triplet cliques. These results therefore confirm that the probability of formation of closed triplets is much higher than random within protein contact networks. Thus, these 3-cliques could be regarded as 'clustering units' and has received detailed attention in terms of geometry and composition, to be discussed in later sections.

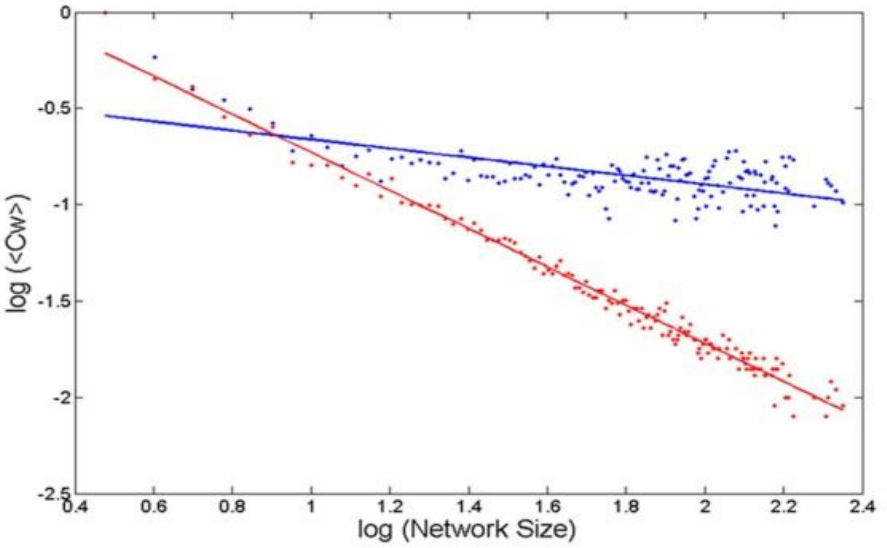


Figure 4. Protein contact networks are locally cohesive. Weighted average clustering coefficients (<Cw>) of contact networks for ASCN (blue) with their corresponding values for random networks (<Cr>: red) plotted against network size in a log-log scale.

3.2. Packing Motif:

One of the central concepts formulated in this study is that of a 'packing motif'. To start with, a packing motif is defined as a graph with a limited number of nodes (3-7), consisting of unique topological connections, which can be found either in isolation or can appear as a component or an induced subgraph, embedded within a larger graph. It follows that no two distinct motifs are super-imposable onto each other. In other words two motifs are identical (or topologically isomorphic) if there exists a one-to-one correspondence between their vertex sets which preserve adjacency. The same motif can be found in different proteins and since a node (in the motif) does not conventionally represent any particular amino acid, it could stand for different sets of residues associated with diverse inter-residue geometries in the actual three dimensional assemblies. Thus a packing motif is a reduced representation of three dimensional residue clusters, rather

analogous to super-secondary structural motifs where, for example, different combination of residues in unrelated proteins can fold into (say) a helix-turn-helix.

In order to aid numerical manipulations, each motif was uniquely represented by a linear array of numbers (motif identifier) which can be regarded as a *complete set of invariants* between any two isomorphic graphs. Initially each node of a given motif was assigned a string of numbers (of length *(d+1)* where *d* is its degree) starting with its own degree; followed by the degrees of its direct neighbors sorted in descending order. These numeric strings were collected as elements of an array and further sorted in descending order. Finally these sorted strings were concatenated, separated by a delimiter **(Figure 5)**. This identifier-string representation of each motif facilitated the computational detection, classification and clustering of motifs from **DB1**.

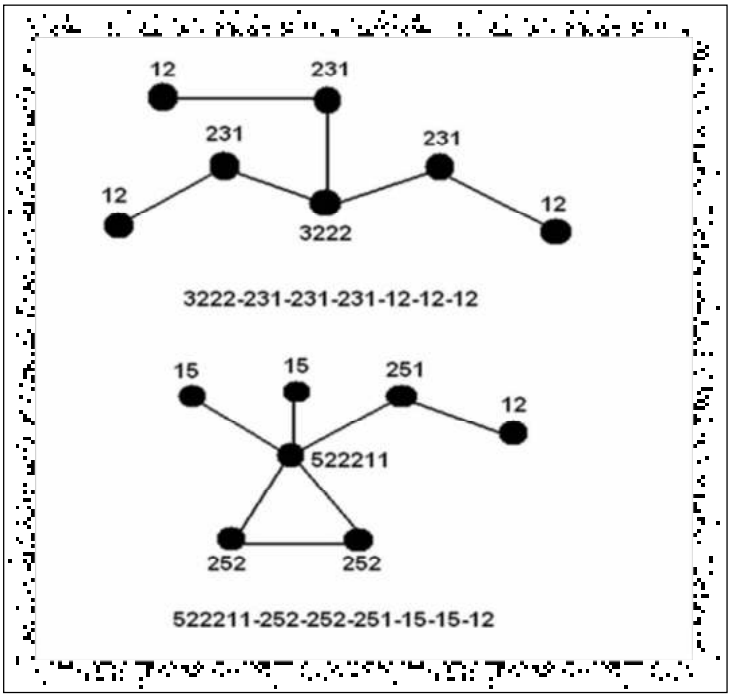


Figure 5. A novel numerical scheme to identify graphs with unique topology. Graphs (packing motifs) along with a unique number-string (motif identifier) displayed below each motif. Each node is assigned a concatenated numeric where the first digit stands for its own degree followed by degrees of its immediate neighbor sorted in a descending order.

As has been mentioned, one of the primary objectives of this study is to (1) identify recurrent motifs in smaller graphs and (2) to ascertain whether larger graphs can be constituted by an assembly of suitable motifs with appropriate topological connections. Firstly, all contact networks observed in the **DB1** were sorted according to their size (n). For smaller graphs ranging from 3-7 nodes (or may be up to 10), each set (with nodes n = 3, 4 … etc) is expected to populate a limited number of motifs **(Table 2)**.

Table 2. Frequency distributions of small (3-10 nodes) networks with their corresponding number of motifs. For a given network size, the number of networks observed in the database and the corresponding number of unique motifs have been tabulated. For example, 12 motifs were observed for 187 networks (ASCN) constituted of 5 nodes (3rd entry of the table). Cutoffs in surface complementarity (S_m) and overlap (Ov) respectively are given in bold within parentheses (for ASCN).

	APCN		**ASCN (0.4, 0.08)**	
Network Size	**Networks**	**Motifs**	**Networks**	**Motifs**
3	1168	2	707	2
4	614	5	349	5
5	433	13	187	12
6	273	37	147	28
7	198	60	90	47
8	148	76	71	55
9	125	93	58	46
10	99	91	53	51

Thus a motif could essentially be viewed as a prototype, while the actual networks observed in proteins as members belonging to a specific type of motif. To estimate the maximum number of possible motifs in networks of a given size (n), a series of random graphs were generated, conditioned by the fact that all nodes had to be connected to at least one other node in the graph **(**see **Materials and Methods)**. Since any amino acid side chain can sustain only a limited number of contacts from its surrounding

environment, it follows that there is a definite upper bound to the maximum number of edges a node can have. Therefore the highest degree (for a given network size, n) was determined from the set of actual protein contact networks, and this number was used to constrain the maximum attainable degree for the corresponding random graphs. Good agreement between the actual number of motifs observed in the database and the possible number of motifs from the random graphs (with no cutoffs on the maximum attainable degree of a node) were found for $n = 3, 4$ with rapidly increasing divergence for $n \geq 5$ **(Table 3)**. Most probably this was due to systematic over estimation in the number of unique random graphs. Thus for graphs with n ranging from 7 to 10, the number of possible motifs were recalculated by varying the maximum allowable degree from 4 (for smaller side chains) to the highest observed value in corresponding protein contact networks, which happened to be either 6 or 7 (for bulkier residues). However despite lowering the cutoff on the permissible number of edges for a node it appeared that for $n \geq 5$ a diminishing number of possible motifs is actually being realized within proteins **(Table 3)**.

Table 3. Correlation between number of (unique) motifs: observed from database versus simulated from random graphs. For a given network size (n), the number of unique motifs observed in the database is tabulated along with the corresponding number generated from simulated random graphs without and with cutoffs on the highest attainable degree.

Network Size (n)	Highest possible degree (n-1)	Observed Highest Degree		Number of Motifs		Cutoff on the highest attainable degree	Number of unique random graphs
		APCN	ASCN	APCN	ASCN		
3	2	2	2	2	2	2	2
4	3	3	3	5	5	3	6
5	4	4	4	13	12	4	22
6	5	5	5	36	25	5	114
7	6	6	5	61	45	4	315
						5	639
						6	782
8	7	5	6	76	55	4	1179
						5	4300
						6	7151
9	8	7	5	93	46	4	1410
						5	10000
						6	25864
						7	35002
10	9	7	6	91	51	4	400
						5	4512
						6	20701
						7	39654

All networks were systematically searched for size of the maximal clique (n_c) **(s**ee **Materials and Methods)** which interestingly was found to be no more than 4 for embedded cliques (n_c being 3 for a large majority of cases) and not exceeding 3 for complete graphs (or isolated cliques). In fact, the number of networks with a maximal clique of 3 and 4 nodes respectively, were found to be 1548 and 77 in case of ASCN (1662 and 146 in the same order : APCN). Since an n-clique should exactly have ($^nC_2$ – n) diagonal edges, these findings demonstrate that any possible closed-ring topology of n > 4 to be found in the database can have at the most ($^nC_2$ – n –1) diagonal edges. Thus the possible network architectures spanning the space under study is expected to be restricted to a few basic topologies namely linear chains, closed triplets (with or without branching), closed quadruplets (including embedded 4-cliques), higher order ring closures (n > 4) with a restricted number of diagonal edges and possibly a series of non-planar graphs.

For n = 3 there are trivially only two possible motifs (1) the open linear chain (motif id: 211-12-12) and (2) the isolated closed triplet clique (motif id: 222-222-222). Both possibilities are found in protein contact networks, though with considerable difference in the number of their respective occurrences. The overwhelming majority of these three-node graphs are found to be open linear chains (660: ASCN; 1070: APCN) which offer greater flexibility unlike isolated closed triplet cliques (47: ASCN; 98: APCN) which can only occur, satisfying additional geometric constraints. It could also be possible that triplet cliques once formed display an inherent tendency to evolve into larger networks given the fact that a significantly larger number of these cliques are found to be embedded as induced subgraphs in larger graphs (8876 : ASCN; 9102 : APCN) relative to isolated instances. Out of a total of 719 polypeptide chains in the database embedded triplet cliques have been found at least once in 696, 689 for ASCN and APCN respectively whereas for isolated instances the corresponding numbers are 47 (ASCN) and 90 (APCN) .

It is a relatively simple task (at least up to n = 5) to enumerate the possible number of motifs and then find their respective number of members (or the frequency of their occurrence) in the **DB1**. It is however a more complex exercise to propose a sound classification scheme, which leads to the regular ordering of actually observed motifs. To this end two additional concepts were defined namely family and path. Two motifs g(n) and g`(n+1) (with n and n+1 nodes respectively) are related by a path if the motif g`(n+1) can be formed from g(n) such that the node added to g(n) is linked to only one pre-existing node by a single edge. In other words the transformation g(n) → g`(n+1) is a path provided the newly added node has degree of one and the degree of one and only one pre-existing node (to which the new node is connected) in g(n) increases by one. Again, all motifs which can be linked by successive paths: g(n) → g`(n+1) → g``(n+2) … etc. fall within the same family. However, in case the intermediate g`(n+1) was missing, g``(n+2) was still retained in the same family. Thus, essentially a path leads to linear branching(s) about nodes belonging to a basic core topology. It follows that a motif of larger size (greater than 7 nodes) can either belong to an already existing family provided it is appropriately linked by a path or belong to an entirely new family (for example, ring closures of n>7), an occurrence which proved to be remarkably less frequent.

For n = 4, there are six possible motifs, of which five (with the sole exception of isolated quadruplet cliques or complete graphs of 4 nodes) were found to have members. Two motifs (221-221-12-12 and 3111-13-13-13) found to have the highest number of members (205, 85: ASCN and 370, 117: APCN respectively) could be related by a path to open linear chains (n = 3) or family: f1 **(Figure 6)**. A third motif (3221-232-232-13 with 52 members in ASCN and 99 in APCN) was included in the family: f2, originating from closed triplet cliques **(Figure 7)**. The remaining two (222-222-222-222 and 3322-3322-233-233) motifs were closed four membered rings **(**the latter having one diagonal edge) and were put in distinct families (f3a, f3b). Thus up to n = 4, a total of 7 motifs with a total number of 1056 (ASCN) and 1782 (APCN) members were organized into 4

families (f1, f2, f3a, f3b), with the overwhelming majority (1049: ASCN and 1754: APCN) of members incorporated into families f1 and f2.

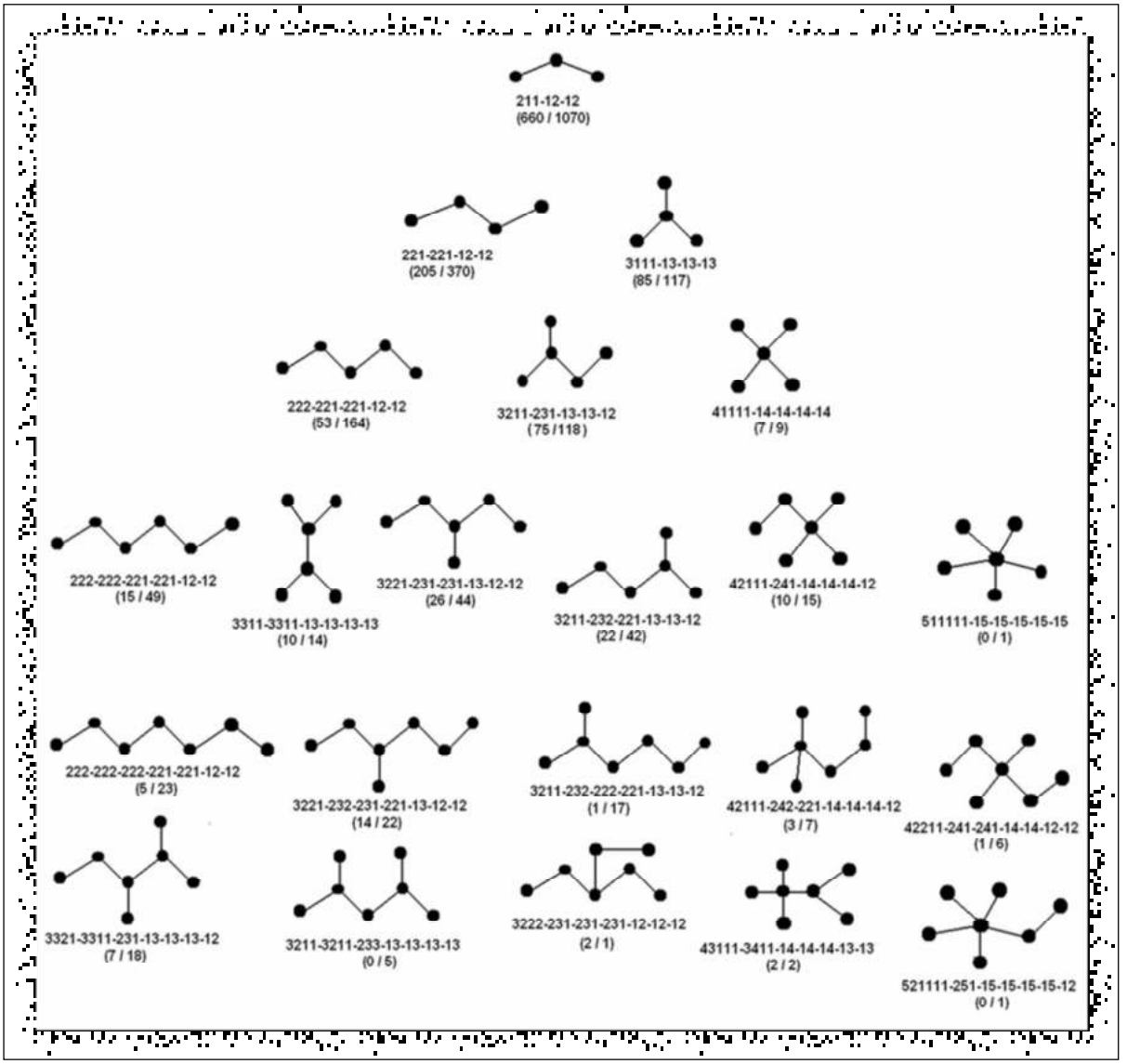


Figure 7. Motifs belonging to family f1. Network diagrams of motifs up to size 7 (nodes) belonging to family f1. The family describes topologies of minimally connected open linear chains. Motif identifier for each motif is displayed below the motif with the number of members for ASCN and APCN respectively in parentheses separated by a front slash.

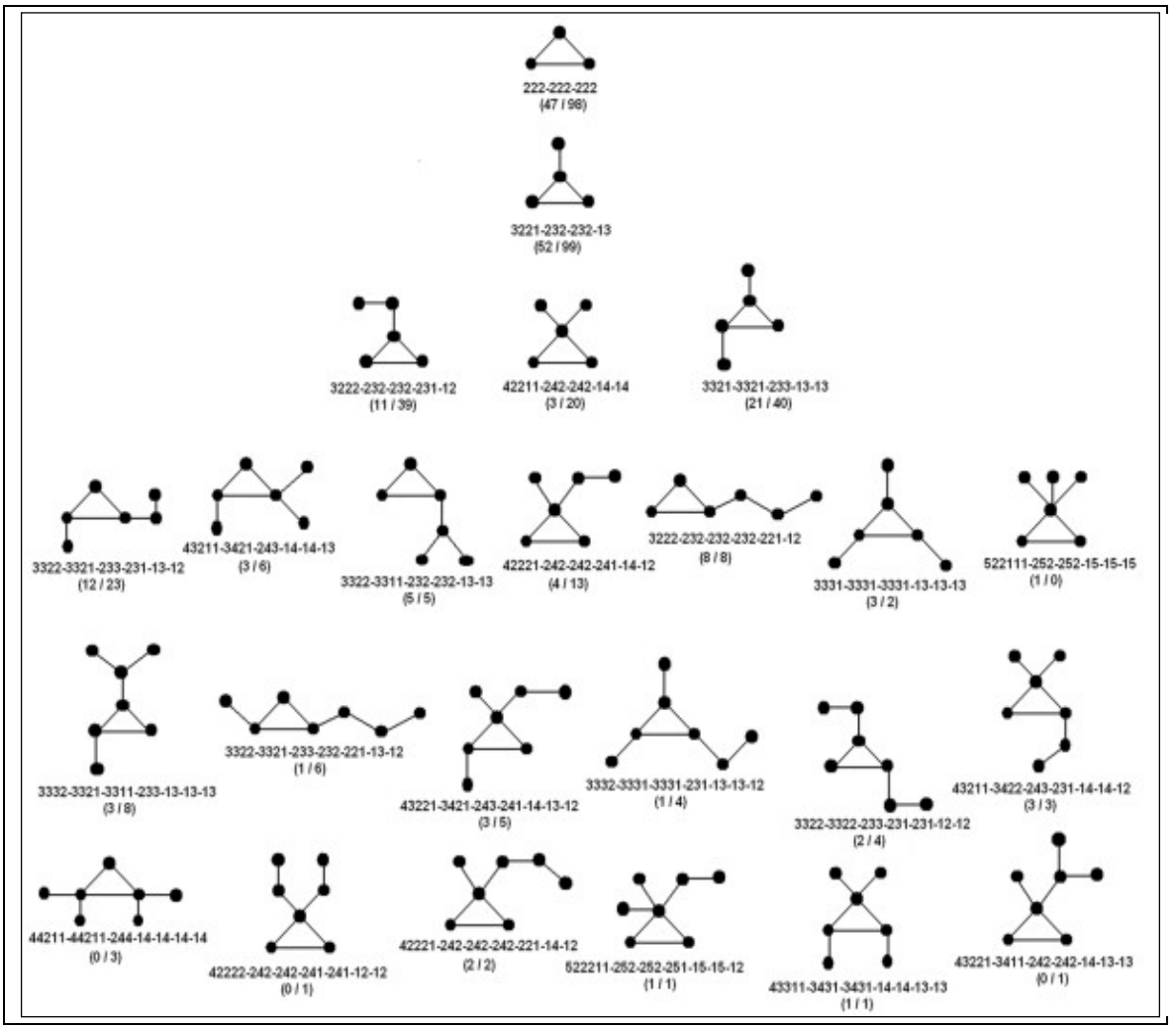


Figure 8. Motifs belonging to family f2. Network diagrams of motifs up to size 7 (nodes) belonging to family f2. The family describes topologies of triplet cliques with or without linear branching. Motif identifier for each motif is displayed below the motif with the number of members for ASCN and APCN respectively in parentheses separated by a front slash.

More or less the same trend was preserved for n = 5 where new motifs with significant number of members were again placed in the families f1 and f2. Additional motifs with marginal membership were included in f3a and f3b, which were essentially

branched four membered rings. Two more families (f4a and f4b) were created at this point, the former (f4a) originating from the closed pentagon (with no diagonals) whereas the latter (f4b) includes the pentagon with a single diagonal edge. Other families at this point include topologies demonstrated by two or more closed triplets; fused along their edges (f5), connected at a node (f6a) or connected by an edge (f6b). Once again, families other than f1 and f2 exhibited negligible memberships. Moving up levels n = 6, 7 led to the inclusion of only five more families: (a) linkage of two four membered rings through a node (f7: 1 member each in ASCN and APCN), (b) embedded quadruplet cliques with additional linear branching (f8a: 1 in ASCN and 4 in APCN) (c) non-planer graphs other than quadruplet cliques (f8b: 1 in ASCN and 3 in APCN), (d) closed six membered ring with or without diagonals edges (f4c : 3 each in ASCN and APCN) and (e) graphs where two non-adjacent nodes are connected by more than two sequences of successively connected nodes. (f8c: 4 each in ASCN and APCN). The addition of nodes from n = 5 to n = 6, 7 primarily led to the addition of motifs in the pre-existing families by, (1) increasing the length and branching of the linear chain (f1), (2) increased linear branching about the triplet cliques (f2), (3) progressive branching and inclusion of diagonal edges of the higher order closed rings (f3a, f3b, f4a, f4b, f4c, f5).

At this stage it became obvious that the initial definitions were leading to a proliferation of families with almost negligible membership. Thus to reduce the number of such families some exceptions were made. For families originating from five membered rings (f4a & f4b), motifs with a closed triplet fused about any two vertices of the pre-existing pentagon (3332-3322-3321-233-232-232-13: f4a & 43322-3432-3432-243-242-232, 533221-3532-3532-253-252-232-15, 44322-44321-3442-244-242-232-12: f4b) were included in the same respective families. Finally up to n = 7, 94 (ASCN) and 117 (APCN) motifs with 1480 (ASCN) and 2686 (APCN) members respectively were organized into 13 families. Diagrams corresponding to families other than f1 and f2 are given in the **Supplementary Information in the CD enclosed.**

The same procedure described above was performed for polypeptide chains in each individual protein class (all alpha, all beta, alpha/beta, alpha+beta), in order to investigate any preference for specific motifs or families. By and large no outstanding

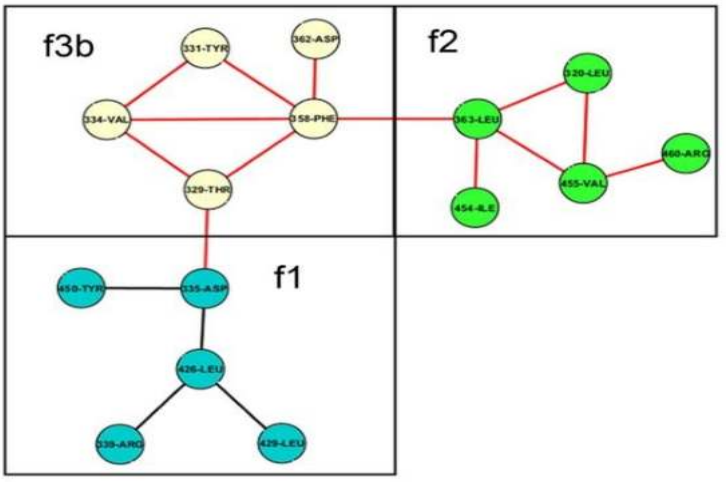
preference was observed (after suitably normalizing for the number of polypeptide chains in each class), though a somewhat reduced frequency was found for family f2 in the case of all alpha proteins. The statistics was not robust for most families barring f1 and f2 due to their extremely low frequency of occurrence.

The overall character of the distributions of motifs into families was not radically changed for different cutoffs on Sm and Ov. Even then, the application of more stringent cutoffs (Sm: 0.5, Ov: 0.1) led to an increase in the population of smaller motifs, predominantly in the f1 family. Most notable was the increase in frequency of motifs with 7 nodes (f1) probably due to the exclusion of few weaker links leading to the 'minimally connected' linear chain. Since from n = 5 to n = 6, 7 a diminishing number of families (only 5) are added with negligible membership, it is highly likely that larger networks (n > 10) will generate motifs either populating already existing families or will be assembled by joining pre-existing motifs following a defined set of rules. Since the same trend of preferential membership in the first two families were followed in networks of size n = 8, 9, 10 it was decided to begin the construction of higher order graphs out of a motif basis set obtained from networks of sizes up to n = 7, with the understanding that motifs of sizes greater than 7 nodes (located in appropriate families) would also be utilized depending on the context of a particular network. Variants of a motif with branching(s) from nodes different from those originally observed (especially with closed ring topologies) though preserving core topology would also be used in the resolution of larger graphs into subgraphs. For n = 10 it was observed that the total number of motifs became comparable to the number of networks or members **(Table 2)**. Thus, the resolution of larger graphs in terms of the proposed basis set was attempted for n greater than 10.

Generally, a graph can be resolved into either a degenerate subset of spanning subgraphs (derived by deleting edges of a graph such that the number of nodes remains conserved) and/or induced subgraphs (by deleting nodes with their corresponding edges such that two nodes adjacent in the subgraph must be adjacent in the original graph) **(Harary, 2001; Cheriyan and Maheshwari., 1988)**. Thus, analogous to a spanning subset, deleting a judiciously chosen set of specific edges of a graph should produce

independent unconnected components. Since, in this study, such isolated components are treated as graphs **(**see **Materials and Methods)** it should be possible to resolve a larger graph into a set of motifs (regarded as components) or their variants from pre-existing families, by deleting specific edges. Such edges, however, strictly exclude those being involved in a closed ring (of any size, n >=3), so that the method does not trivially produce an arbitrary combinations of motifs. On the other hand, in an induced subgraph, there exists an identical topological relation between any two corresponding nodes to that of the original graph. This one-to-one mapping serves as the basis for a computational search for motifs embedded as induced subgraphs in a larger graph. These two fundamental concepts of graph-analysis were successfully implemented to test the hypothesis whether the motif space is by and large adequate in assembling larger graphs. Contact networks for n = 15 (38: ASCN; 47: APCN) were carefully examined using Cytoscape **(Shannon et al., 2003)** and it was found that the majority of (24: ASCN; 28: APCN) networks could be resolved across one or more edges to produce isolated components which were invariably motifs belonging to pre-existing families **(Figure 9)**.

Figure 9. A contact network resolved into components. A contact network of size 15 (from 1OWL.pdb) resolved into isolated components (separated by boxes) belonging to families f1, f2 and f3b.



Other networks could not be resolved into pre-existing motifs by simply cutting across edges and in such instances the majority of possible induced subgraphs embedded in the graph were recognized as pre-existing in the motif basis set **(Figure 10)**.

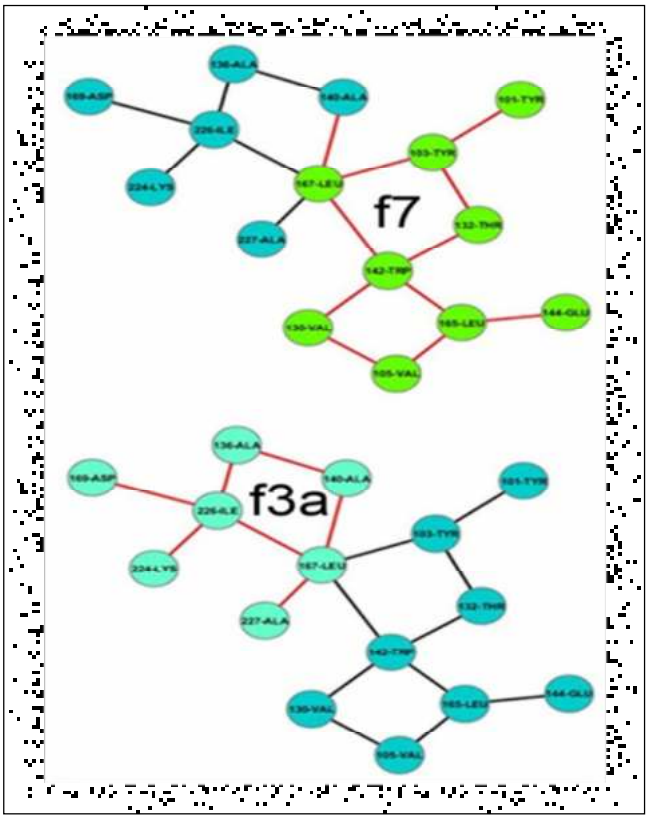


Figure 10. A contact network resolved into induced subgraphs. A contact network of size 15 (from 2HNF.pdb) resolved into induced subgraphs (highlighted by different colors) belonging to families f3a and f7.

Cases were also found where a larger graph was resolved into both components and induced subgraphs (10: ASCN, 15: APCN) **(Figure 11)**.

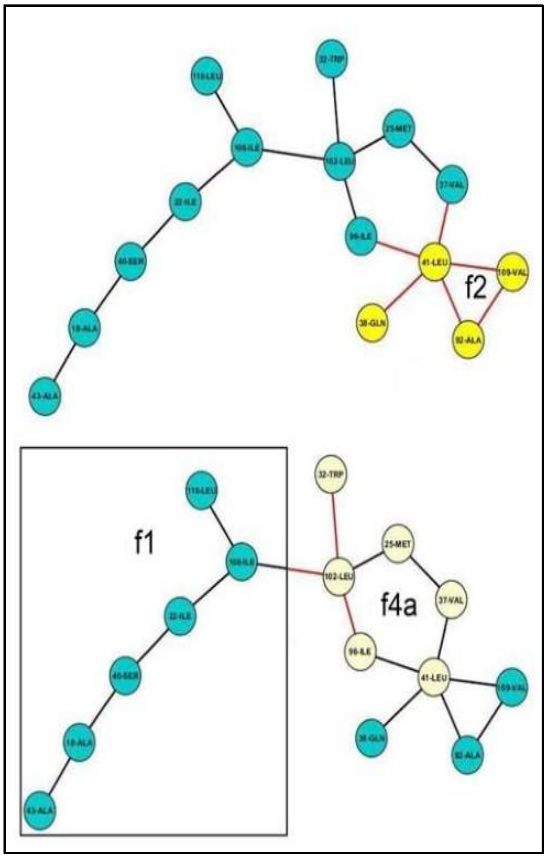


Figure 11. A contact network resolved into induced subgraphs and components. A contact network of size 15 (from 1MQV.pdb) resolved into induced subgraphs (highlighted by different colors) and components (separated by boxes) belonging to families f2, f4a and f1 respectively.

It is to be noted that there can be more than one sequence of steps to assemble a graph from degenerate sets of subgraphs following either of these procedures. As expected, for all cases, newly emerging motifs were restricted only to ring closures of greater than 6 nodes. Thus, predictably, for graphs with more than 15 nodes, new motifs should mostly be closed ring topologies with increasing number of nodes in the ring.

3.3. Triplet Clique:

The classification of motifs into families reveals that the overwhelming majority of contact networks found in protein structures occur in the first two families (f1 + f2) originating from core topologies of either open linear chains or closed triplet cliques. Although the simple rule governing the classification of motifs leads to about thirteen families in all, a significant proportion of these families have such negligible membership that they can be currently disregarded. To investigate whether the most frequently occurring motifs exhibit any preference in their constituent amino acid residues and whether their side chains pack with specific geometry, the frequently occurring closed triplet clique was chosen for further investigation. Analysis of the relative frequencies of isolated and embedded triplet cliques appeared to suggest that isolated cliques (or in other words, complete graphs of three nodes) have an inbuilt tendency for further branching(s) about the three constituent nodes resulting in their being embedded in larger graphs. Thus, to improve the statistics, both isolated triplet cliques and those embedded as induced subgraphs in larger graphs were pooled together. Further, since hydrophobic residues show greater propensity for burial and inclusion into contact networks, only the subspace of triplet cliques composed exclusively of hydrophobic residues (Ala, Val, Leu, Ile, Phe, Tyr, Trp) were considered. The resultant number of triplet cliques thus reduced to 4874, 1545 out of a total of 8923, 9200 for ASCN and APCN respectively. Interestingly, the number of such cliques was found to be significantly higher for ASCN relative to APCN, thus, results from ASCN alone are being discussed, which in any case should give superior statistics.

For a combination of three residues packed in the form of a closed triplet clique **(Figure 12)** three possibilities can be expected: (i) **C1:** all the three constituent residues are non-identical (e.g., Phe-Leu-Val) (ii) **C2:** one residue unique, the other two being identical (e.g., Ala-Ala-Leu) and (iii) **C3:** all three residues identical (Leu-Leu-Leu). Starting with the set of seven hydrophobic residues (listed above) the total number of possible combinations for each case are 35, 42 and 7 respectively. For assemblies of three identical residues (C3), the highest frequencies were observed for Leu-Leu-Leu (55.5%),

followed by Ile-Ile-Ile (~19.2%) and an almost equal proportion for Phe-Phe-Phe and Val-Val-Val (both ~12%) **(Table 4)**. A negligible fraction of triplets was found to be composed exclusively of Ala, Trp and Tyr. In all probability an assembly of three leucines provides optimal conditions, in terms of shape and size for cohesive packing. Ala and Trp represent the opposite ends of the spectrum with regard to volume and the association of tyrosines could be disfavored due to the partial charge of its terminal side chain oxygen (OH). A similar trend was observed for triplet cliques with one unique and two identical residues (C2). 40% of all triplets in this category were composed of two leucines, with X-Ile-Ile, X-Phe-Phe and X-Val-Val exhibiting frequencies 20.2%, 16.8% and 16.7% respectively. Predictably, X-Ala-Ala, X-Trp-Trp and X-Tyr-Tyr were rarely found. For hydrophobic clusters with three non-identical residues (C1) the most frequent composition is that of Ile-Leu-Val (~15.4%). It is notable that the most frequent triplet clique in this category can also be considered to be an exception as the overwhelming majority of triplets consist of at least one aromatic residue (Trp, Tyr or Phe : 79.2%). Even here, occurrence of only one aromatic in the triplet clique appears to be preferred over two, whereas cliques composed exclusively of aromatics seldom occur **(Figure 12)**. Examination of the position of the residues along the polypeptide chain showed that the contacts were mostly non-local (spatially located greater than 25 residues apart) in character.

Table 4. Triplet cliques constituted of hydrophobic residues exhibit preferences in their amino acid composition. Frequency distributions of triplet clique compositions in categories (a) C1 (all three residues different), (b) C2 (two residues identical) and (c) C3 (all three identical) are tabulated respectively.

(a)

Composition	Frequency	Composition	Frequency
ILE-LEU-VAL	322	TRP-PHE-VAL	25
PHE-ILE-LEU	276	PHE-VAL-ALA	22
PHE-LEU-VAL	246	TRP-TYR-LEU	21
TYR-ILE-LEU	151	TRP-ILE-VAL	21
PHE-ILE-VAL	150	TRP-TYR-PHE	18
TYR-PHE-LEU	117	TRP-TYR-ILE	12
TYR-LEU-VAL	98	TRP-TYR-VAL	10
TYR-PHE-ILE	85	PHE-ILE-ALA	9
TYR-PHE-VAL	77	TYR-LEU-ALA	9
TYR-ILE-VAL	69	TYR-PHE-ALA	8
TRP-PHE-LEU	56	TYR-VAL-ALA	6
LEU-VAL-ALA	50	TRP-VAL-ALA	5
TRP-ILE-LEU	46	TYR-ILE-ALA	4
TRP-LEU-VAL	41	TRP-LEU-ALA	4
TRP-PHE-ILE	33	TRP-PHE-ALA	3
ILE-LEU-ALA	32	TRP-ILE-ALA	3
ILE-VAL-ALA	30		
PHE-LEU-ALA	27	TOTAL	2086

(b)

Composition	Frequency	Composition	Frequency
ILE-LEU-LEU	291	LEU-TYR-TYR	20
VAL-LEU-LEU	268	TRP-PHE-PHE	20
PHE-LEU-LEU	237	TRP-ILE-ILE	19
LEU-ILE-ILE	187	VAL-ALA-ALA	17
LEU-PHE-PHE	162	ILE-TYR-TYR	13
VAL-ILE-ILE	134	VAL-TYR-TYR	12
LEU-VAL-VAL	128	LEU-TRP-TRP	11
ILE-VAL-VAL	119	ALA-ILE-ILE	11
ILE-PHE-PHE	105	LEU-ALA-ALA	9

TYR-LEU-LEU	104	ILE-ALA-ALA	8
PHE-ILE-ILE	94	PHE-TRP-TRP	6
PHE-VAL-VAL	88	TRP-VAL-VAL	6
VAL-PHE-PHE	67	ILE-TRP-TRP	5
TYR-PHE-PHE	53	TRP-TYR-TYR	4
TRP-LEU-LEU	47	ALA-PHE-PHE	3
TYR-ILE-ILE	46	VAL-TRP-TRP	3
TYR-VAL-VAL	35	TYR-TRP-TRP	3
ALA-VAL-VAL	31	TRP-ALA-ALA	1
PHE-TYR-TYR	29		
ALA-LEU-LEU	28	TOTAL	2434

(c)

Composition	**Frequency**	**Composition**	**Frequency**
LEU-LEU-LEU	202	ALA-ALA-ALA	4
ILE-ILE-ILE	70	TRP-TRP-TRP	2
PHE-PHE-PHE	43	TYR-TYR-TYR	1
VAL-VAL-VAL	42	TOTAL	364

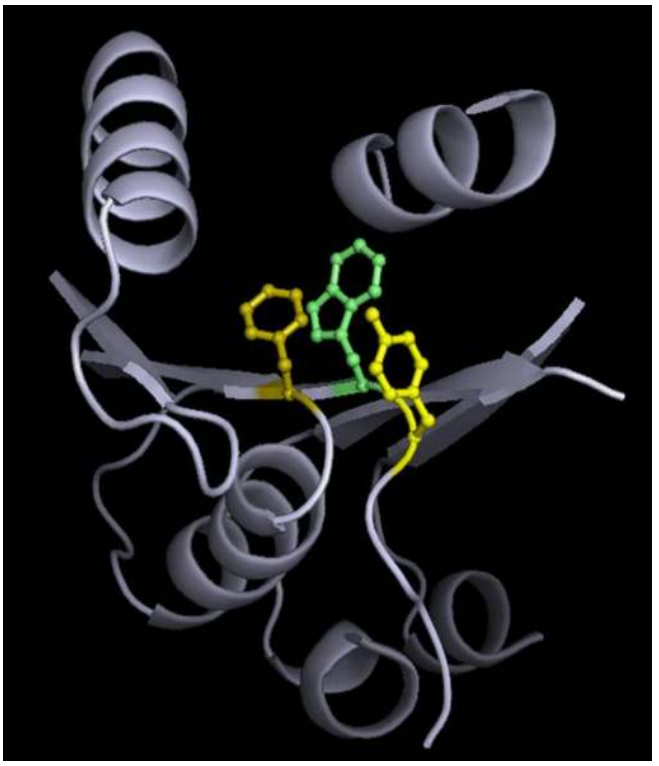


Figure 12. A three-residue clique, embedded in the protein interior. An embedded triplet clique (from 3F67.pdb) constituted of 119-Phe (Olive), 142-Trp (Lime) and 143-Tyr (Yellow) displayed as sticks in a background of broken stretches of the backbone being displayed as cartoon (Cyan).

Thus, the data indicates that even though most of the possible residue combinations are realized in local closed triplets within proteins, there is a wide divergence in their respective frequencies. Some residue combinations definitely appear to be preferred over others. Moreover, since only a subspace has been studied, the compositional propensities appear to be fairly pronounced, rather than outstanding. Without the use of surfaces and careful classification of triplet cliques (based on their compositions) these could well be overlooked. Even then, the formation of well packed three residue cliques in proteins appears to be constrained in terms of the total volume occupied by the triplet and probably their inter-residue geometry. The question then is what are the geometrical constraints imposed on these three-residue cliques?

Extending the methodology established by Singh and Thornton **(Singh and Thornton, 1985)** (to study inter-residue geometry between two amino acids) an internal right-handed Cartesian frame of reference was defined for each of the three residues constituting the triplet clique. Connecting the origins of the three internal frames of reference constructs a triangle, which can be considered to be a reduced geometric representation of the assembly. A global frame of reference was then defined on the triangle plane (see **Materials and Methods**), the global Z axis being the normal to the plane and the origin being set at the centroid of the triangle. However, there can always be two degenerate directions of the normal (Z axis). Therefore in order to secure uniformity among all the reference frames the following conventions were adopted:

1. For **C1** (all three residues different), the three residues (R1, R2, R3) were first sorted on the basis of their side chain volume R1 > R2 > R3. Let the vector directed from the origin of R1 to R2 be **v1** and that from R1 to R3 be **v2**. Then the global Z axis was defined as **v1** × **v2** and the global X axis as the unit vector directed from the global origin towards the origin of R1.

2. For a given composition in **C2** (e.g., Leu–Phe–Phe) one specific example was arbitrarily chosen whose unique residue was designated R1 and R2, R3 were assigned such that the identical procedure outlined above (in procedure 1) resulted in an acute angle being subtended between the global Z and the internal Z of R1. All other triangles with the same composition were superposed onto this template. The calculations were repeated starting from different templates to confirm that the results were not artifacts of this geometrical procedure.

3. In case of **C3** (e.g., Leu–Leu–Leu), a randomly chosen triplet was arbitrarily assigned R1, R2, R3 and the global frame was defined following procedure 1. All other triangles of the same composition were superposed onto this template. To select for the

best possible superposition in each case 6 combinatorial possibilities were checked. Similar to C2, the calculation was repeated for different starting templates.

For almost all the compositions the lengths of the triangular edges and the internal angles were severely constrained, with standard deviations ranging from ~ 0.5 - 0.6 Å and ~ 5 - 10° in lengths and angles respectively. In almost all the cases, the triangle approximates to being equilateral with the average of all the three sides lying between 5-6 Å and angles close to 60° (± 10°). Inclusion of bulky residues in the triplet cliques (Tyr-Phe-Leu), (Tyr-Phe-Ile) did not appear to significantly alter the overall trends observed in these triangular parameters. The longest average lengths were observed for Ile-Leu-Leu (6.3 ± 0.6), Leu-Ile-Ile (6.4 ± 0.5), Ile-Val-Val (6.3 ± 0.5) and Ile-Phe-Phe (6.3 ± 0.6) **(Table 5)**.

Table 5. The triangle constructed from the associated residues in a triplet clique approximates to being equilateral. Average lengths of sides (Å) and internal angles (°) (along with their standard deviations in parentheses) of the triangle formed by joining the origins of the three internal frames corresponding to the constituent residues in a triplet clique are tabulated. Only those compositions have been given whose frequencies are greater than equal to 50.

Composition			Frequency	$\langle r_{12} \rangle$	$\langle r_{13} \rangle$	$\langle r_{23} \rangle$	$\langle \Omega_1 \rangle$	$\langle \Omega_2 \rangle$	$\langle \Omega_3 \rangle$
ILE	LEU	VAL	322	5.9 (0.7)	5.9 (0.7)	5.6 (0.5)	56.5 (7.5)	61.6 (9.2)	61.9 (8.6)
ILE	LEU	LEU	291	6.3 (0.6)	5.6 (0.6)	5.5 (0.6)	55.0 (7.9)	55.9 (5.9)	69.1 (7.1)
PHE	ILE	LEU	276	5.8 (0.8)	5.4 (0.6)	5.9 (0.6)	63.2 (9.5)	55.5 (9.3)	61.3 (10.9)
VAL	LEU	LEU	268	5.4 (0.5)	5.9 (0.4)	5.6 (0.6)	59.2(7.8)	65.2 (5.6)	55.5 (5.4)
PHE	LEU	VAL	246	5.5 (0.6)	5.3 (0.6)	5.6 (0.6)	61.6 (8.2)	57.7 (9.4)	60.7 (9.4)
PHE	LEU	LEU	237	5.1 (0.5)	5.8 (0.5)	5.6 (0.5)	62.1 (8.4)	65.7 (7.6)	52.1 (5.7)
LEU	LEU	LEU	202	5.2 (0.5)	6.1 (0.3)	5.7 (0.4)	59.5 (4.2)	67.8 (4.9)	52.6 (4.7)
LEU	ILE	ILE	187	6.4 (0.5)	5.7 (0.6)	6.3 (0.8)	63.4 (11.7)	52.8 (7.0)	63.8 (7.6)
LEU	PHE	PHE	162	5.1 (0.5)	5.8 (0.5)	5.5 (0.5)	60.7 (9.3)	66.1 (7.0)	53.3 (5.9)
TYR	ILE	LEU	151	5.7 (0.8)	5.3 (0.6)	6.0 (0.6)	65.5 (8.8)	53.9 (9.5)	60.5 (11.1)
PHE	ILE	VAL	150	5.7 (0.7)	5.4 (0.7)	5.9 (0.7)	64.9 (10.6)	55.6 (10.7)	59.5 (10.3)
VAL	ILE	ILE	134	5.5 (0.6)	6.3 (0.6)	6.2 (0.8)	63.1 (11.1)	64.9 (7.5)	52.0 (7.2)
LEU	VAL	VAL	128	5.9 (0.4)	5.3 (0.5)	5.7 (0.5)	61.1 (8.0)	54.8 (4.7)	64.1 (6.1)
ILE	VAL	VAL	119	6.3 (0.5)	5.5 (0.6)	5.7 (0.5)	57.8 (8.1)	54.4 (6.4)	67.8 (5.9)
TYR	PHE	LEU	117	5.5 (0.6)	5.4 (0.6)	5.4 (0.5)	59.6 (9.0)	59.6 (10.1)	60.7 (9.5)
ILE	PHE	PHE	105	6.3 (0.6)	5.5 (0.7)	5.5 (0.6)	54.6 (8.5)	54.9 (7.3)	70.5 (7.2)
TYR	LEU	LEU	104	4.9 (0.4)	5.9 (0.5)	5.5 (0.6)	60.9 (8.5)	67.9 (7.1)	51.2 (6.3)
TYR	LEU	VAL	98	5.4 (0.6)	5.4 (0.7)	5.5 (0.4)	60.7 (7.6)	59.4 (10.1)	59.9 (10.4)
PHE	ILE	ILE	94	5.5 (0.6)	6.3 (0.7)	6.4 (0.9)	66.0 (11.3)	63.1 (8.5)	50.8 (6.7)
PHE	VAL	VAL	88	4.9 (0.6)	5.8 (0.5)	5.7 (0.5)	63.5 (10.0)	65.7 (7.7)	50.8 (6.8)
TYR	PHE	ILE	85	5.6 (0.5)	5.7 (0.9)	5.8 (0.8)	61.8 (12.1)	59.9 (13.3)	58.2 (9.9)
TYR	PHE	VAL	77	5.5 (0.6)	5.5 (0.6)	5.5 (0.7)	59.9 (10.7)	59.3 (10.0)	60.7 (9.9)
ILE	ILE	ILE	70	5.5 (0.6)	6.4 (0.5)	7.0 (0.6)	72.1 (6.2)	59.9 (4.5)	48.0 (5.8)
TYR	ILE	VAL	69	5.8 (0.9)	5.4 (0.6)	5.9 (0.7)	63.4 (10.6)	55.0 (10.3)	61.5 (11.7)
VAL	PHE	PHE	67	5.8 (0.5)	5.0 (0.5)	5.4 (0.5)	59.8 (9.0)	52.3 (6.9)	67.8 (8.0)
TRP	PHE	LEU	56	5.5 (0.6)	5.6 (0.6)	5.4 (0.6)	58.7 (8.5)	61.6 (10.3)	59.7 (9.3)
TYR	PHE	PHE	53	5.9 (0.5)	5.1 (0.4)	5.5 (0.5)	59.8 (9.2)	52.6 (5.7)	67.5 (8.1)
LEU	VAL	ALA	50	5.7 (0.5)	4.7 (0.4)	4.7 (0.4)	53.4 (5.3)	53.1 (6.8)	73.5 (7.7)

Relative geometries of the three constituent residues from the perspective of the abstract triangle defined above were analyzed by means of two more angles, namely tilt and swivel. Dot product of the global Z axis defined on the triangle plane with Z axes (Z_1, Z_2, Z_3) of the internal frames of the three residues defines the tilt angle (θ_t). It essentially describes the orientation of the residue (principal) plane **(**see **Materials and Methods)** with respect to the triangle plane **(Figure 13)**. As is well known, the angular distribution of two randomly oriented vectors should fall of as a function of sin θ` dθ`/2 (where θ` is the angle between the two vectors) **(Singh and Thornton, 1985)** and the deviation of an actual observed distribution from one which is random can be estimated by means of χ^2. Examination of χ^2 of θ_t shows that for triplet cliques composed of at least one aromatic, their corresponding tilt angles ($\theta 1_t$) exhibit significant deviation from randomness. Compositions such as Phe-Leu-Leu (χ^2 ($\theta 1_t$) = 72.1), Phe-Leu-Val (60.0), Phe-Ile-Leu (48.5), Tyr-Leu-Leu (46.7), Tyr-Phe-Leu (44.6), Tyr-Ile-Leu (39.0), Phe-Ile-Val (37.7), Tyr-Leu-Val (35.5) etc **(Table 6)** indicates a preferred orientation of the aromatic ring plane with respect to the global triangle plane.

Table 6. χ^2 for angular variables for triplet clique compositions exhibiting specific geometry. χ^2 of tilt angles ($\theta 1_t$, $\theta 2_t$, $\theta 3_t$) and swivel angles ($\varphi 1_s$, $\varphi 2_s$, $\varphi 3_s$) of residues constituting the clique (where 1,2,3 corresponds to the same sequence of residues given in the table e.g., PHE → 1, ILE → 2, LEU → 3 for the first entry) for compositions showing significant deviation in $\theta 1_t$ from a random distribution. $\chi^2_{0.05}$ for three-bin and six-bin models are 5.991 and 11.071, respectively**.** Compositions which have a predicted frequency of less than 5 for any particular angular bin, assuming a random distribution are marked with an asterisk (*). This minimal number (of data points) is 37 for a 3-bin and 74 for a 6-bin model for tilt (θ_t) angles and 30 for a 6-bin model for swivel (φ_s) angles.

Composition			Frequency	$\chi^2(\theta 1_t)$	$\chi^2(\theta 2_t)$	$\chi^2(\theta 3_t)$	$\chi^2(\varphi 1_s)$	$\chi^2(\varphi 2_s)$	$\chi^2(\varphi 3_s)$
PHE	ILE	LEU	276	48.5	11.2	35.8	15.3	14.5	5.8
PHE	LEU	VAL	246	60.0	20.6	5.2	14.0	8.5	6.8
PHE	LEU	LEU	237	72.1	25.5	13.5	13.9	20.8	3.8
TYR	ILE	LEU	151	39.0	2.9	7.3	20.5	16.0	10.0
PHE	ILE	VAL	150	37.7	9.3	16.7	19.4	4.9	19.5
TYR	PHE	LEU	117	44.6	6.2	2.8	18.5	5.7	10.6
TYR	LEU	LEU	104	46.7	15.2	8.6	6.4	5.3	16.4
TYR	LEU	VAL	98	35.5	8.7	9.5	7.4	3.1	6.6
PHE	VAL	VAL	88	21.6	24.9	5.1	14.1	4.2	4.5
TYR	PHE	VAL	77	20.3	3.0	4.0	9.0	10.2	2.0
TYR	ILE	VAL	69	23.4	2.5	3.8	4.5	2.0	4.7
TRP	LEU	LEU	47	29.5*	4.0*	5.1*	3.9	17.7	2.9
TRP	LEU	VAL	41	23.9*	5.3*	3.0*	6.3	6.3	3.9

The actual distribution of the angles (θ_t) shows the angular bins 60-90°, 90-120° to be preferentially populated (with respect to a random distribution) in contrast to ranges 0-30°, 150-180°, 30-60°, 120-150°, which exhibit a corresponding depletion **(Table 7)**.

Table 7. Distribution in $\theta 1_t$ for triplet clique compositions exhibiting high χ^2. Angular distribution of $\theta 1_t$ in different angular bins (3-bin models for Phe and Tyr and 6-bin model for Trp: 30° bins) for compositions that have shown significant deviations from a random distribution:

Composition			$\chi^2(\theta 1_t)$	% Occupancy in bins with θ (deg.) range					
				0-30	**30-60**	**60-90**	**90-120**	**120-150**	**150-180**
Random (6bin):				6.7	18.3	25.0	25.0	18.3	6.7
Random (3bin):				13.4	36.6	50.0	-	-	-
PHE	ILE	LEU	48.5	4.0	26.1	69.9	-	-	-
PHE	LEU	VAL	60.0	4.5	21.1	74.4	-	-	-
PHE	LEU	LEU	72.1	2.1	21.1	76.8	-	-	-
TYR	ILE	LEU	39.0	2.0	23.8	74.2	-	-	-
PHE	ILE	VAL	37.7	4.0	21.3	74.7	-	-	-
TYR	PHE	LEU	44.6	1.7	17.9	80.4	-	-	-
TYR	LEU	LEU	46.7	0.0	17.3	82.7	-	-	-
TYR	LEU	VAL	35.5	2.0	18.3	79.7	-	-	-
PHE	VAL	VAL	21.6	0.0	28.4	71.6	-	-	-
TYR	PHE	VAL	20.3	3.9	20.8	75.3	-	-	-
TYR	ILE	VAL	23.4	1.4	20.3	78.3	-	-	-
TRP	LEU	LEU	29.5	0.0	2.1	44.7	44.7	6.4	2.1
TRP	LEU	VAL	23.9	2.4	7.4	43.9	43.9	2.4	0

Thus, both the angular distribution and visual inspection of the triplet cliques indicate that for bulky aromatics, their normals (to the residue plane) tend to be perpendicular to the global Z axis, as if the side-chain tends to enclose the volume demarcated by the edges of the triangle, rather than penetrating into its perimeter **(Figure 2)**. The other residues (Ile, Leu, Val) however did not exhibit any consistent specificity in their tilt.

In order to investigate the rotation of the residue planes (XY plane of the residue-internal frames) about an axis parallel to their own internal Z, the component of the global Z axis

of the triangle was projected onto the respective planes and the orientation of this vector (Zp) with respect to the internal X axis (defined as the swivel angle φ_s ranging from 0-360°) was computed. Since the angle φ_s is restricted to a plane, each quadrant is expected to be equally populated for a random distribution. χ^2 in φ_s did not appear to show any significant preferences for any residue. Therefore, for a given tilt, the residue plane can adopt multiple orientations about an axis perpendicular to it.

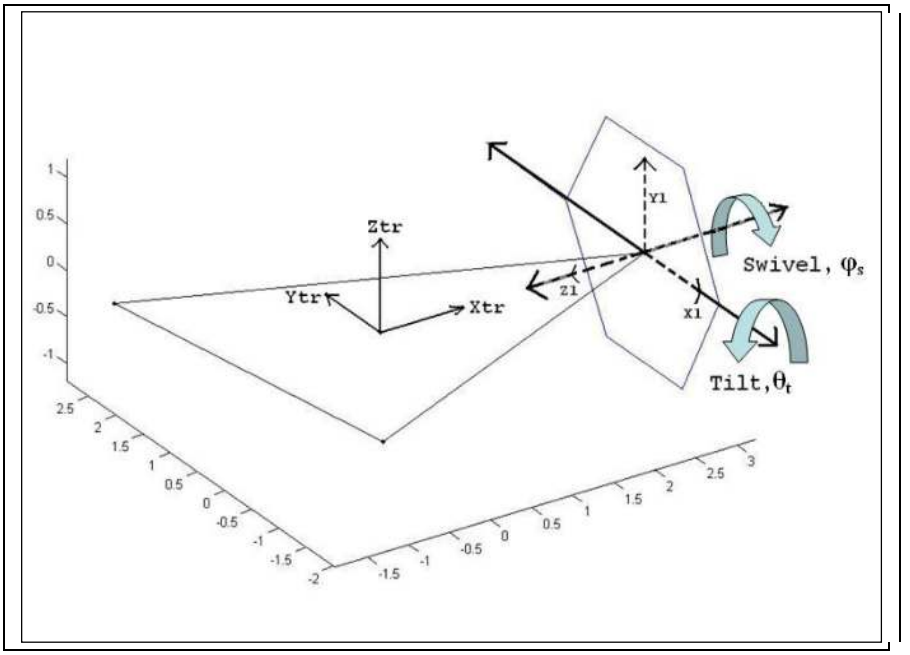


Figure 13. Tilt and Swivel angles: The tilt angle defined to be the relative orientation of the residue principal plane with respect to the triangle plane whereas the swivel angle is the rotation of the residue principal plane about an axis perpendicular to it.

3.4. Packing Density:

Investigations were also carried out to quantify local packing densities (see **Materials and Methods**, section: Packing density) in and around triplet cliques and also in their absence. Plots of packing density (f(x)) versus burial ratio (x) (see **Materials and Methods**, section: Burial ratio) exhibited an almost identical correlation for all the residue types **(Figure 14)**, decaying as a cubic polynomial

($f(x) = a.x^3 + b.x^2 + c.x + d$), demonstrating loose packing with higher exposure to the solvent.

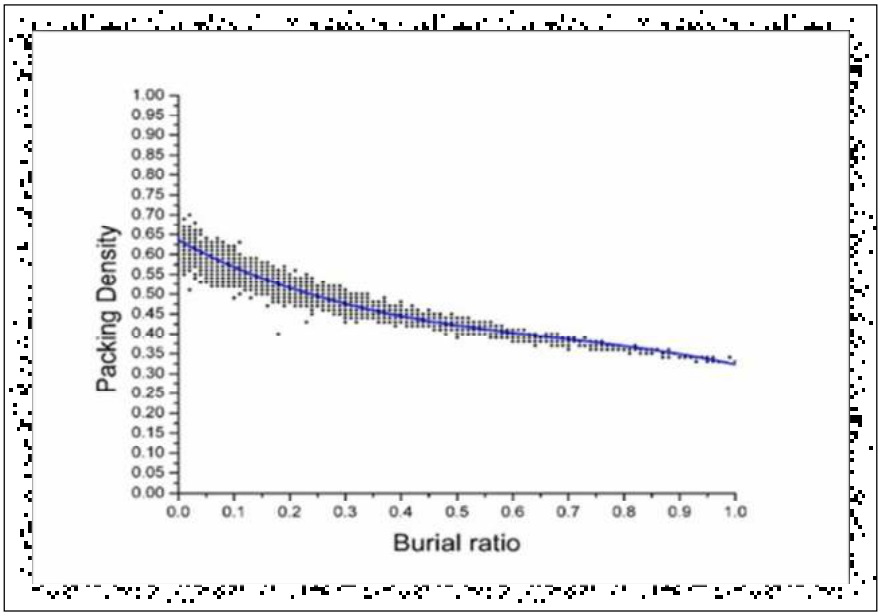


Figure 14. Packing density as a function of burial ratio. Packing density decays with increasing burial ratio (which is an index of the exposure to the solvent) following a cubic polynomial (plotted for tyrosine).

Networks were distributed into two major categories, those with triplet cliques and those devoid of them. The former were further subcategorized into the set of clique-nodes alone and that of the other non–clique members. It was evident from the results that the clique-nodes are predominantly completely buried (burial ratio <= 0.05) and thus on an average, more tightly packed than the other non-clique members whose average exposure to the solvent was consistently found to be higher **(Figure 15)**. These regions of high local packing densities occur at or near the cliques with gradual decrease at the periphery. On the other hand, networks devoid of triplet cliques are on an average less tightly packed as a consequence of higher exposure to solvent.

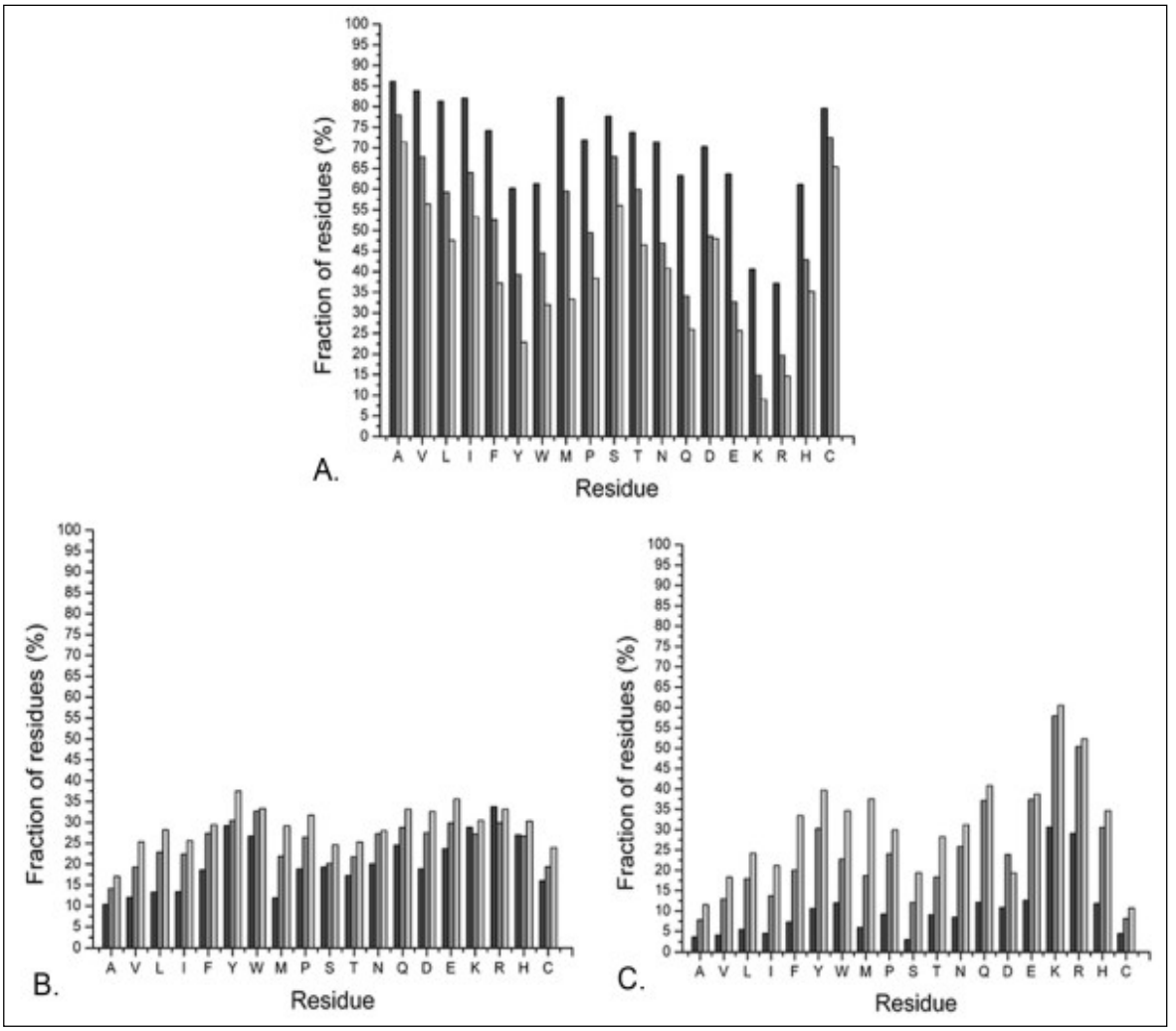


Figure 15. Nodes of a clique exhibit greater propensity to get completely buried. Percentage fraction of individual residues categorized into clique-nodes (dark gray), non-clique nodes of clique containing networks (gray) and nodes of networks devoid of cliques (light gray) sorted according to their exposure to solvent in terms of burial ratio: (**A**). for burial ratio < = 0.05 (completely buried), **(B)**. for 0.05 < burial ratio < = 0.15 (partially buried with lower exposure to the solvent), **(C)**. for 0.15 < burial ratio < = 0.3 (partially buried with higher exposure to the solvent).

4. Conclusion

The chapter is based on the confluence of two related though distinct ideas, (1) some network topologies are preferred within protein interiors, leading to the concept of packing motifs (2) the 'jigsaw puzzle' model can be successfully extended into the domain of protein contact networks. The implementation of both these ideas depends partly on representing the internal architecture of proteins in terms of surfaces rather than point atoms. A previous study from this laboratory provided simple well defined criteria to identify those contacts which definitely constrain inter-residue geometry of the associating amino acid side chains **(Banerjee et al., 2003)**. Networks based on surface contacts (with appropriate cut offs on S_m and Ov) is in effect a straightforward extension of the jigsaw puzzle model. In the search for compositional or geometrical bias, surface contact networks appear to be indispensable. In particular, triplet cliques composed exclusively of hydrophobic residues had a frequency 3 fold higher in ASCN than APCN starting from a comparable (total) number of triplet cliques. Furthermore, compositional preferences along with strong geometrical constraints were far better explored by surfaces than point atoms. One feature which appears to be more or less conserved in surface contact networks (irrespective of the cutoff criteria in surface complementarity and overlap) is the almost ubiquitous presence of smaller networks (3-10 nodes) in all proteins which probably coalesce to produce larger networks specific to the particular folds. Thus, the distribution in network sizes and topologies appear to favor a nucleation-condensation phenomenon in protein packing wherein open linear chains, closed triplet cliques and other closed ring topologies could serve as basic packing units which could either get linked or recruit neighboring residues to grow into networks of larger size. This notion of packing units led to the definition of 'packing motifs', which could serve as a 'basis set' in the assembly of extended graphs. Based on these basis set of motifs, graphs were organized into families (or set of similar graphs with gradual addition of nodes following a path such that the core topology remains unaltered) and it soon became clear that some families were overwhelmingly preferred in protein topological space. These families emanated from the 'minimally connected' open linear chains and three residue

cliques (regarded as clustering units) and cast their dominant influence in frequency distribution of motifs. Other families occurred with such abysmally low frequencies that they could be considered oddities rather than the rule. Thus, in accord with the inductive approach of the current work, it was felt that larger graphs ($n > 10$) would either fall into pre-existing families or could be assembled by known motifs or their variants. This possibility was explored for networks of 15 nodes and the observations tended to support the hypothesis. The next step was to enquire whether packing motifs exhibited any preferences in terms of their constituent residues and geometry. For this, triplet cliques were selected due to their ubiquitous presence primarily as induced subgraphs embedded in larger graphs. It soon became evident that in the sub-space of hydrophobic residues, regular trends of propensities favoring specific residues or their combination do indeed exist and certain geometrical features exhibit very strong constraints (especially the approximately equilateral triangle connecting the three residue-origins and the tilt angles of aromatic residues).

References

Aftabuddin Md, Kundu S (2007). **Hydrophobic, Hydrophilic, and Charged Amino Acid Networks within Protein.** *Biophys. J.*, **93**: 225–231.

Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S (2004). **Network analysis of protein structures identifies functional residues.** *J. Mol. Biol.* **344**: 1135-1146.

Banerjee R, Sen M, Bhattacharya D, Saha P (2003). **The Jigsaw Puzzle Model: Search for Conformational Specificity in Protein Interiors.** *J. Mol. Biol.* **333**: 211–226.

Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004). **The architecture of complex weighted networks.** *Proc. Natl. Acad. Sci. USA.* **101**: 3747–3752.

Bagler G, Sinha S (2007). **Assortative mixing in protein contact networks and protein folding kinetics.** *Bioinformatics.* **23**:1760–1767.

Berman HM, Westbrook J, Feng Z, et al. (2000). **The protein data bank.** Nucleic Acids Res, **28**: 235–242.

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

Brinda KV, Vishveshwara S (2005). **A network representation of the protein structures: implications for protein stability.** *Biophys. J.* **89**: 4159-4170.

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM (Jr), Ferguson, DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995). **A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.** *J. Am. Chem. Soc.* **117**: 5179-5197.

Cheriyan J, Maheshwari SN (1988). **Finding nonseparating induced cycles and independent spanning trees in 3-connected graphs.** *J. Algorithms.* **9**: 507-537.

Crick FHC (1953). **The packing of a-helices: simple coiled coils.** *Acta Crystallog.* **6**: 689–697.

Goede A, Preissner R, Frommel C (1997) **Voronoi cell: new method for allocation of space among atoms: elimination of avoidable errors in calculation of atomic volume and density.** *J. Comput. Chem.* **18**: 1113-1123.

Greene LH, Higman VA (2003). **Uncovering Networks within protein structures.** *J. Mol. Biol.* **334:** 781-791.

Harary F (2001). **Graphs.** In *Graph Theory.* 10th Reprint. Addison-Wesley Publishing company Inc, USA & Narosa Publishing House, New Delhi, India; 10-13.

Lawrence MC, Colman PM (1993). **Shape complementarity at protein/protein interfaces.** *J. Mol. Biol.* **234**: 946–950.

Li J, Wang J, Wang W (2007). **Identifying folding nucleus based on residue contact networks of proteins.** *Proteins.* **71**: 1899-1907.

Lee B, Richards FM (1971). **The interpretation of protein structure: Estimation of static accessibility.** *J. Mol. Biol.* **55**: 379-400.

Pembroke JT (2000). **Bio-molecular modeling utilizing RasMol and PDB resources: a tutorial with HEW lysozyme.** *Biochem. Mol. Biol. Educ.* **28**: 297-300.

Punta M, Rost B (2005). **Protein folding rates estimated from contact predictions.** *J. Mol. Biol.* **348**: 507-512.

Plaxco KW, Simons KT, Baker D (1998). **Contact order, transition state placement and the refolding rates of single domain proteins.** *J. Mol. Biol.* **277**: 985-994.

Richards FM (1974). **The interpretation of protein structures total volume, group volume distributions and packing density.** *J. Mol. Biol.* **82**: 1-14.

Rother K, Hildebrand PW, Geodge A, Gruening B, Preissner R (2009). **Voronoia: analyzing packing in protein structures.** *Nucleic. Acids. Res.* **37**: D393-D395.

Singh J, Thornton JM (1985). **The interaction between phenylalanine rings in proteins.** *FEBS Letters*. **191**: 1–6.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003). **Cytoscape: A software environment for integrated models of biomolecular interaction networks.** *Genome. Res.* **13**: 2498–2504.

Vendruscolo M, Pacl E, Dobson CM, Karplus M (2001). **Three key residues from a critical contact network in a protein folding transition state.** *Nature.* **409**: 642-645.

Watts DJ, Strogatz SH (1998). **Collective dynamics of 'small-world' networks.** *Nature.* **393**: 441-442.

Word JM, Lovell SC, Richardson JS, Richardson DC (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* 285: 1735-1747.

Probing Electrostatic Complementarity within protein interiors

1. Introduction

As has been discussed in the chapter 1, complementarity in biomolecular recognition has a dual aspect and the concept appears to be particularly appealing for protein-protein interactions, due to their large interfacial surface areas (~1600 $Å^2$ on average) buried upon stereo-specific associations. The earlier chapter (Chapter 2) described the specific modes of side-chain packing geometry explored within the protein interior probed by surface complementarity. The other aspect of this specific match between the two interacting surfaces is the complementarity mediated by non-local long-range electric fields due to charged or partially charged atoms. Once the solution continuum electrostatic model **(Gilson et al., 1988)** became available for proteins, electrostatic potentials were calculated for many protein structures **(Green and Tidor, 2005; Morreale et al., 2007; Radhakrishnan and Tidor, 2008; LeMaster et al., 2009; Shibata et al., 2009)** by iteratively solving the Poisson-Boltzmann equation (PBE) as implemented in the program DelPhi **(Nichollos and Honig, 1991).** In 1997, complementarity in both charge and electrostatic potential were first probed at protein-protein interfaces **(McCoy et al., 1997),** and the complementarity was found to be significant for potential rather than charge. A partially desolvated model (a protein is partially desolvated by the volume of the other protein in the complex, thus, leaving a low dielectric region in the close vicinity of the interacting molecule) was also preferred as a more accurate and suitable method to compute electrostatic potential at protein-protein interfaces than a fully solvated model **(McCoy et al., 1997)**. In this chapter, the computation of complementarity in electrostatic potential between buried amino acids and the rest of the polypeptide chain has been discussed.

One central problem for continuum electrostatic calculations is the precise assessment of the dielectric within proteins. As is well known, dielectric of a medium is contributed by induced polarization and dipolar reorientation **(Gilson and Honig, 1986)**. Generally, the interior of a protein is considered to have a low (ε_p = 2 to 4) dielectric which significantly increases at solvent exposed surfaces (ε_p = 20 to 40). External

dielectric of the surrounding aqueous solvent is very high (generally set to 80) and primarily contributed by high degree of dipolar rotations **(Gilson and Honig, 1986)**. Dipoles in proteins are polar substituents, primarily found at backbone amide (…CO=NH…) and polar side-chains. Being constrained within the protein structure, they have low degree of freedom particularly at the interior of the molecule leading to the characteristic low internal dielectric **(Gilson and Honig., 1986)**. However, the degree of freedom increases with increasing solvent exposure and so does the dielectric at the solvent exposed residue surfaces. Another related problem is the ionic strength of counter ions in the surrounding aqueous medium of the protein which is related to the choice of whether linear or non-linear methods to solve the Poisson-Boltzmann potential will be appropriate. The original non-linear PBE is implemented for multi-dielectric systems with defined charge densities contributed both by the protein and the counter-ions, whereas, for systems, that do not involve high charge densities, a simplified linearized form of the PBE is preferred and more rapidly evaluated **(Mandell et al., 2001)**. Both these problems were adequately addressed in the course of the calculation, details of which will be found below. In parallel, surface complementraity was also estimated for interior residues of proteins for sake of comparison of the two (short and long range) complementarity measures.

1. Materials and Methods

2.1. Databases

A subset of the database, **DB1** discussed in the earlier chapter (chapter 2), consisting of 400 polypeptide chains **(DB2)** was assembled by removing proteins with deeply embedded prosthetic groups (e.g., cytochromes) and any missing atoms. **DB2** (composed of 65 all α, 70 all β, 106 α|β, 124 α+β and 35 multi-domain proteins : see **Supplementary Information in the CD enclosed**) was used in the calculation of electrostatic complementarity **(E_m)** of amino acid residues and their related statistics. Of these, 62 proteins were found to contain metal ions as an integral part of their structure.

Hydrogen atoms were geometrically fixed to all structures by the program REDUCE **(Word et al., 1999).**

2.2. Partial Charge Assignment

Prior to the calculation of electrostatic potential, partial charges and atomic radii for all protein atoms were assigned from the AMBER94 all atom molecular mechanics force field **(Cornell et al., 1995)**. Asp, Glu, Lys, Arg, doubly-protonated histidine (Hip) and both the carboxy, amino terminal groups were considered to be ionized. Crystallographic water molecules and surface bound ligands were excluded from the calculations and thus modeled as bulk solvent. Ionic radii were assigned to the bound metal ions according to their charges **(Shannon, 1976)**.

2.3. Van der Waals Surface and solvent accessibility

Van der Waals surfaces of the polypeptide chains were sampled at 10 dots / $Å^2$. The details of the surface generation have been discussed in a previous chapter **(**Chapter 2**)**. The exposure of individual atoms to solvent was estimated by rolling a probe sphere of radius 1.4 Å over the protein atoms **(Lee and Richards, 1971)** and burial (***Bur***) of individual residues was estimated by the ratio of solvent accessible surface areas of the amino acid X in the polypeptide chain to that of an identical residue located in a Gly–X–Gly peptide fragment with a fully extended conformation.

2.4. Calculation of Electrostatic Potential

The finite difference Poisson–Boltzmann method as implemented in Delphi (version 4) **(Gilson et al., 1988)** was used to compute the electrostatic potential of the molecular surface along the polypeptide chain. The protein interior was considered to be a low (dielectric constant of 2) and the surrounding solvent, a high dielectric medium (dielectric constant of 80). Ionic strength was set to zero as adoption of physiological strength has been found to have little effect on the final electrostatic solution **(Radhakrishnan and Tidor, 2008; Jackson and Sternberg, 1994)** and calculations were performed at 298 K. The dielectric boundary and the partial charges were mapped

onto a cubic grid of size either 151× 151 × 151 or 201 × 201 × 201 grid points / side, the latter for those proteins which exhibited pronounced asymmetry in their physical dimensions. The percentage grid fill was set to 80% with a scale of 1.2 grid points / Å. Boundary potentials were approximated by the Debye–Huckel potential of the dipole equivalent to the molecular charge distribution. A probe radius of 1.4 Å was used to delineate the dielectric boundary. The linearized Poisson-Boltzmann equation was then solved iteratively until convergence; the number of cycles to convergence being automatically determined by the program (the convergence threshold based on the maximum change in potential being set to 0.0001 kT/e) and monitored by examining a plot of convergence in the output log file.

2.5. Electrostatic Complementarity (E_m) at the interface

Delphi requires a set of surface points on which the electrostatic potentials are to be computed along with a set of atoms contributing to the potential. Subsequent to the generation of the van der Waals surface of the entire polypeptide chain, the dot surface points of the individual amino acids (targets) were identified and fed to the program along with the selected set of (charged) atoms. The electrostatic potential for each residue surface was then calculated twice, 1) due to the atoms of the particular target residue and 2) from the rest of the protein excluding the selected amino acid. In either case, the atoms not contributing to the potential (dummy atoms) were only assigned their radii with zero charge, to maintain the scaling and orientation of the molecule on the grid. Thus each dot surface point of the (selected) residue was tagged with two values of electrostatic potential. Adapted from the function EC, originally proposed by McCoy et al. (for protein-protein interfaces) **(McCoy et al., 1997)**, electrostatic potential complementarity **(E_m)** of an amino acid residue (within protein) was then defined as the negative of the correlation coefficient (Pearsons) between these two sets of potential values,

$$E_m = -\left(\frac{\sum_{i=1}^{N} (\varphi(i) - \overline{\varphi})(\varphi`(i) - \overline{\varphi}`)}{\left(\sum_{i=1}^{N} (\varphi(i) - \overline{\varphi})^2 \sum_{i=1}^{N} (\varphi`(i) - \overline{\varphi}`)^2 \right)^{1/2}} \right) \quad (1)$$

where, for a given residue consisting of a total of N dot surface points, $\varphi(i)$ is the potential on its ith point realized due to its own atoms and $\varphi`(i)$, due to the rest of the protein atoms, $\overline{\varphi}$ and $\overline{\varphi}`$ are the mean potentials of $\varphi(i)$, i = 1…N and $\varphi`(i)$, i = 1…N respectively.

Subsequent to the calculation of electrostatic potentials, the values corresponding to N dot surface points were also divided into two distinct sets, based on whether the dot point was obtained from main chain or side chain atoms of the target residue and E_m calculated separately for each set. Thus for a given residue, electrostatic potential complementarity was estimated for the entire residue (E_m^{all}, as described above), the side chain surface points (E_m^{sc}) and the main chain surface points (E_m^{mc}).

2.6. Surface Complementarity (S_m) at the interface

The calculation of surface complementarity has been discussed extensively in the previous chapter (Chapter 2). Briefly, surface complementarity **(S_m)** can be calculated between the side chain surface points of a target residue and the all other dot points in its immediate neighborhood (within a distance of 3.5 Å), contributed by the rest of the protein. Any dot surface point (which is essentially an area element) is characterized by its coordinates (x, y, z) and the direction cosines of its normal (dl, dm, dn). The surface complementarity measure **(S_m)** is then defined (following Lawrence and Colman **(Lawrence and Colman., 1993)**) to be the median of the distribution $\{S(a,b)\}$, $S(a,b)$ being calculated by the following equation:

$$S(a,b) = \mathbf{n}_a \cdot \mathbf{n}_b . \exp(-w.d_{ab}^{\ 2}) \qquad (2)$$

where $\mathbf{n}_a$ and $\mathbf{n}_b$ are two unit normal vectors, corresponding to the dot surface point a (located on the side chain surface of the target residue) and b (the dot point nearest to a, within 3.5 Å) respectively, with d_{ab} the distance between them and w, a scaling constant set to 0.5. Subsequent to identifying nearest neighbors, the side chain surface points of

the specified residue can also be partitioned into two sets by virtue of their neighbors coming from either side chain or main chain atoms, and S_m calculated separately for each set. Thus, every target residue (side chain) has three measures of S_m based on the choice of its nearest neighbors (surface points), whether obtained from side chain (S_m^{sc}), main chain (S_m^{mc}) atoms alone or all atoms (S_m^{all}). Since glycines lack any non-hydrogen side chain atom; they were excluded as targets from all calculations.

2. Results and Discussion

3.1. Independence of electrostatic potential on ionic strength

Linearized Poisson-Boltzmann equation (LPBE) as implemented in Delphi **(Nichollos and Honig, 1991)** was iteratively solved until convergence to compute the electrostatic potential at the protein interior and estimation of electrostatic complementarity **(E_m)** adapted **(**see **Eq. 1)** from a method proposed by McCoy et al. for protein-protein interfaces **(McCoy et al., 1997)**. Nonlinear Poisson-Boltzmann equation at non-zero ionic strengths is preferred for highly charged molecules like DNA **(Nichollos and Honig, 1991)**, microtubules and ribosomal subunits **(Baker et al., 2001)**. Globular proteins, however, have appreciably low net charge densities and LPBE has been used extensively to compute electrostatic potentials at protein-protein interfaces and solvent exposed residue-surfaces **(Radhakrishnan and Tidor, 2008; Green and Tidor, 2005).** Electrostatic potentials estimated by nonlinear PBE (in a trial calculation involving 150 polypeptide chains) under physiological counter-ionic strength (0.15 M NaCl, ion exclusion radii: 2.0 Å) were virtually identical to those calculated by LBPE **(Figure 1)**. Similar results were also previously obtained **(Jackson and Sternberg, 1994; Radhakrishnan and Tidor, 2008)** where calculations using LPBE and non-linear PBE with nonzero (physiological) ionic strength were in good agreement for biologically relevant charge magnitudes.

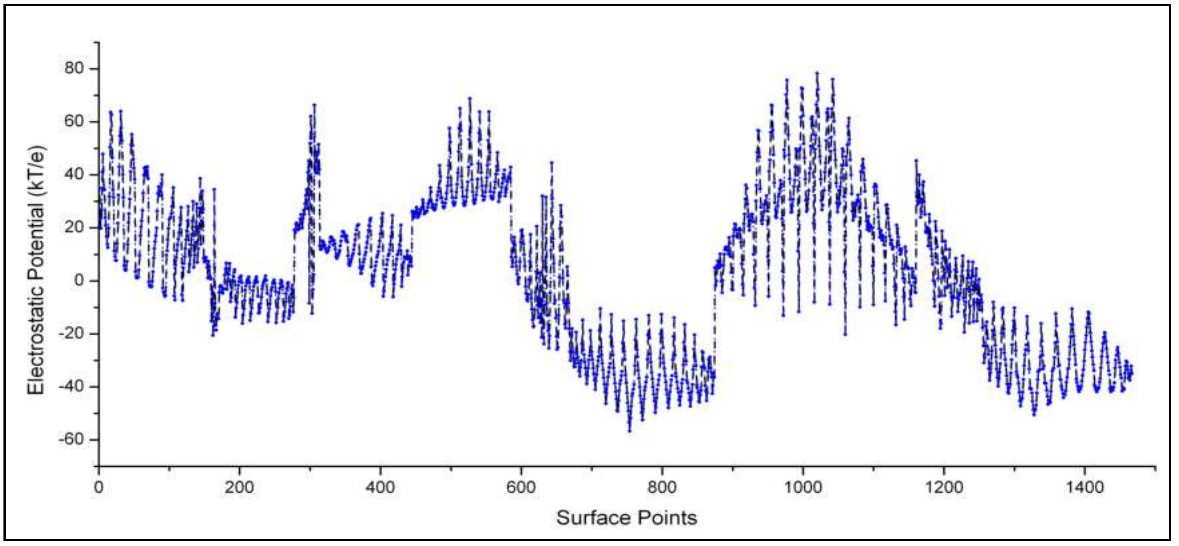


Figure 1. Electrostatic potentials computed by linearized Poisson-Boltzmaan equation (at zero ionic strength) and non-linear PBE at physiological counter-ionic strength (0.15 M NaCl) are virtually identical. Potentials calculated on individual surface points of a completely buried asparagine (58-Asn: 2HAQ) plotted in **blue:** LPBE; **black:** nonlinear PBE. The plot shows practically identical values in potential (obtained from the two methods) realized due to the atoms of the selected residue. Similar agreement has also been obtained for potentials realized due to the rest of the charged protein atoms.

3.2. Invariance of electrostatic complementarity on the internal dielectric

$\boldsymbol{E_m}$ was estimated for all residues at the protein interior (burial $\leq$ 0.30; see **Materials and Methods**) from a database of 400 polypeptide chains **(DB2)**. In order to test the sensitivity of $\boldsymbol{E_m}$ with respect to the internal dielectric of the continuum (ε_p), all calculations were repeated thrice setting ε_p to 2, 4 and 10 respectively. The root mean square deviations between these three sets of $\boldsymbol{E_m}$ values for different residues were negligible, indicating the invariance of $\boldsymbol{E_m}$ at least in the commonly used ranges of ε_p **(Figure 2)**. Identical calculations performed with higher internal dielectric (ε_p = 20 & 40) also preserved the overall trends in the results. It should be noted that $\boldsymbol{E_m}$ estimates the correlation between potentials generated by the two sets of atoms (over a collection of surface points), regardless of their magnitude.

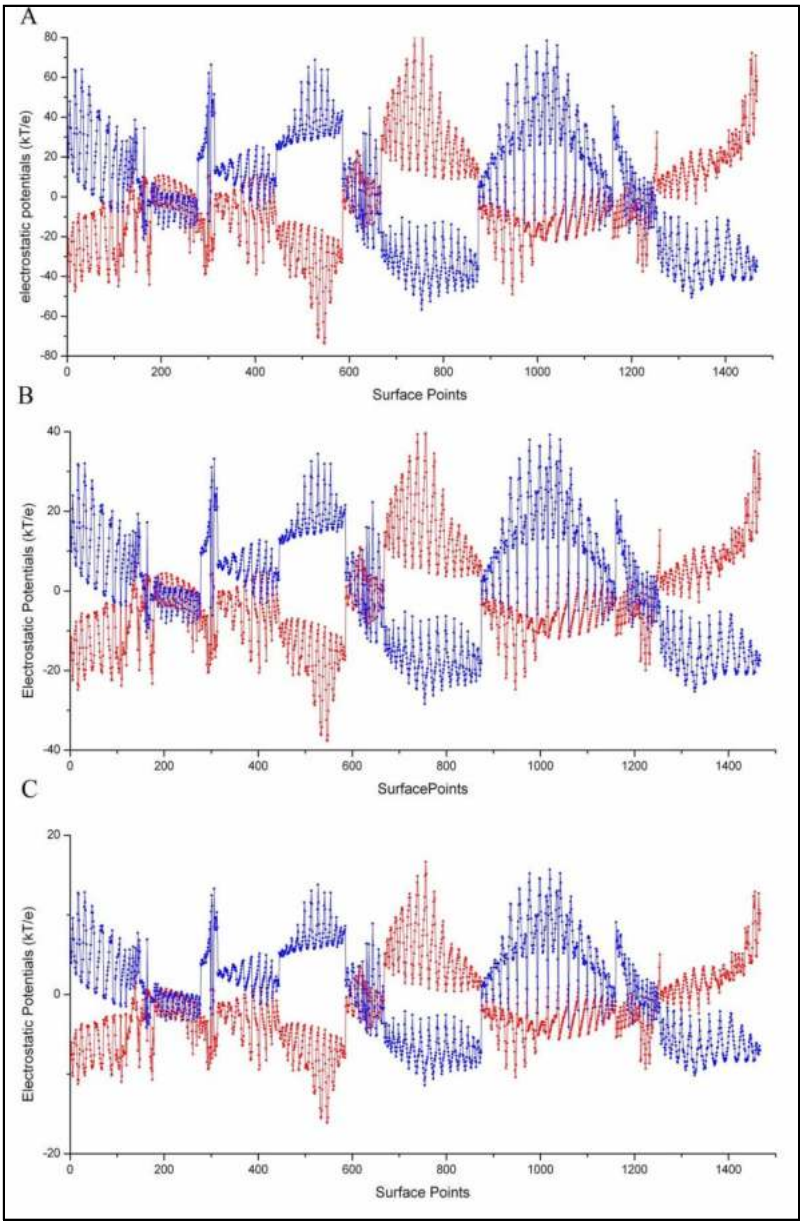


Figure 2. Variation in the internal dielectric of the continuum does not alter electrostatic complementarity for interior residues. Electrostatic potentials computed at the commonly used ranges of internal dielectrics: ε_p = 2 (A), 4 (B), 10 (C) for a completely buried asparagine (58-Asn: 2HAQ) (**blue:** due to the atoms of the selected residue, **red:** due to the rest of the charged protein atoms). As expected, change in ε_p changes merely the scale of the potentials thereby leading to conserved E_m values.

3.3. Anti-correlated surface electrostatic potential at the protein interior

Prior to statistical analysis, all completely / partially buried (target) residues were distributed in three burial bins (burial: 0.0-0.05, 0.05-0.15, 0.15-0.30, see **Materials and Methods**). Enumeration of the average E_m values, in each burial bin for different amino acids (targets), calculated over the entire residue surface ($\overline{E_m^{all}}$), revealed a fairly uniform distribution among the different residues, within the range ~ 0.5 – 0.7 **(Table 1)**. The high positive values of E_m^{all} throughout the protein interior suggest that individual residues buried within proteins have anti-correlated (complementary) surface electrostatic potentials **(Figure 3)** similar to protein-protein interfaces **(McCoy et al., 1997)**.

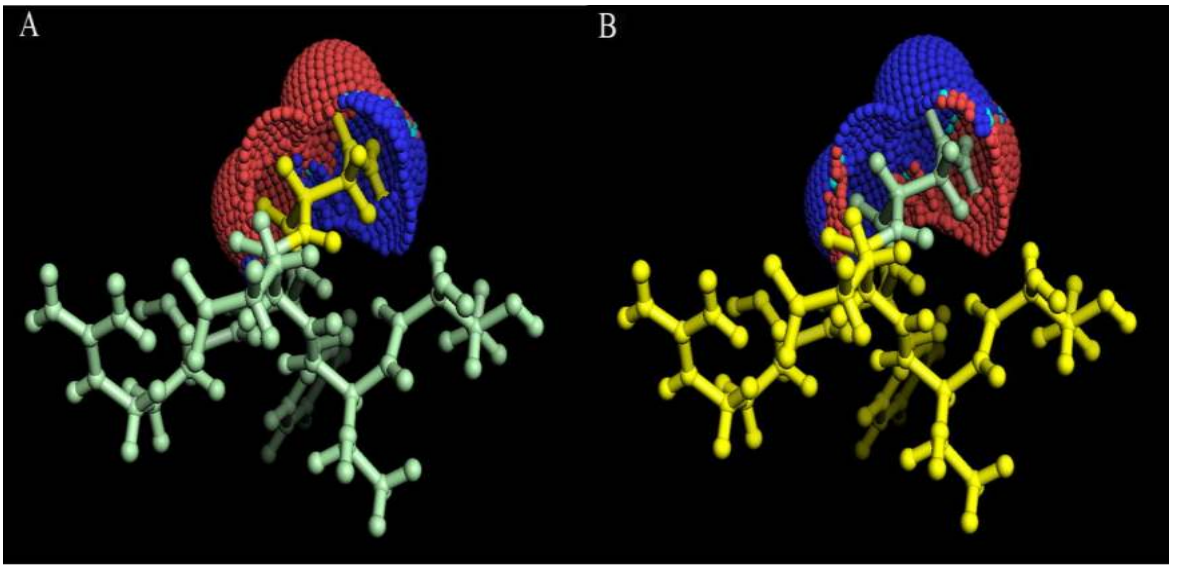


Figure 3. Molecular surface of an individual buried residue (target) exhibiting anti-correlated (complementary) electrostatic potentials. The figure shows the van der Waals surface (displayed as non-bonded spheres) of a completely buried asparagine (58-Asn from 2HAQ) along with its own atoms accompanied by a few more residues along the polypeptide chain as representative of the 'rest of the protein atoms' (sticks). Atoms (along with their interconnecting bonds) are colored by 'bright yellow' when 'charged' and by 'pale green' when 'uncharged'. Surface coloring follows standard conventions where patches of positive potentials are colored by 'blue', negative potentials by 'red' and neutral (0.0 ± 0.5 kT/e) by 'cyan'. (A) shows potentials realized due to the charged atoms of the residue itself whereas (B) shows potentials due to the rest of the charged protein atoms. Figure constructed by **PyMol** [http://www.pymol.org/].

In fact, $\overline{E_m^{all}}$ for hydrophobic residues were comparable to those for polar and charged amino acids. From these observations, it was thought that the main chain surface points could be contributing predominantly to E_m^{all}, especially for hydrophobic residues. In order to test this hypothesis, the surface points were segregated by virtue of their residence on main chain / side chain atoms and $\boldsymbol{E_m}$ calculated separately for each set, namely, E_m^{sc}, E_m^{mc} for side and main chain surface points respectively. As expected, $\overline{E_m^{mc}}$ were again uniform for all the amino acids and comparable in magnitude to $\overline{E_m^{all}}$. Interestingly, even for hydrophobic residues, $\overline{E_m^{sc}}$ were also found to exhibit fairly significant values. However, differences were observed in $\overline{E_m^{sc}}$, between hydrophobic (Val: 0.48, Leu: 0.46, Ilu: 0.48, Phe: 0.41) and charged / polar residues (Asn: 0.67, Gln: 0.64, Asp: 0.61, Glu: 0.63, Lys: 0.62, Arg: 0.56), though within one standard deviation (~ 0.1 – 0.25) **(Table 1)**. Somewhat reduced values were obtained for sulphur containing amino acids (Cys: 0.34, Met: 0.32) and proline (0.34). A similar pattern was observed in all three burial bins indicating that within protein interior, the distribution in $\boldsymbol{E_m}$ appear to be independent of the exposure of a residue to solvent.

Table 1. Native electrostatic complementarities of completely buried residues: Average $\boldsymbol{E_m}$ and their standard deviations (in parentheses) for different residues in the 1st burial bin ($0.0 \leq \boldsymbol{Bur} \leq 0.05$) calculated from all atoms on (1) the entire residue surface ($\overline{E_m^{all}}$), (2) on side chain surface ($\overline{E_m^{sc}}$) and (3) on main chain surface ($\overline{E_m^{mc}}$).

Residue	$\overline{E_m^{all}}$	$\overline{E_m^{sc}}$	$\overline{E_m^{mc}}$
ALA	0.68 (0.17)	0.48 (0.25)	0.72 (0.17)
VAL	0.62 (0.16)	0.48 (0.18)	0.72 (0.16)
LEU	0.61 (0.16)	0.46 (0.19)	0.73 (0.16)
ILE	0.61 (0.16)	0.48 (0.17)	0.72 (0.16)
PHE	0.56 (0.15)	0.41 (0.16)	0.70 (0.17)
TYR	0.58 (0.15)	0.50 (0.19)	0.69 (0.18)
TRP	0.57 (0.15)	0.50 (0.17)	0.68 (0.20)
SER	0.64 (0.18)	0.59 (0.27)	0.67 (0.18)
THR	0.62 (0.16)	0.55 (0.23)	0.68 (0.18)
CYS	0.51 (0.18)	0.34 (0.22)	0.66 (0.21)
MET	0.45 (0.13)	0.32 (0.16)	0.72 (0.16)
ASP	0.63 (0.22)	0.61 (0.26)	0.62 (0.17)
GLU	0.64 (0.25)	0.63 (0.28)	0.66 (0.19)
ASN	0.68 (0.17)	0.67 (0.22)	0.68 (0.17)
GLN	0.66 (0.17)	0.64 (0.21)	0.70 (0.18)
LYS	0.72 (0.17)	0.62 (0.22)	0.75 (0.15)
ARG	0.68 (0.16)	0.56 (0.19)	0.75 (0.15)
PRO	0.53 (0.20)	0.34 (0.23)	0.65 (0.19)
HIS	0.54 (0.26)	0.50 (0.28)	0.65 (0.21)

3.4. Contribution of the native main-chain trajectory on Electrostatic Complementarity

In order to assess the relative contribution of side or main chain atoms to $\boldsymbol{E_m}$, four more sets of calculations were performed based on the choice of residue surface (target: side chain / main chain) on which to calculate the electrostatic potentials and the atoms (side chain / main chain) contributing to the potential.

S1: Main chain surface, main chain atoms;

S2: Side chain surface, main chain atoms;

S3: Side chain surface, side chain atoms;

S4: Side chain surface, side chain atoms of the target and all atoms from the 'rest of the polypeptide chain'.

But for the choice of surfaces and atoms the method for calculating $\boldsymbol{E_m}$ was identical to the one outlined above. As expected, **S1** gave a uniform distribution in $\overline{E_m}$ with elevated values for all residues **(Table 2)**. For **S2,** fairly significant values of $\overline{E_m}$ were still retained for hydrophobic residues (Ala: 0.43, Val: 0.44, Leu: 0.42, Ile: 0.43, Phe: 0.36, Met: 0.38), which is a reflection of the long range electric fields generated by the main chain atoms overwhelmingly contributing to the complementarity attained on hydrophobic side chain surfaces. This was confirmed by the comparison of $\overline{E_m}$ in **S2** and $\overline{E_m^{sc}}$ whereby both sets of values were almost identical for hydrophobic residues **(Table 1 & Table 2)**, while, polar / charged residues exhibited a marked reduction in **S2** compared to $\overline{E_m^{sc}}$, since the contribution of side chain atoms carrying high partial charges were disregarded in **S2**. For both **S3** and **S4,** $\overline{E_m}$ for hydrophobic residues were practically negligible **(Table 2)**, whereas polar / charged residues gave consistently high values for **S4,** while distinctly reduced for **S3**. The substantial increase in $\overline{E_m}$ for **S4** relative to **S3** (except for alanine) was indicative of the considerable role being played by the main chain atoms (contributed by the rest of the polypeptide chain) in the overall determination of $\overline{E_m}$. This holds true even for hydrophilic amino acids where the main chain atoms contribute appreciably to the neutralization of the electric fields generated by polar / charged side chain atoms.

Table 2. Assessment of the relative contributions of main chain and side chain atoms on electrostatic complementarity. Average $\boldsymbol{E_m}$ ($\overline{E_m}$) and their standard deviations (in parentheses) for different residues ($0.00 \leq \boldsymbol{Bur} \leq 0.05$) calculated from different combinations of atomic sets and surfaces: **S1:** main chain atoms (target) versus main chain atoms ('rest of the protein') on main chain surface (of the target residue); **S2:** main chain atoms (target) versus main chain atoms (rest) on side chain surface (target). **S3:** side chain atoms (target) versus side chain atoms (rest) on side chain surface (target); **S4:** side chain atoms (target) versus all atoms (rest) on side chain surface (target).

Residue	$\overline{E_m}$			
	S1	**S2**	**S3**	**S4**
ALA	0.63 (0.25)	0.43 (0.22)	-0.04 (0.24)	-0.08 (0.18)
VAL	0.65 (0.22)	0.44 (0.23)	-0.02 (0.25)	0.15 (0.19)
LEU	0.65 (0.22)	0.42 (0.24)	0.01 (0.26)	0.14 (0.21)
ILE	0.65 (0.23)	0.43 (0.22)	-0.01 (0.28)	0.14 (0.24)
PHE	0.63 (0.24)	0.36 (0.28)	0.13 (0.16)	0.22 (0.17)
TYR	0.61 (0.25)	0.33 (0.27)	0.25 (0.27)	0.43 (0.22)
TRP	0.59 (0.28)	0.28 (0.26)	0.30 (0.22)	0.42 (0.19)
SER	0.53 (0.28)	0.25 (0.34)	0.24 (0.37)	0.53 (0.28)
THR	0.55 (0.28)	0.30 (0.29)	0.16 (0.34)	0.45 (0.28)
CYS	0.58 (0.27)	0.36 (0.23)	0.06 (0.29)	0.18 (0.25)
MET	0.64 (0.23)	0.38 (0.22)	0.12 (0.22)	0.21 (0.17)
ASP	0.47 (0.30)	0.17 (0.28)	0.27 (0.41)	0.55 (0.33)
GLU	0.56 (0.28)	0.27 (0.28)	0.31 (0.42)	0.54 (0.37)
ASN	0.54 (0.27)	0.23 (0.30)	0.33 (0.38)	0.64 (0.23)
GLN	0.57 (0.29)	0.29 (0.28)	0.32 (0.37)	0.60 (0.24)
LYS	0.62 (0.24)	0.29 (0.23)	0.40 (0.37)	0.58 (0.24)
ARG	0.59 (0.25)	0.23 (0.22)	0.28 (0.30)	0.53 (0.21)
PRO	0.49 (0.26)	0.25 (0.28)	-0.15 (0.35)	-0.05 (0.28)
HIS	0.55 (0.29)	0.30 (0.28)	0.28 (0.36)	0.49 (0.31)

It is thus evident that, the long range electric fields generated by main chain atoms cast their shadow over the side chain surface in such a manner, that all residues regardless of their hydrophobicity and burial attain a fairly uniform level of overall complementarity. Polar / charged (side chain) atoms of hydrophilic residues additionally contribute to the elevated complementarity attained on their side chain surfaces.

3.5. Comparison of surface and electrostatic complementarity for buried residues

To the best of our knowledge this is the first time that electrostatic complementarity has been calculated within proteins. In parallel, surface complementarity has also been computed for interior residues in order to compare between the two (short-range and long-range) complementarity measures. The results show that one of the universal characteristics of correctly folded proteins is the almost uniformly elevated values in S_m^{sc} and E_m^{sc} attained by all deeply buried residues **(Figure 4)**.

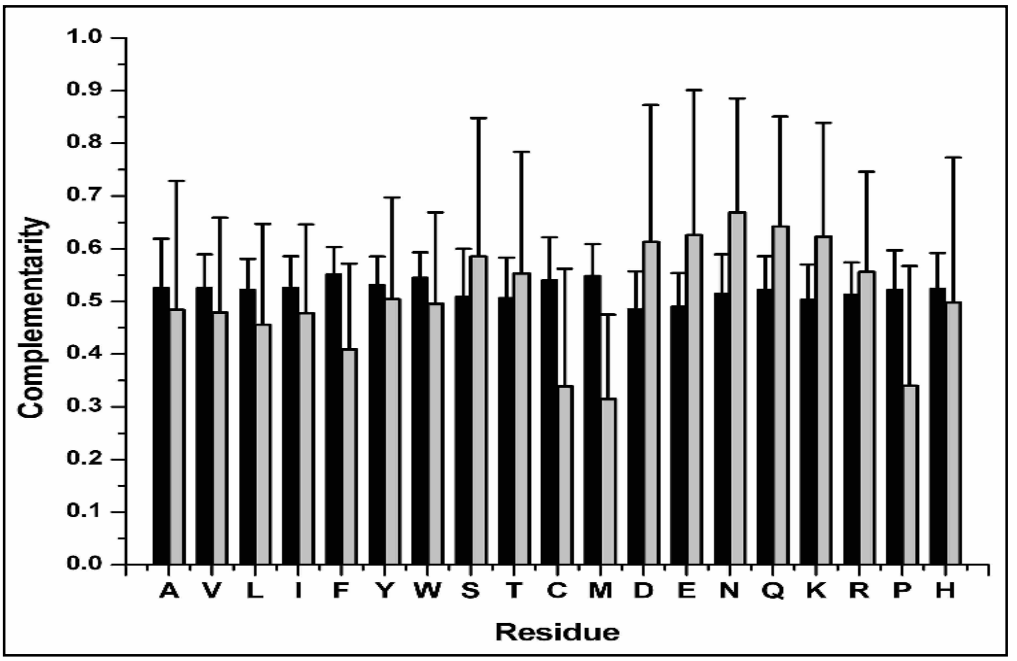


Figure 4. Trends in surface and electrostatic complementarities for deeply buried amino acid side chains. Mean $\overline{S_m^{sc}}$ (black), $\overline{E_m^{sc}}$ (gray), plotted as filled thick bars along with their standard deviations (represented by error bars) for different residues in the 1st burial bin ($0.0 \leq \boldsymbol{Bur} \leq 0.05$).

However, the constraints in S_m^{sc} appear to be more stringent relative to E_m^{sc}, given its reduced standard deviation, compared to the latter. The nature of short and long range forces which determine the values of S_m^{sc}, E_m^{sc} also gives rise to their contrasting features. S_m^{sc} is a function of burial whereas E_m^{sc} is not **(Table 3)** and the primary determinants of S_m^{sc} are side chain atoms (for all residues) and main chain atoms in the case of E_m^{sc} for hydrophobic residues, while both side chain and main chain atoms contribute equally to the E_m^{sc} of hydrophilic residues.

Table 3: Comparison of the values between electrostatic and shape complementarities: Mean $\overline{E_m^{sc}}$, $\overline{S_m^{sc}}$ and their standard deviations (in parenthesis) tabulated for different amino acid residues distributed in three burial bins (bin1: $0.00 \leq$ ***Bur*** ≤ 0.05; bin2: $0.05 <$ ***Bur*** ≤ 0.15; bin3: $0.15 <$ ***Bur*** ≤ 0.30) where '***Bur***' stands for the burial ratio.

Residue	$\overline{E_m^{sc}}$			$\overline{S_m^{sc}}$		
	bin1	**bin2**	**bin3**	**bin1**	**bin2**	**bin3**
ALA	0.48 (0.25)	0.46 (0.25)	0.50 (0.26)	0.53 (0.09)	0.43 (0.12)	0.30 (0.15)
VAL	0.48 (0.18)	0.45 (0.17)	0.46 (0.18)	0.53 (0.06)	0.45 (0.08)	0.35 (0.11)
LEU	0.46 (0.19)	0.41 (0.20)	0.43 (0.21)	0.52 (0.06)	0.45 (0.08)	0.34 (0.10)
ILE	0.48 (0.17)	0.43 (0.18)	0.46 (0.18)	0.53 (0.06)	0.46 (0.08)	0.35 (0.10)
PHE	0.41 (0.16)	0.40 (0.18)	0.42 (0.18)	0.55 (0.05)	0.49 (0.07)	0.40 (0.10)
TYR	0.50 (0.19)	0.48 (0.20)	0.45 (0.20)	0.53 (0.05)	0.46 (0.07)	0.36 (0.09)
TRP	0.50 (0.17)	0.46 (0.20)	0.47 (0.18)	0.55 (0.05)	0.48 (0.06)	0.38 (0.09)
SER	0.59 (0.26)	0.54 (0.28)	0.54 (0.29)	0.51 (0.09)	0.41 (0.12)	0.28 (0.14)
THR	0.55 (0.23)	0.54 (0.24)	0.52 (0.26)	0.51 (0.08)	0.42 (0.10)	0.30 (0.12)
CYS	0.35 (0.22)	0.32 (0.20)	0.29 (0.23)	0.54 (0.08)	0.45 (0.12)	0.34 (0.14)
MET	0.32 (0.16)	0.28 (0.15)	0.27 (0.15)	0.55 (0.06)	0.46 (0.08)	0.33 (0.11)
ASP	0.61 (0.26)	0.63 (0.22)	0.63 (0.23)	0.49 (0.07)	0.39 (0.09)	0.26 (0.10)
GLU	0.63 (0.28)	0.61 (0.22)	0.57 (0.22)	0.49 (0.06)	0.39 (0.08)	0.27 (0.09)
ASN	0.67 (0.22)	0.63 (0.27)	0.59 (0.25)	0.52 (0.07)	0.43 (0.09)	0.31 (0.11)
GLN	0.64 (0.21)	0.63 (0.19)	0.53 (0.23)	0.52 (0.06)	0.42 (0.08)	0.30 (0.10)
LYS	0.62 (0.22)	0.56 (0.25)	0.51 (0.23)	0.50 (0.07)	0.42 (0.08)	0.30 (0.10)
ARG	0.56 (0.19)	0.60 (0.19)	0.58 (0.19)	0.51 (0.06)	0.43 (0.08)	0.30 (0.10)
PRO	0.34 (0.23)	0.38 (0.21)	0.40 (0.22)	0.52 (0.07)	0.43 (0.10)	0.31 (0.12)
HIS	0.50 (0.28)	0.50 (0.25)	0.46 (0.29)	0.53 (0.07)	0.45 (0.08)	0.34 (0.10)

4. Conclusion

Electrostatic complementarity was computed (for the first time to our knowledge) for residues buried within the native protein interior. All amino-acids irrespective of their hydrophobicity and charge seemed to attain elevated level of complementarity. The most interesting and insightful finding was that the native main-chain trajectories appeared to cast their shadow over the entire side-chain in order to attain substantial complementarity. This was particularly evident for the case of hydrophobic residues. Thus, for a correctly folded globular protein, the entire polypeptide chain meticulously balance the electric fields arising from different parts of the folded chain, so as to neutralize all destabilizing electrostatic effects. Unlike surface complementarity, being a long range effect, electrostatic complementarity also found to be independent of solvent exposure of residues.

Reference

Baker NA, Sept A, Joseph S, Holst MJ, McCammon JA. (2001). **Electrostatics of nanosystems: Application to microtubules and the ribosome.** *Proc. Nat. Acad. Sci. USA*. **98**: 10037–10041.

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995). **A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.** *J Am Chem Soc* **117**: 5179-5197.

Gilson M, Honig B (1986). **The dielectric constant of a folded protein.** *Biopolymers.* **25**: 2097-2119.

Gilson M, Sharp K, Honig B (1988). **Calculation of the Total Electrostatic Energy of a Macromolecular System: Solvation Energies, Binding Energies, and Conformational Analysis.** *Proteins: Struct. Func. Genet.* **4**: 7-18.

Green DF, Tidor B (2005). **Design of Improved Protein Inhibitors of HIV-1 Cell Entry: Optimization of Electrostatic Interactions at the Binding Interface.** *Proteins* **60**: 644–657.

Jackson RM, Sternberg MJE (1994). **Application of scaled particle theory to model the hydrophobic effect: implications for molecular association and protein stability.** *Protein. Eng.* **7**: 371-383.

Lee B, Richards FM (1971). **The interpretation of protein structure: Estimation of static accessibility.** *J. Mol. Biol.* **55**: 379-400.

LeMaster DM, Anderson JS, Hernandez G (2009). **Peptide conformer acidity analysis of protein flexibility monitored by hydrogen exchange.** *Biochemistry* **48**: 9256-9265.

Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF (2001). **Protein docking using continuum electrostatics and geometric fit.** *Protein Eng.* **14**: 105–113.

McCoy AJ, Epa VC, Colman PM (1997). **Electrostatic complementarity at protein/protein interfaces.** *J. Mol. Biol.* **268**: 570-584.

Nichollos A, Honig B (1991). **A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation.** *J. Comput. Chem.* **12**: 435-445.

Radhakrishnan ML, Tidor B (2008). **Optimal drug cocktail design: methods for targeting molecular ensembles and insights from theoretical model systems.** *J. Chem. Inf. Model* **48**: 1055–1073.

Shannon RD (1976). **Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides.** *Acta Cryst* A **32**: 751-767.

Shibata N, Ueda Y, Takeuchi D, Haruyama Y, Kojima S, Sato J, Niimura Y, Kitamura M, Higuchi Y (2009). **Structure analysis of the flavoredoxin from Desulfovibrio vulgaris Miyazaki F reveals key residues that discriminate the functions and properties of the flavin reductase family.** *FEBS Journal* **276**: 4840–4853.

Word JM, Lovell SC, Richardson JS, Richardson DC (1999). **Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation.** *J Mol Biol,* **285**: 1735-1747.

Application of the combined use of shape and electrostatic complementarity in protein fold recognition: an attempt to bridge the gap between binding and folding

1. Introduction

As discussed in the earlier chapter (chapter 3), surface and electrostatic complementarities were probed for residues interior to proteins from a representative database. The main idea behind the study was to test whether the concept could serve as a common conceptual platform in order to discuss binding and folding. As also briefly mentioned in chapter 1, for small molecule ligands or cofactors binding to proteins, the concept of complementarity appears to be only partially true. The same ligand not only adopts a wide range of conformations upon binding to different proteins, but also the binding pocket exhibits variability in their shapes and physicochemical characteristics, than can be accounted for by the multiple conformations adopted by the ligand **(Stockwell and Thronton, 2006; Kahraman et al., 2007; Kahraman et al., 2010)**. For protein-protein interfaces, however, the concept appears to have greater plausibility and wider appeal due to the relatively larger size of protein-protein interfaces (~1600 $Å^2$ on average) **(Lo Conte et al., 1999)**. A variety of shape correlation and electrostatic complementarity measures incorporated into docking algorithms have been effective in predicting the interfaces between interacting proteins **(Mandell et al., 2001, Heifetz et al., 2002).** Colman and McCoy along with coworkers have formulated and estimated shape correlation **(S_c)** and electrostatic complementarity **(EC)** measures for a wide range of proteins in quaternary association, protein-inhibitor and antigen-antibody complexes **(Lawrence and Colman, 1993; McCoy et al., 1997)**. It thus appears reasonable that threshold values of geometric and electrostatic complementarities will have to be satisfied for the stereo-specific association between two polypeptide chains. Within proteins, surface complementarity **(S_m)** has been used to enumerate specific modes of packing between amino acid side chains **(Basu et al., 2011)** and somewhat analogous to protein interfaces, all residues upon burial achieve uniformly high measures of surface fit **(Banerjee et al., 2003)**.

Although the notion of complementarity lends itself naturally to the characterization of inter-protein association, of late there have been several suggestions in

the literature to approach both binding and folding from a common conceptual platform **(Bahadur and Chakrabarti, 2009)**. The native conformation adopted by the polypeptide chain leads to the stereo-specific packing of its buried side chains and optimal electrostatic interactions due to the strategic three dimensional placements of charges. Thus, folding can possibly be described as the self recognition of the polypeptide chain as it collapses onto itself. However, one inherent problem in equating binding with folding lies in the different characteristics of protein interiors compared to interfaces. Barring dimers, interfaces resemble protein surfaces rather than interiors, both in their composition and spatial distribution of amino acid residues. Unlike hydrophobic clusters found within proteins, nonpolar residues are found in isolation at protein-protein interfaces, surrounded by polar or charged amino acids. However, despite these differences, the fact remains that both interfacial **(Lawrence and Colman, 1993)** and interior atoms **(Banerjee et al., 2003; Basu et al., 2011)** have to satisfy fairly stringent packing requirements and at least for the interfaces, significant values of electrostatic complementarity have been found **(McCoy et al., 1997)**. To explore the similarities or equivalence between binding and folding (in terms of complementarity), the current chapter describes the design and utility of scoring functions (based on the combined use of the two complementarity measures) for fold recognition much similar to functions that discriminate between multiple solutions in a protein-protein docking exercise.

2. Materials and Methods

Calculation of surface (S_m^{sc}) and electrostatic (E_m^{sc}) complementarity for buried and partially buried residues within proteins had been extensively discussed in the previous chapter. Two scoring functions (based on the amino acid identity (Res), burial (Bur), E_m^{sc} and S_m^{sc}) were formulated in order to identify the native fold amidst a set of decoys. Residues that were completely ($0.00 \leq \text{Bur} \leq 0.05$) or partially buried ($0.05 < \text{Bur} \leq 0.3$) were only considered. Initially the average and standard deviation for both S_m^{sc} ($\overline{S_m^{sc}}$, σ_S) and E_m^{sc} ($\overline{E_m^{sc}}$, σ_E) were estimated (over their respective databases **DB1** &

DB2), separately for different amino acid residues (Ala, Val etc.) distributed into three bins based on their burial (bin1: 0.0 $\leq$ Bur $\leq$ 0.05; bin2: 0.05 < Bur $\leq$ 0.15; bin3: 0.15 < Bur $\leq$ 0.30). The center (mode: E_0^{sc}) and the half width at half maximum height (γ_E) were also computed for individual residues (in different burial bins) from the normalized frequency distributions in E_m^{sc} by numerical curve fitting. For the first measure, $S_m^{sc}(i)$, $E_m^{sc}(i)$ were computed for all buried residues (i = 1….N; Bur $\leq$ 0.30) of a given polypeptide chain and the following expression was calculated:

$$CS_{gl} = \frac{1}{N}\sum_{\substack{i=1,\\ Bur\leq 0.3}}^{N}\left(\frac{1}{\sqrt{2\pi}\sigma_S}\exp\left(-\frac{1}{2}\left(\frac{S_m^{sc}(i)-\overline{S_m^{sc}}}{\sigma_S}\right)^2\right)\right)\cdot\left(\frac{1}{\pi}\left(\frac{\gamma_E}{(E_m^{sc}(i)-E_0^{sc})^2+\gamma_E{}^2}\right)\right)$$

$$= \frac{1}{N}\sum_{\substack{i=1,\\ Bur\leq 0.3}}^{N} Gaussian\,(S_m^{sc}(i))\cdot Lorentzian\,(E_m^{sc}(i)) \qquad (1)$$

The second scoring function was based on the conditional probability distributions of E_m^{sc} and S_m^{sc} for each residue type within a particular burial bin. As in the previous case three burial bins were considered. Distributions of E_m^{sc} and S_m^{sc} for a given residue type in a particular burial bin were then divided into intervals of 0.05. Conditional probability distributions of E_m^{sc} and S_m^{sc} were then defined as:

$$P(C_m^{sc}(i)\,|\,\{\mathrm{Res(i)}, Bur(i)\}) = \frac{N(C_m^{sc}(i)\cap \mathrm{Res(i)}\cap Bur(i))}{N(\mathrm{Res(i)}\cap Bur(i))} \qquad (2)$$

for the ith residue along the polypeptide chain where C_m^{sc} stands for either E_m^{sc} or S_m^{sc} and *N* denotes the count of residues in the specified sets.

Thus, for example,

$$P(S_m^{sc}: 0.45-0.5 \mid \{Valine, Bur: 0.0-0.05\}) = \frac{N\,(Valine \cap (0.0 \le Bur \le 0.05) \cap (0.45 < S_m^{sc} \le 0.5))}{N\,(Valine \cap (0.0 \le Bur \le 0.05))}$$

For any given polypeptide chain, the product of the conditional probabilities in S_m^{sc} and E_m^{sc} for each (i^{th}) residue (i = 1…N, Bur ≤ 0.30) were then summed and divided by the total number of buried residues (N) giving rise to the following measure:

$$CS_{cp} = \frac{1}{N} \sum_{\substack{i=1, \\ Bur \le 0.3}}^{N} P(S_m^{sc}(i) \mid \{\mathrm{Res}(i), \mathrm{Bur(i)}\}) \cdot P(E_m^{sc}(i) \mid \{\mathrm{Res}(i), \mathrm{Bur(i)}\}) \tag{3}$$

Z-scores corresponding to the native structure (along with its rank) for the complementarity scores (CS_{gl}, CS_{cp}) were calculated in a multiple decoy set by the following equation:

$$Z_{CS} = \frac{CS_{native} - \overline{CS}}{\sigma} \tag{4}$$

where CS_{native} is the score obtained for the parameter CS_{gl} or CS_{cp} from the native structure and $\overline{CS}$ and σ are the mean and standard deviations for the scores in the decoy set. Average Z_{CS} (<Z>) was calculated for the successful hits (native at rank 1) in a decoy set.

3. Results and Discussion

3.1. Application of S_m and E_m in protein fold recognition

As described in the earlier section, two scoring functions were designed based on the combined use of the complementarity measures obtained for different residues distributed in the aforementioned burial bins. Plots of the normalized frequency distributions in S_m^{sc}, E_m^{sc} for the individual residues in each burial bin (i.e., $P(S_m^{sc} \mid \{\mathrm{Res, Bur}\})$, $P(E_m^{sc} \mid \{\mathrm{Res, Bur}\})$) gave characteristic curves (symmetric for S_m^{sc}

and negatively skewed for E_m^{sc}), which fitted best to Gaussian and Lorentzian functions for S_m^{sc} and E_m^{sc} respectively (goodness of fit, $R^2 \geq 0.85$ for all cases, **Figure. 1**). From these observations, the first scoring function (CS_{gl}) was designed based on Gaussian for S_m^{sc} and Lorentzian for E_m^{sc} (see **Eq. 3.**). The second function (CS_{cp}) directly multiplies the conditional probabilities $P(S_m^{sc} \mid \{\text{Res, Bur}\})$ and $P(E_m^{sc} \mid \{\text{Res, Bur}\})$ for each residue along the polypeptide chain to obtain the joint probability of their co-occurrence. These individual probabilities were averaged over all buried residues (Bur $\leq$ 0.3) in the polypeptide chain to give the final score (see **Eq. 5.**). The conditional probabilities had been estimated previously (see **Materials and Methods)**.

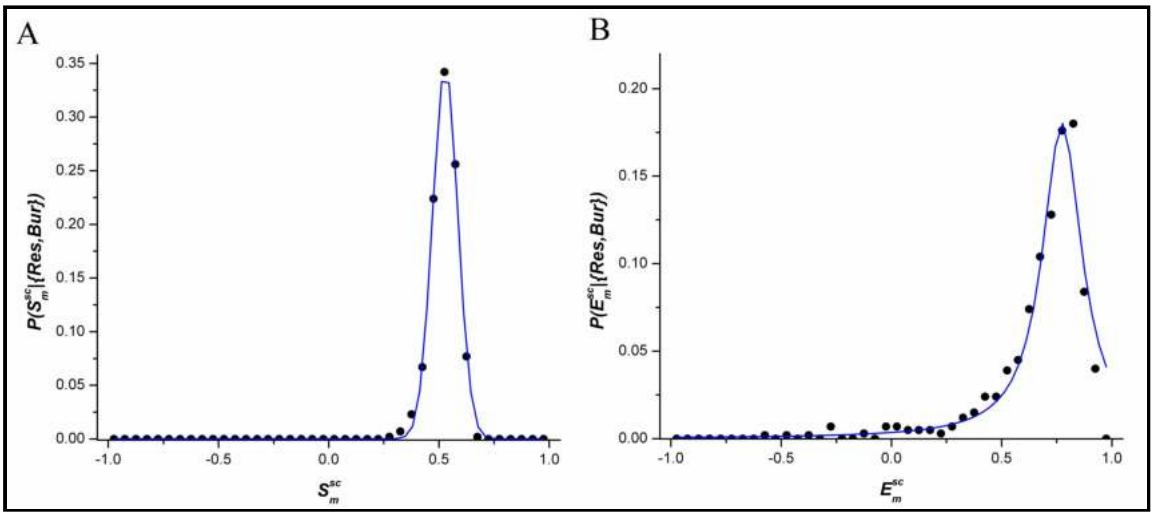


Figure 1. Normalized frequency distributions of S_m^{sc}, E_m^{sc} gives characteristic curves which fit best to Gaussian, Lorentzian functions respectively. These normalized frequencies for a given burial bin (Bur) and residue type (Res) can also be interpreted as conditional probabilities $P(S_m^{sc} \mid \{\text{Res, Bur}\})$ and $P(E_m^{sc} \mid \{\text{Res, Bur}\})$. **(A)** the distribution in S_m^{sc} for leucine ($0.0 \leq \text{Bur} \leq 0.05$) fitted to a Gaussian function ($R^2 = 0.997$) and **(B)** the distribution in E_m^{sc} for asparagine (same burial) fitted to a Lorentzian function ($R^2 = 0.948$). Similar curves were obtained for all completely / partially buried amino acids for all three burial bins.

It is to be noted that both CS_{gl} and CS_{cp} are averages of individual scores given by all the completely / partially buried residues in a protein and thus are independent of the polypeptide chain length. Thus, for any given native structure, one would expect their values to cluster around optimum numbers characteristic of native folds. The distributions of CS_{gl} and CS_{cp} computed for the native folds (in **DB2**) had a very good linear correlation between themselves (R^2 = 0.94**, Figure 2**) and gave mean values of 3.7 (± 0.437) and 0.015 (± 0.0017) respectively. Thus for the native folds, these functions do exhibit a reduced scatter about the mean, whereas for decoys, reduced scores for both the functions are to be expected. The decoy sets used to benchmark and validate the scoring functions included both single and multiple decoys, with Z–scores being calculated for the latter (see **Eq. 6.**). Since both the knowledge based scoring functions were parameterized on crystal structures alone, NMR structures were excluded in their validation.

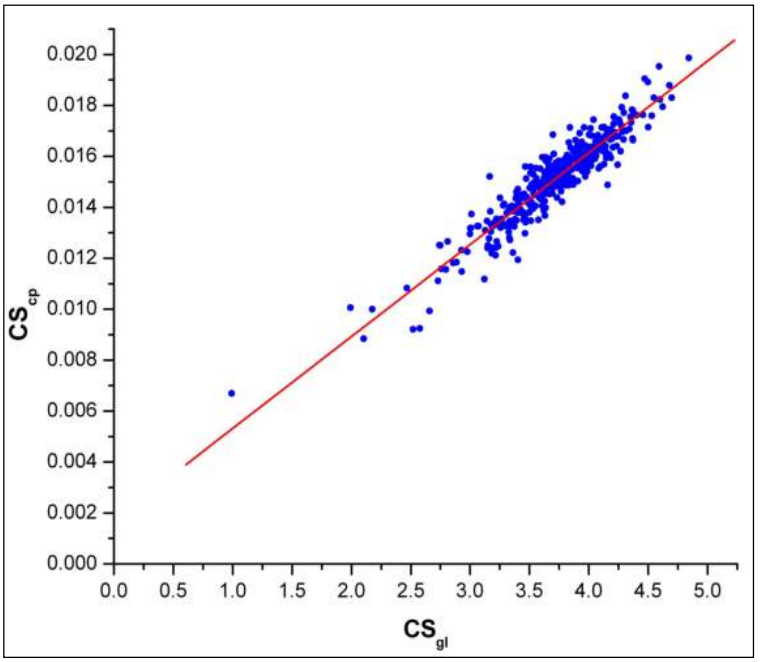


Figure 2. The two Complementarity Scores are linearly correlated. The figure shows the plot of CS_{gl} versus CS_{cp} computed for the 400 native folds (in **DB2**) linearly fitting to each-other (R^2 = 0.94).

3.2. Identification of the native crystal structure from decoys

Among the single decoy sets tested, 'Misfold' **(Holm and Sander, 1992)** consists of 26 pairs of structures. In each pair, the native sequence is threaded onto an unrelated fold to generate the decoy. 25 pairs were considered in the calculation (with the exception of 1CBH which is an NMR structure). The 'Pdberr' decoy set **(Branden and Jones, 1990)** consists of three correctly solved X-ray crystal structures along with their erroneous decoy counterparts, whereas 'sgpa' **(Avbelj et al., 1990)** contains the experimental structure of *Streptomyces griseus* Protease A (2SGA) and its two corresponding decoys, generated by molecular dynamics simulations. For the three data sets both functions successfully identified the native structure from their corresponding decoys for all cases (**Table 1)**. Comparison with other knowledge based scoring functions (**Table 2)** shows the performance of the complementarity scores in single decoy sets to be as efficient as or better than the other functions.

Table 1. Performance of CS_{gl} and CS_{cp} in single decoy sets. Scores obtained by native and decoy structures have been tabulated for both functions in the decoy sets **(A)** Misfold **(Holm and Sander, 1992)** and **(B)** Pdberr **(Branden and Jones, 1990)** and sgpa **(Avbelj et al., 1990)**. For **(A)** Misfold, sequence of a native structure (e.g., 1BP2: 1st row) have been threaded onto the template from an unrelated fold (2PAZ) to generate the corresponding decoy (1BP2on2PAZ). In **(B)** MDC1 and MDC2 refer to the two molecular dynamic simulation snapshots of 2SGA. The decoy sets have been downloaded from the database '**Decoys 'R' Us**': [http://dd.compbio.washington.edu/].

(A).

Native	Decoy	Resolution / R-factor	CS_{gl}		CS_{cp}	
			Native	decoy	native	decoy
1BP2	1BP2on2PAZ	1.7/0.17	2.64	0.88	0.0113	0.0039
1FDX	1FDXon 5RXN	2.0/0.19	1.91	1.58	0.0071	0.0065
1HIP	1HIPon 2B5C	2.0/0.24	2.12	0.58	0.0097	0.0024
1LH1	1LH1on 2I1B	2.0/0.00	2.55	1.05	0.0111	0.0044
1P2P	1P2Pon1RN3	2.6/0.24	2.08	1.35	0.0085	0.0047
1PPT	1PPTon1CBH	1.4/0.00	2.37	0.94	0.0103	0.0033
1REI	1REIon5PAD	2.0/0.24	2.63	0.84	0.0108	0.0038
1RHD	1RHDon2CYP	2.5/0.00	1.58	0.84	0.0066	0.0035
1RN3	1RN3on1P2P	1.5/0.22	2.86	1.15	0.0120	0.0047
1SN3	1SN3on2CI2	1.2/0.19	2.13	0.60	0.0085	0.0032
1SN3	1SN3on2CRO	1.2/0.19	2.13	0.91	0.0085	0.0034
2B5C	2B5Con1HIP	2.0/0.16	3.48	0.97	0.0149	0.0034
2CDV	2CDVon2SSI	1.8/0.18	1.13	0.95	0.0047	0.0033
2CI2	2CI2on1SN3	2.0/0.20	4.09	1.05	0.0174	0.0045
2CI2	2CI2on2CRO	2.0/0.20	4.09	0.81	0.0174	0.0032
2CRO	2CROon1SN3	2.4/0.20	3.49	0.82	0.0130	0.0032
2CRO	2CROon2CI2	2.4/0.20	3.49	0.92	0.0130	0.0038
2CYP	2CYPon1RHD	1.7/0.22	2.84	0.84	0.0115	0.0033
2I1B	2I1Bon1LH1	2.0/0.17	2.90	0.96	0.0119	0.0034
2PAZ	2PAZon1BP2	1.6/0.18	3.21	1.42	0.0137	0.0053
2SSI	2SSIon2CDV	2.3/0.19	1.04	0.86	0.0046	0.0036
2TMN	2TMNon2TS1	1.6/0.18	2.71	0.96	0.0107	0.0037
2TS1	2TS1on2TMN	2.3/0.23	2.74	1.02	0.0114	0.0039
5PAD	5PADon1REI	2.8/0.00	2.26	1.06	0.0093	0.0042
5RXN	5RXNon1FDX	1.2/0.14	2.34	1.31	0.0117	0.0048

(B).

Native	Decoy	Chain length	CS_{gl}		CS_{cp}	
			Native	decoy	native	decoy
2F19	1F19	435	2.172	0.617	0.009	0.003
3HFL	2HFL	556	2.329	1.805	0.009	0.007
5FD1	2FD1	106	2.951	0.464	0.011	0.002
2SGA	MDC1	181	2.546	1.952	0.011	0.008
2SGA	MDC2	181	2.546	1.754	0.011	0.007

Table 2. Comparison in the performances of different knowledge-based scoring functions on single decoy sets. The functions include R_s, R_p **(Bahadur and Chakrabarti, 2009)**, RAPD, CDF **(Samudrala and Moult, 1998)**, Surfield **(Arab et al., 2011),** Atomic Knowledge Based Potential (AKBP) **(Lu and Skolnick, 2001)**, Residue Contact Potential (RCP) **(Skolnick et al., 2000)** along with the complementarity scores (CS_{gl}, CS_{cp}) developed in this study. The number of successful hits / total number of trials are tabulated.

Scoring Functions	Misfold	Pdberr and sgpa
R_s	24/24	5/5
R_p	20/24	5/5
RAPD	24/24	5/5
CDF	19/24	5/5
Surfield	23/23	-
AKBP	24/24	5/5
RCP	24/24	4/5
CS_{gl}	25/25	5/5
CS_{cp}	25/25	5/5

The '4-state reduced' decoy set **(Park and Levitt, 1996)** consists of 7 sequences (chain length ranging from 54-75 residues), each with nearly 600-700 decoys that include structures with RMSD (C^{α} atoms) ranging from 0.8 to 9.4 Å from the native. Out of the 7 sequences, 6 native structures were correctly identified (rank 1) by CS_{gl}, CS_{cp} with significant Z-scores **(Table 3.A)**. In the case of 4RXN (all β class), the native structure was found to be at ranks 10, 15 respectively for CS_{gl}, $CS_{cp.}$ Further investigation revealed that 4RXN has negligible side chain packing between its secondary structural elements.

The decoy set, 'Fisa' **(Simons et al., 1997)** contains 4 small (43-76 residues) all-α proteins with 500 decoys for each set. Major failures were encountered for this decoy set where both CS_{gl} and CS_{cp} were successful in detecting the native at the top-rank in two out of the four proteins **(Table 3.B).** 1HDD-C was detected at rank 4 (CS_{gl}) and 5 (CS_{cp}), however, for 1FC2, both the functions failed entirely, leading to insignificant or negative Z-scores. This was due to minimal packing between their helices for both these low resolution structures (2.8Å). It is notable **(Table 4)** that for 1HDD-C, 1FC2 and 4RXN failure is quite common even for the other functions.

Table 3. Performance of CS_{gl} and CS_{cp} in multiple decoy sets of small proteins. Results tabulated for the decoy sets **(A)** 4-state reduced **(Park and Levitt, 1996)**, **(B)** Fisa **(Simons et al., 1997)**. **Resol / R** stands for resolution / R-factor of the native crystal structures whereas $\mathbf{N_{dec}}$ refers to the number of decoys. For **(B)** Fisa, there are 500 decoys for each native structure respectively. All proteins in Fisa belong to the all α class. $\mathbf{Z_{CS}}$ denotes the native Z-scores for the corresponding functions ($\boldsymbol{CS_{gl}}$/$\boldsymbol{CS_{cp}}$).

(A).

PDB ID	Length (aa)	class	Resol (Å) / R	N_{dec}	RMSD (Å) range of decoys	CS_{gl}		CS_{cp}	
						Rank	Z_{CS}	Rank	Z_{CS}
1CTF	68	α+β	1.70/0.17	630	1.3 - 9.1	1	7.9	1	7.1
1R69	63	All α	2.0/0.19	675	0.9 - 8.3	1	6.4	1	6.1
1SN3	65	α+β	1.2/0.19	660	1.3 - 9.1	1	5.6	1	4.7
2CRO	65	All α	2.4/0.20	674	0.8 - 8.3	1	5.9	1	5.0
3ICB	75	All α	2.3/0.18	653	0.9 - 9.4	1	6.7	1	4.9
4PTI	58	α+β	1.5/0.16	687	1.4 - 9.3	1	3.6	1	3.9
4RXN	54	All β	1.2/0.13	677	1.4 - 8.1	10	2.7	15	2.3

(B).

PDB ID	Length (aa)	Resol (Å) / R	RMSD (Å) range of decoys	CS_{gl}		CS_{cp}	
				Rank	Z_{CS}	Rank	Z_{CS}
1FC2	43	2.8/0.22	3.1 - 10.5	206	0.2	293	-0.2
1HDD-C	57	2.8/0.24	2.8 - 12.9	4	3.5	5	3.0
2CRO	65	2.4/0.20	4.3 - 12.6	1	7.2	1	6.3
4ICB	76	1.6/0.19	4.8 - 14.1	1	5.8	1	5.1

Table 4. Comparison in the performances of different knowledge based scoring functions on multiple decoy sets. The functions include DFIRE **(Zhang et al., 2004)**, Rosetta **(Misura et al., 2006)**, ModPipe-Pair (MPP), ModPipe-Surf (MPS) **(Melo et al., 2002)**, TE13, LHL **(Li et al., 2003)**, Force Model (FM) **(Mirzaie et al., 2009)**, DOPE **(Shen and Sali, 2006)**, MJ **(Miyazawa and Jernigan, 1996),** Surfield **(Arab et al., 2011),** R_s, R_p **(Bahadur and Chakrabarti, 2009)** along with the complementarity scores (CS_{gl}, CS_{cp}). All entries in the table refer to the rank of the native structure as detected by the corresponding method.

Decoy Set	PDB ID	DFIRE	Rosetta	MPP	MPS	TE13	LHL	FM	DOPE	MJ	Surfield	R_s	R_p	CS_{gl}	CS_{cp}
4state reduced	1CTF	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1R69	1	2	1	17	1	1	8	1	1	1	1	19	1	1
	1SN3	1	1	1	7	6	1	23	1	2	1	5	23	1	1
	2CRO	1	5	1	103	1	1	4	1	1	1	1	1	1	1
	3ICB	4	6	15	33	-	5	2	1	-	1	1	6	1	1
	4PTI	1	1	1	71	7	1	13	1	3	1	1	1	1	1
	4RXN	1	1	1	18	16	51	85	1	1	1	1	1	10	15
Fisa	1FC2	254	158	491	1	-	-	1	357	-	1	-	-	206	293
	1HDD	1	90	293	18	-	-	1	1	-	1	-	-	4	5
	2CRO	1	26	11	146	-	-	1	1	-	1	-	-	1	1
	4ICB	1	1	196	2	-	-	1	1	-	1	-	-	1	1

'Hg_structal' is a decoy set composed of 29 globins **(Samudrala and Levitt, 2000)** where each globin has been built by comparative modeling using 29 other globins as templates with their C^{α} RMSD's ranging from 1.96 to 8.57 Å. Thus, for each native globin chain there are 29 decoys. In 23 out of 29 globins, both CS_{gl} and CS_{cp} were able to correctly detect the native at the top rank (<Z>: 3.23, 3.24 respectively, **Table 5.A**). For similar decoy sets, 'ig_structal' CS_{gl} and CS_{cp} were successful in 48 and 50 cases (<Z>: 3.89, 3.91) out of 61 immunoglobulins, whereas for 'ig_structal_hires' (subset of 20 high resolution structures), 100 % success was achieved for both (**Table 5.B, C**).

Table 5. Performance of CS_{gl} and CS_{cp} in decoy sets composed by homology modeling. Results tabulated for the decoy sets **(A)** Hg_structal **(Samudrala and Levitt, 2000)**, **(B)** Ig_structal **(Samudrala and Levitt, 2000)** and **(C)** Ig_structal_hires **(Samudrala and Levitt, 2000)**. **Resol / R** stands for resolution / R-factor of the native crystal structure whereas $\mathbf{N_{dec}}$ refers to the number of decoys. For **(A)** Hg_structal, **(B)** Ig_structal and **(C)** Ig_structal_hires there are 29, 61 and 19 decoys for each native structure respectively. **(A)** Hg_structal is constituted of globins whereas **(B)** Ig_structal and **(C)** Ig_structal_hires contains immunoglobulins. $\mathbf{Z_{CS}}$ denotes the native Z-scores for the corresponding function ($\boldsymbol{CS_{gl}}$ /$\boldsymbol{CS_{cp}}$). The decoy sets have been downloaded from the database '**Decoys 'R' Us**': [http://dd.compbio.washington.edu/].

(A).

PDB ID	Length (aa)	Resol (Å) / R	RMSD (Å) range of decoys	CS_{gl}		CS_{cp}	
				Rank	Z_{CS}	Rank	Z_{CS}
1ASH	147	2.2/0.18	2.222 - 6.947	1	4.1	1	3.6
1BAB-B	146	1.5/0.16	0.702 - 6.920	1	3.4	1	3.1
1COL-A	197	2.4/0.18	12.399 - 30.284	1	4.8	1	4.6
1CPC-A	162	1.7/0.18	6.835 - 13.957	1	4.6	1	4.5
1ECD	136	1.4/0.00	1.471 - 6.188	1	3.8	1	4.1
1EMY	153	1.8/0.15	0.735 - 9.281	1	2.9	1	2.8
1FLP	142	1.5/0.17	1.734 - 7.227	1	4.1	1	4.1
1GDM	153	1.7/0.16	2.609 - 8.371	1	3.2	1	3.3
1HBG	147	1.5/0.15	2.050 - 6.896	1	3.8	1	4.1
1HBH-A	142	2.2/0.16	0.958 - 6.347	1	1.9	1	2.3
1HBH-B	146	2.2/0.16	1.024 - 7.330	1	2.3	1	2.4
1HDA-A	141	2.2/0.19	0.487 - 5.794	1	2.6	1	2.7
1HDA-B	145	2.2/0.19	0.545 - 5.644	1	2.9	1	2.9
1HLB	157	2.5/0.15	2.891 - 7.001	1	1.6	1	2.4
1HLM	158	2.9/0.19	2.973 - 8.737	20	-0.5	17	-0.3
1HSY	153	1.9/0.16	0.795 - 9.681	1	2.6	1	2.2
1ITH-A	141	2.5/0.15	1.638 - 6.071	1	4.5	1	4.5
1LHT	153	2.0/0.18	0.814 - 9.736	1	2.5	1	2.4
1MBA	146	1.6/0.19	1.829 - 7.314	1	3.9	1	3.8
1MBS	153	2.5/0.00	1.698 - 9.304	29	-1.2	30	-1.3
1MYG-A	153	1.8/0.20	0.479 - 9.562	1	2.6	1	2.5
1MYJ-A	153	1.9/0.21	0.623 - 7.944	1	2.9	1	3.0
1MYT	146	1.7/0.18	1.014 - 10.043	1	3.0	1	3.0
2DHB-A	141	2.8/0.00	0.648 - 6.358	14	-0.1	13	0.2
2DHB-B	146	2.8/0.00	0.858 - 7.062	13	0.2	12	0.3
2LHB	149	2.0/0.14	3.022 - 8.080	1	2.9	1	3.0
2PGH-A	141	2.8/0.15	0.707 - 6.485	16	-0.1	14	-0.1
2PGH-B	146	2.8/0.15	0.769 - 7.479	11	0.4	14	0.2
4SDH-A	145	1.6/0.16	2.273 - 6.429	1	3.4	1	3.3

(B).

PDB ID	Length (aa)	Resol (Å) / R	RMSD (Å) range of decoys	CS_{gl}		CS_{cp}	
				Rank	Z_{CS}	Rank	Z_{CS}
1ACY	232	3.0/0.21	1.167 - 4.455	1	5.7	1	5.7
1BAF	222	2.9/0.20	1.123 - 3.973	1	3.8	1	4.0
1BBD	231	2.8/0.19	1.610 - 4.537	1	3.5	1	4.2
1BBJ	221	3.1/0.18	0.797 - 4.047	1	2.3	1	2.3
1DBB	231	2.7/0.21	0.993 - 4.203	3	1.9	1	2.2
1DFB	231	2.7/0.18	1.383 - 4.826	2	1.8	3	1.8
1DVF	223	1.9/0.19	0.706 - 4.176	1	5.2	1	5.2
1EAP	225	2.5/0.19	1.557 - 4.537	1	4.9	1	4.7
1FAI	231	2.7/0.19	1.258 - 4.619	1	3.3	1	3.6
1FBI	229	3.0/0.19	1.322 - 4.993	2	2.5	1	1.9
1FGV	227	1.9/0.18	0.977 - 4.502	1	3.7	1	3.7
1FIG	227	3.0/0.22	1.423 - 4.388	19	0.5	12	0.8
1FLR	228	1.9/0.19	1.090 - 4.284	1	3.7	1	3.8
1FOR	225	2.8/0.17	0.998 - 4.200	1	2.6	1	2.9
1FPT	231	3.0/0.23	0.891 - 4.332	7	1.4	9	0.9
1FRG	233	2.8/0.19	0.952 - 4.134	1	5.5	1	5.7
1FVC	229	2.2/0.18	1.505 - 4.970	1	3.8	1	4.1
1FVD	227	2.5/0.17	0.798 - 4.137	1	3.3	1	3.7
1GAF	221	2.0/0.24	0.841 - 4.034	1	3.6	1	2.9
1GGI	226	2.8/0.18	0.978 - 4.247	1	3.2	1	3.5
1GIG	231	2.3/0.19	1.424 - 4.497	1	5.6	1	5.7
1HIL	233	2.0/0.19	0.910 - 4.303	1	5.2	1	5.2
1HKL	221	2.7/0.18	0.762 - 3.996	6	1.5	9	1.2
1IAI	228	2.9/0.21	1.069 - 4.594	5	1.7	5	1.5
1IBG	232	2.7/0.20	1.483 - 4.713	1	3.7	1	3.7
1IGC	227	2.6/0.16	0.960 - 4.237	12	0.9	14	0.8
1IGF	231	2.8/0.18	0.927 - 4.241	1	3.5	1	3.7
1IGI	231	2.7/0.17	1.579 - 4.266	1	3.7	1	4.0
1IGM	227	2.3/0.20	1.077 - 4.146	1	2.0	1	2.5
1IKF	233	2.5/0.16	1.330 - 5.105	1	4.6	1	4.8
1IND	222	2.2/0.18	1.183 - 4.213	1	3.7	1	3.5
1JEL	230	2.8/0.19	1.032 - 4.193	1	3.9	1	3.8
1JHL	224	2.4/0.21	1.100 - 4.083	1	3.7	1	3.7
1KEM	231	2.2/0.18	1.386 - 4.671	1	2.2	1	2.2
1MAM	227	2.5/0.21	1.473 - 4.170	1	3.3	1	3.8
1MCP	235	2.7/0.22	1.186 - 4.520	1	4.5	1	4.4
1MFA	229	1.7/0.16	2.053 - 4.805	1	4.6	1	4.7
1MLB	223	2.1/0.18	0.899 - 4.186	1	3.4	1	3.9
1MRD	225	2.4/0.19	1.311 - 4.150	1	3.8	1	3.5
1NBV	232	2.0/0.24	1.205 - 4.237	1	2.6	1	2.4
1NCB	227	2.5/0.16	1.167 - 4.385	1	3.1	1	3.1
1NGQ	229	2.4/0.19	1.407 - 4.065	1	3.8	1	3.9
1NMB	231	2.5/0.21	1.496 - 5.568	1	3.6	1	3.5
1NSN	224	2.9/0.19	1.163 - 4.033	15	0.8	19	0.5

1OPG	232	2.0/0.16	1.379 - 4.795	8	1.1	8	1.1
1PLG	228	2.8/0.16	1.036 - 4.168	1	4.3	1	3.9
1RMF	231	2.8/0.18	1.563 - 4.457	7	1.3	8	0.8
1TET	228	2.3/0.14	0.768 - 4.119	1	4.0	1	4.3
1UCB	228	2.5/0.20	1.183 - 4.346	1	2.9	1	3.2
1VFA	224	1.8/0.15	1.135 - 4.080	1	4.7	1	4.6
1VGE	231	2.0/0.18	1.619 - 5.004	1	5.4	1	5.2
1YUH	225	3.0/0.19	1.265 - 4.292	11	1.1	30	0.3
2CGR	228	2.2/0.21	0.925 - 4.226	1	3.7	1	4.3
2FB4	236	1.9/0.18	1.667 - 6.135	1	4.7	1	4.9
2FBJ	224	2.0/0.19	1.078 - 4.204	1	4.3	1	4.1
2GFB	227	3.0/0.21	1.639 - 4.123	1	4.1	1	4.2
3HFL	223	2.6/0.19	1.512 - 5.056	1	2.7	1	3.6
3HFM	220	3.0/0.24	1.029 - 3.993	5	1.5	8	1.0
6FAB	228	1.9/0.20	1.193 - 4.636	1	4.3	1	3.5
7FAB	219	2.0/0.16	1.845 - 4.547	1	4.7	1	4.8
8FAB	228	1.8/0.17	1.913 - 6.816	1	4.4	1	4.9

(C).

PDB ID	**Length (aa)**	**Resol (Å) / R**	**RMSD (Å) range of decoys**	***CS_{gl}***		***CS_{cp}***	
				Rank	**Z_{CS}**	**Rank**	**Z_{CS}**
1DVF	223	1.9/0.19	0.706 - 4.176	1	3.4	1	3.4
1FGV	227	1.9/0.18	0.977 - 4.502	1	2.9	1	2.9
1FLR	228	1.9/0.19	1.090 - 4.284	1	3.0	1	3.0
1FVC	229	2.2/0.18	1.505 - 4.970	1	2.9	1	2.9
1GAF	221	2.0/0.24	0.841 - 4.034	1	3.1	1	2.8
1HIL	233	2.0/0.19	0.910 - 4.303	1	3.6	1	3.5
1IND	222	2.2/0.18	1.183 - 4.213	1	3.1	1	2.9
1KEM	231	2.2/0.18	1.386 - 4.671	1	2.1	1	2.1
1MFA	229	1.7/0.16	2.053 - 4.805	1	3.3	1	3.3
1MLB	223	2.1/0.18	0.899 - 4.186	1	2.6	1	2.8
1NBV	232	2.0/0.24	1.205 - 4.237	1	2.3	1	2.1
1OPG	232	2.0/0.16	1.379 - 4.795	1	1.1	1	1.2
1VFA	224	1.8/0.15	1.135 - 4.080	1	3.2	1	3.2
1VGE	231	2.0/0.18	1.619 - 5.004	1	3.5	1	3.3
2CGR	228	2.2/0.21	0.925 - 4.226	1	2.9	1	3.1
2FB4	236	1.9/0.18	1.667 - 6.135	1	3.3	1	3.4
2FBJ	224	2.0/0.19	1.078 - 4.204	1	3.1	1	3.1
6FAB	228	1.9/0.20	1.193 - 4.636	1	2.8	1	2.4
7FAB	219	2.0/0.16	1.845 - 4.547	1	3.3	1	3.4
8FAB	228	1.8/0.17	1.913 - 6.816	1	3.2	1	3.5

The ROSETTA all-atom decoy sets are built for small single domain proteins by the fragment insertion-simulated annealing strategy. The latest ROSETTA decoy set **(Tsai et al., 2003)** contains over 75000 decoys for 41 proteins (of which 25 are X-ray structures, the number of decoys in each set ranging from 1610 to 1934), sampling a wide variety of topological folds and polypeptide chain lengths ranging from 35 to 85 amino acids. CS_{gl}, CS_{cp} were able to rank the native in 23, 24 instances (out of 25). The high average Z-scores (7.24, 6.98) also demonstrate the discriminatory ability of both the scoring functions (**Table 6**). The only major failure was encountered for 1CC5 (detected at rank: 36, 58) which is a cytochrome C molecule with an embedded Fe^{+2} containing protoporphyrin IX ring. Since only protein atoms were considered, a false picture of interior atomic packing was available to the scoring functions.

Table 6. Performance of CS_{gl} and CS_{cp} in identifying the native crystal structures in the Rosetta all atom decoy set (20). N_{dec} and **Resol** stands for the number of decoys for each native structure and its crystallographic resolution. **RMSD** refers to the C^{α}-rms deviation of the decoy closest to the native in the set. $\mathbf{Z_{CS}}$ denotes the native Z-scores for the corresponding function ($\boldsymbol{CS_{gl}}$ /$\boldsymbol{CS_{cp}}$). The dataset was downloaded from: http://trimer.tamu.edu/~daniel/decoys_11-14-01.tar.gz.

PDB ID	Length (aa)	class	Resol (Å)	N_{dec}	RMSD (Å)	CS_{gl}		CS_{cp}	
						Rank	Z_{CS}	Rank	Z_{CS}
1A32	65	All α	2.1	1610	0.90	1	5.3	1	5.0
1AIL	67	All α	1.9	1807	1.97	1	6.1	1	6.4
1AM3	57	All α	1.7	1898	1.80	1	7.4	1	6.5
1BQ9	53	All β	1.2	1825	2.79	1	7.1	1	6.8
1CC5	76	All α	2.5	1892	4.31	36	1.6	58	1.3
1CEI	85	All α	1.8	1897	4.57	1	8.6	1	8.5
1CSP	64	All β	2.45	1809	3.24	1	8.5	1	7.2
1CTF	67	α|β	1.70	1922	2.66	1	9.7	1	9.1
1DOL	62	α+β	2.4	1871	3.76	1	7.1	1	6.8
1HYP	75	All α	1.8	1893	4.05	1	6.3	1	6.1
1LFB	69	All α	2.8	1893	2.47	1	5.2	1	6.1
1MSI	60	All β	1.25	1894	5.40	1	10.6	1	10.1
1MZM	71	All α	1.78	1934	2.69	1	6.1	1	5.8
1ORC	56	α+β	1.54	1883	2.81	1	8.3	1	9.6
1PGX	57	α+β	1.66	1851	1.48	1	7.3	1	7.3
1PTQ	43	All β	1.95	1885	5.42	1	7.1	1	7.1
1R69	61	All α	2.00	1733	1.38	1	6.0	1	5.7
1TIF	59	α+β	1.80	1849	2.60	1	8.5	1	9.1
1TUC	61	All β	2.02	1894	4.48	1	6.6	1	6.6
1UTG	62	All α	1.34	1897	3.36	1	7.1	1	4.7
1VCC	77	α+β	1.60	1857	3.85	1	8.1	1	9.0
1VIF	48	All β	1.80	1896	0.44	1	5.7	1	6.4
2FXB	81	α|β	0.92	1800	5.48	2	3.3	1	3.7
5ICB	72	All α	1.50	1870	2.98	1	8.0	1	7.8
5PTI	55	α+β	1.00	1853	3.94	1	5.9	1	6.1

CASP9 **(Moult et al., 2011)** is probably the most challenging test as the decoys are the best predicted near-native models submitted by different groups participating in the CASP experiment. CASP9 (conducted in July – August, 2010) consisted of 111 valid targets with 90 X-ray crystal structures. T0543 (2XRQ) and T0605 (3NMD) were not considered in the calculation, the former due to its excessively huge chain length (887 residues) and the latter being a single standalone helix. For the remaining 88 targets (with

a total of 9197 models, chain length ranging from 83 to 611 residues) CS_{gl} and CS_{cp} detected the native at the top rank in 70, 72 and 85, 86 within rank 5 (<Z>: 3.65, 3.95) respectively **(Table 7)**.

Table 7. Performance of CS_{gl} and CS_{cp} in CASP9 dataset. CASP9 experiment (conducted in July – August, 2010) consisted of 129 accepted targets out of which 18 were cancelled during the experiment. Of the remaining 111, 90 were X-ray crystal structures. T0543-2XRQ and T0605-3NMD were not considered in the calculation, the former due to its huge chain length (887 residues) and the latter being a single standalone helix. Results are tabulated for all other valid crystal structure targets. $\mathbf{N_{mod}}$ stands for the number of 'first models' used in the calculations. $\mathbf{Resol/R_{obs}}$ represent the crystallographic resolution and R-factor (observed) of the native structure respectively. **RMSD** refers to the C^{α}-rms deviation of the model closest to the native in the set, calculated by Dali server **(Holm and Rosenstrom, 2010)**. $\mathbf{N_{res}}$ refers to the number of residues to be modeled for each target. **Rank** and $\mathbf{Z_{CS}}$ denotes the native rank and Z-scores for the corresponding function ($\boldsymbol{CS_{gl}}$ /$\boldsymbol{CS_{cp}}$). For the target T0602-3NKZ, none of the models was superposible to the native by Dali server **(Holm and Rosenstrom, 2010)** (RMSD: N/A).

Target ID	N_{res} (aa)	PDB ID	Resol (Å) / R_{obs}	N_{mod}	RMSD (Å)	CS_{gl}		CS_{cp}	
						Rank	Z_{CS}	Rank	Z_{CS}
T0515	365	3MT1	2.50, 0.194	130	2.2	1	6.0	1	6.1
T0516	229	3NO6	1.65, 0.167	89	2.0	1	3.8	2	3.8
T0517	159	3PNX	1.92, 0.191	127	1.6	1	4.0	1	4.0
T0518	288	3NMB	2.40, 0.172	79	1.5	1	3.8	1	4.0
T0520	189	3MR7	2.60, 0.184	144	2.0	3	2.6	3	2.6
T0521	179	3MSE	2.10, 0.224	85	1.3	1	4.2	1	4.5
T0522	134	3NRD	2.06, 0.170	87	0.8	1	3.6	1	3.4
T0523	120	3MQO	1.70, 0.226	131	1.1	2	3.5	2	3.4
T0524	325	3MWX	1.45, 0.149	82	1.8	1	5.4	1	5.5
T0525	215	3MQZ	1.30, 0.150	82	2.1	1	5.1	1	5.1
T0526	290	3NRE	1.59, 0.162	124	2.0	1	4.4	1	4.3
T0527	142	3MR0	2.35, 0.224	85	1.9	1	3.5	1	3.6
T0528	388	3N0X	1.50, 0.149	89	2.5	1	5.8	1	5.7
T0529	569	3MWT	1.98, 0.183	50	1.3	1	2.4	1	2.4
T0530	115	3NPP	2.15, 0.181	83	1.6	1	3.8	2	3.5
T0532	506	3MX3	2.00, 0.167	50	1.9	1	3.0	1	3.0
T0534	384	3N8U	1.44, 0.168	126	2.7	1	3.7	1	3.7
T0537	381	3N6Z	1.30, 0.167	145	2.6	1	5.7	1	5.7
T0540	90	3MX7	1.76, 0.178	138	2.3	1	5.2	1	4.9
T0542	590	3N05	2.35, 0.233	50	1.5	1	2.7	1	2.7

T0547	611	3NZP	3.00, 0.183	50	2.8	1	2.0	1	2.0
T0548	106	3NNQ	2.69, 0.234	86	3.0	1	2.9	1	2.8
T0550	339	3NQK	2.61, 0.200	107	2.7	1	5.3	1	5.3
T0558	294	3NO2	1.35, 0.147	129	2.2	1	6.7	1	6.7
T0561	161	2XSE	1.90, 0.176	129	2.9	1	2.7	1	3.0
T0563	279	3ON7	2.20, 0.187	88	1.9	1	5.1	1	5.2
T0565	326	3NPF	1.72, 0.142	87	1.8	1	5.8	1	5.7
T0566	156	3N72	1.77, 0.191	140	2.1	1	3.2	1	3.7
T0567	145	3N70	2.80, 0.233	91	1.8	3	1.3	5	1.2
T0568	158	3N6Y	1.50, 0.182	114	2.5	1	3.8	1	4.1
T0570	258	3N70	2.80, 0.233	91	1.6	1	4.7	1	4.9
T0571	344	3N91	2.40, 0.183	108	2.3	1	3.9	1	3.8
T0573	311	3OOX	1.44, 0.170	92	1.8	1	4.0	1	4.2
T0574	126	3NRF	1.50, 0.186	113	2.2	1	4.4	1	4.4
T0575	216	3NRG	2.56, 0.209	93	1.6	3	2.7	1	2.8
T0576	172	3NA2	2.29, 0.190	110	2.1	1	5.9	1	6.0
T0578	164	3NAT	2.92, 0.206	107	2.1	1	3.4	1	3.4
T0580	105	3NBM	1.30, 0.134	141	1.3	1	3.1	1	3.0
T0581	136	3NPD	1.60, 0.162	93	1.8	1	3.3	1	3.3
T0582	222	3O14	1.70, 0.154	128	1.9	1	3.9	1	4.0
T0584	352	3NF2	2.20, 0.233	145	1.7	1	2.9	1	2.8
T0585	234	3NE8	1.24, 0.161	94	1.6	3	2.7	4	2.6
T0586	125	3NEU	1.58, 0.195	145	1.0	1	3.3	1	3.3
T0588	400	3NFV	1.95, 0.158	138	2.8	1	4.5	1	4.5
T0589	465	3NET	2.70, 0.219	87	1.8	1	3.8	1	3.8
T0591	406	3NRA	2.15, 0.156	88	1.8	1	3.8	1	3.8
T0592	144	3NHV	2.50, 0.199	141	1.3	1	3.1	1	2.8
T0593	208	3NGW	2.31, 0.187	88	2.2	1	3.8	1	3.9
T0594	140	3NI8	2.50, 0.226	141	1.4	1	3.3	1	3.4
T0596	213	3NI7	2.78, 0.250	143	1.3	7	1.7	7	1.8
T0597	429	3NIE	2.30, 0.225	86	1.9	1	2.8	1	2.7
T0598	161	3NJC	1.69, 0.188	135	2.5	7	1.5	5	1.4
T0599	399	3OS6	2.40, 0.173	89	1.3	1	3.5	1	3.5
T0601	449	3QTD	2.70, 0.227	82	1.4	1	3.4	1	3.4
T0602	123	3NKZ	2.11, 0.158	148	**N/A**	4	1.6	5	1.5
T0603	305	3NKD	1.95, 0.220	90	1.6	2	3.8	1	3.9
T0604	549	3NLC	2.15, 0.218	50	1.5	1	2.5	1	2.5
T0606	169	3NOH	1.60, 0.161	123	1.7	1	4.4	1	4.7
T0607	471	3PFE	1.50, 0.139	88	2.5	1	2.9	1	2.9
T0608	279	3NYY	1.60, 0.150	134	2.1	1	4.6	1	4.7
T0609	340	3OS7	1.80, 0.145	87	2.1	1	3.8	1	3.8
T0610	186	3OT2	1.96, 0.217	138	1.8	1	4.6	1	4.7
T0611	227	3NNR	2.49, 0.232	90	2.1	1	3.2	1	3.2
T0613	287	3OBI	1.95, 0.236	91	1.1	2	2.2	2	2.2
T0615	179	3NQW	2.90, 0.213	92	1.9	2	2.3	3	2.2
T0616	103	3NRT	2.54, 0.201	104	3.2	2	2.8	2	2.6
T0617	148	3NRV	2.00, 0.198	91	1.7	1	3.2	1	3.2
T0618	182	3NRH	1.80, 0.198	133	2.9	1	2.9	1	2.8
T0619	111	3NRW	1.70, 0.187	148	1.1	1	3.4	1	3.6
T0620	312	3NR8	2.80, 0.216	85	1.2	1	3.9	1	4.1
T0621	172	3NKG	2.00, 0.172	65	2.9	1	3.5	1	3.4

T0622	138	3NKL	1.90, 0.169	151	1.6	2	2.6	2	2.3
T0623	220	3NKH	2.50, 0.158	88	1.8	1	3.7	1	3.9
T0624	81	3NRL	1.90, 0.219	96	2.2	2	2.6	1	2.9
T0625	233	3ORU	1.11, 0.127	123	1.9	1	3.6	1	3.5
T0626	283	3O1L	2.20, 0.175	91	1.2	1	5.0	1	5.3
T0627	261	3OQL	2.54, 0.178	142	2.1	10	0.7	10	0.8
T0628	295	3NUW	2,09, 0.174	142	2.8	1	4.1	1	4.1
T0629	216	2XGF	2.20, 0.182	95	2.1	1	5.1	1	4.8
T0632	168	3NWZ	2.57, 0.233	90	1.0	3	2.5	1	3.1
T0634	140	3N53	2.20, 0.252	91	1.3	4	2.4	4	2.3
T0635	191	3N1U	1.80, 0.178	85	0.6	2	3.3	2	3.6
T0636	336	3P1T	2.60, 0.202	87	1.5	1	5.5	1	5.5
T0638	269	3NXH	2.58, 0.247	90	1.6	1	3.5	1	3.3
T0639	128	3NYM	1.90, 0.175	81	1.3	1	3.9	1	3.8
T0640	250	3NYW	2.16, 0.236	90	2.0	1	4.6	1	4.6
T0641	296	3NYI	1.90, 0.171	89	1.8	1	4.2	1	4.2
T0643	83	3NZL	1.20, 0.150	134	2.1	1	3.2	1	3.4

3.3. Discrimination between good and bad RMSD models

In order to test the sensitivity of the functions with respect to deviations from the experimentally determined coordinates of the side chain atoms, 10 native (top ranked) targets from CASP9 along with their corresponding models were selected. Subsequent to superposition of the models onto the native structure by Dali server **(Holm and Rosenstrom, 2010)**, the rms deviation of the side chain atoms were calculated at one-to-one atomic correspondence w.r.t. the native. Local Deviations (in C^{α}) greater than 10 Å were considered to be so large, as to lose all structural relationship with the corresponding region of the native, as also models which were non–superposable (by Dali) and these were thus not included in the calculation. CS_{gl}, CS_{cp} of the native structure and ~ 60 models per target were then plotted **(Figure 3)** as a function of their RMSD's (ranging from ~ 1.5 to 10 Å). Although the scores generally fell with increase in RMSD especially in the range of 1.5 – 5 Å, there was substantial scatter amongst the points which belied the expectation of obtaining a functional relationship between the two variables. However, as these RMSD's contain contributions from both main and side chain deviations, a second calculation (with 10 structures: **Figure 3**) was performed, wherein the backbone coordinates were held fixed with errors being incorporated in the side chain conformations by three distinct methods: a) randomizing the side-chain χ

angles (50 erroneous models) **(Basu et al., 2011)**, b) the same 50 models as in (a) subjected to an energy minimization protocol (using CHARMM **(Brooks et al., 1983)**) described previously **(Basu et al., 2011)** and c) an unique solution determined by SCWRL4.0 **(Krivov et al., 2009)** upon threading. Two distinct clusters were obtained for (a) and (b) with energy minimization significantly improving scores in (b) relative to (a). The models derived from SCWRL4.0 (c) generally gave values closest to the native **(Figure 3)** while rarely a few structures from (b) gave similar / slightly better scores than (c). Thus, the scores indeed reflect errors in side chain coordinates as estimated by RMSD (w.r.t. to native) and generally drop with increase in error.

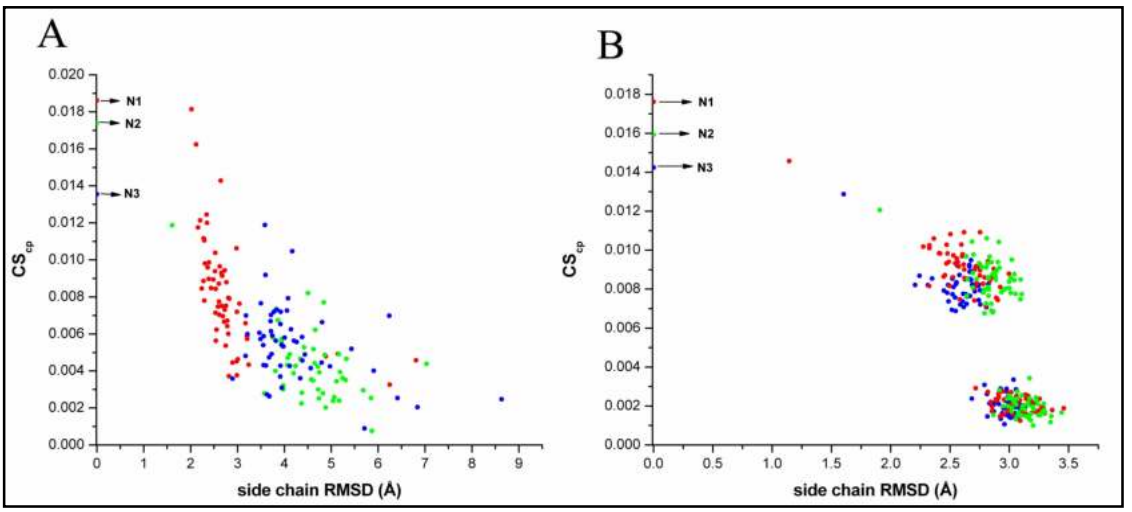


Figure 3. Complementarity Scores drop with increased errors in side chain coordinates. (A) CS_{cp} values as a function of side chain RMSD's for 3 CASP9 targets (native and models). N1, N2, N3 correspond to the native crystal structures of T0522 (3NRD), T0623 (3NKH) and T0586 (3NEU) and their corresponding models plotted in red, green and blue. **(B)** CS_{cp} values as a function of side chain RMSD's for 3 globular proteins and their models (see Text). N1, N2, N3 and S1, S2, S3 correspond to the native structures and the unique solutions generated by SCWRL4.0 respectively for 2OEB (red), 3COU (green) and 2HAQ (blue). The two distinct clusters are for structures produced by randomization of the side chain conformers (with lower values) and energy minimization of the same set of randomized conformers (higher). Similar patterns have been obtained for CS_{gl}. Barring 2HAQ, all other structures are from **DB2.**

3.4. Fold recognition by cross-threading

The scoring functions were also tested for protein pairs belonging to the same fold though with low sequence identity upon alignment. 100 such pairs (sequence identities ranging from 6–30% : **Dataset S1** in **Supplementary Information** in CD enclosed) sampling diverse folds were selected from the PREFAB4.0 database **(Edgar, 2004)**. The sequence identities upon structural alignment for each pair were determined by Dali Server **(Holm and Rosenstrom, 2010)** and their folds assigned according to the SCOP database **(Murzin et al., 1995)**. For every pair, the two native sequences were aligned using CLUSTAL W **(Thompson et al., 1994)** and insertions in the sequence to be threaded onto the main chain (of its partner) were excised whereas deletions were padded with glycine, in order to maintain the correct position of the threaded residues consistent with the alignment. For the cross-threaded sequences, padded poly-glycine stretches at the N/C termini were also excised prior to the calculations. When the fold was part of a larger polypeptide chain (domain), two possibilities were considered. If the fold was found to be completely separated from the other domains in the chain, then it was considered in isolation for all subsequent calculations, whereas if the fold was found to be integrally embedded in the composite structure; the entire chain was used to calculate S_m^{sc}, E_m^{sc} and the relevant residues in the domain were then used to compute CS_{gl}, CS_{cp}. For all pairs, the native structures gave characteristic similar scores for both CS_{gl} and CS_{cp}. The two sequences were then cross-threaded onto the backbone of each other, with their side chain torsions being set to values, determined by SCWRL4.0 **(Krivov et al., 2009)**. For each such pair, 100 random sequences (≤ 15% identity between any two sequences in a set) were threaded onto each of the two corresponding templates following the same protocol. Hydrogen atoms were geometrically fixed by REDUCE **(Word et al., 1999)** in all models. In large majority of the cases, the average score of the two cross-threaded structures was found to be markedly lower than their native counterparts yet noticeably higher (Z >= 2 for 86, 87 pairs) than those obtained from the random decoys (<Z>: 3.43, 3.33 for CS_{gl}, CS_{cp} respectively). However, below 15 % sequence identity, there was a drop in the Z-scores (less than 1.5 for 5 out of 21

such pairs) primarily due to large mismatches in structural (Dali server) and sequence (CLUSTL W) alignments. In general, large variations were observed in the Z-scores (ranging from 0.4 to 8.0) for different folds.

4. Conclusion

The earlier chapter described that, fairly stringent constraints both in terms of shape and electrostatic complementarities are to be satisfied for interior residues of a correctly folded polypeptide chain similar to residues at the protein-protein interface. This was used to predict the native fold of a sequence. Both functions (CS_{gl}, CS_{cp}) based on the probability distributions in S_m^{sc} and E_m^{sc} performed successfully in state-of-the-art decoy sets. This could be considered analogous to protein-protein docking wherein both surface and electrostatic complementarities rise to their optimum values upon the interlocking of interacting protein molecules in their correct stereo-specific geometry of association. That is to say, folding can be envisaged as the 'docking' of interior residues to their respective native environments consistent with short and long range forces. The fact that the performance of both the functions were comparable to or better than the best scoring functions currently available in the literature, demonstrates the practical application of complementarity in the area of protein folding and structure prediction. The functions were also found to be useful in correctly identifying the same fold for two sequences with low sequence identity. Thus, indeed the concept of complementarity provides a common conceptual platform to discuss folding and binding.

References

Arab S, Sadeghi M, Eslahchi C, Pezeshk H, Sheari A (2010). **A pairwise residue contact area-based mean force potential for discrimination of native protein structure.** *BMC Bioinformatics.* **11**: 16.

Avbelj F, Moult J, Kitson DH, James MN, Hagler AT (1990). **Molecular dynamics study of the structure and dynamics of a protein molecule in a crystalline ionic environment: Streptomyces griseus protease A.** *Biochemistry.* **29**: 8658-8676.

Basu S, Bhattacharya D, Banerjee R (2011). **Mapping the distribution of packing topologies within protein interiors shows predominant preference for specific packing motifs.** *BMC Bioinformatics.* **12**: 195.

Banerjee R, Sen M, Bhattacharya D, Saha P. (2003). **The jigsaw puzzle model: search for conformational specificity in protein interiors.** *J. Mol. Biol.* **333**: 211–226.

Bahadur RP, Chakrabarti P. (2009). **Discriminating the native structure from decoys using scoring functions based on the residue packing in globular proteins.** *BMC Structural Biology*. **9**: 76.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983). **CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations.** *J Comp Chem*. **4**: 187-217.

Branden CI, Jones TA (1990). **Between objectivity and subjectivity.** *Nature.* **343**: 687-689.

Edgar RC (2004). **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nuc. Acids. Res.* **32**: 1792-1797.

Holm L, Rosenstrom P (2010). **Dali server: conservation mapping in 3D.** *Nuc. Acids. Res.* **38**: W545-549.

Holm L, Sander CJ (1992). **Evaluation of protein models by atomic solvation preference.** *J. Mol. Biol.* **225**: 93-105.

Heifetz A, Katchalski-katzir E, Eisenstein M (2002). **Electrostatics in protein–protein docking.** *Protein Sci,* **11**: 571–587.

Kahraman A, Morris RJ, Laskowski RA, Thronton JM (2007). **Shape variation in protein pockets and their ligands.** *J. Mol. Biol.* **368**: 283-301.

Kahraman A, Morris RJ, Laskowski RA, Favia AD, Thronton JM (2010). **On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins.** *Proteins.* **78**: 1120-1136.

Krivov GG, Shapovalov MV, Dunbrack RL (2009). **Improved prediction of protein side-chain conformations with SCWRL4.** *Proteins.* **77**: 778-795.

Lo Conte L, Chothia C, Janin J (1999). **The atomic structure of protein-protein recognition sites.** *J Mol Biol.* **285**: 2177-2198.

Lawrence MC, Colman PM (1993). **Shape complementarity at protein/protein interfaces.** *J Mol Biol.* **234**: 946–950.

Li X, Hu C, Liang J (2003). **Simplicial edge representation of protein structures and alpha contact potential with confidence measure.** *Proteins.* **53**: 792-805.

Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF (2001). **Protein docking using continuum electrostatics and geometric fit.** *Protein Eng.* **14**: 105–113.

Mirzaie M, Eslahchi C, Pezeshk H, Sadeghi M (2009). **A distance-dependent atomic knowledge-based potential and force for discrimination of native structures from decoys.** *Proteins.* **77**: 454-463.

Misura KM, Chivian D, Rohl CA, Kim DE, Baker D (2006). **Physically realistic homology models built with ROSETTA can be more accurate than their templates.** *Proc. Natl. Acad. Sci. USA.* **103**: 5361-5366.

Miyazawa S, Jernigan RL (1996). **Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading.** *J. Mol. Biol.* **256**: 623-644.

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

McCoy AJ, Epa VC, Colman PM (1997). **Electrostatic complementarity at protein/protein interfaces.** *J Mol Biol.* **268**: 570-584.

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995). **SCOP: a structural classification of protein database for the investigation of sequences and structures.** *J Mol Biol*. **247**: 536-540.

Moult J, Fidelis K, Kryshtafovych A, Tramontano A (2011). **Critical assessment of methods of protein structure prediction (CASP)—Round IX.** *Proteins* **79**:1–5.

Park B, Levitt M (1996). **Energy functions that discriminate X-ray and near-native folds from well-constructed decoys.** *J Mol Biol*, **258**: 367-392.

Stockwell GR, Thronton JM (2006). **Conformational diversity of ligands bound to proteins**. *J Mol Biol.* **356**: 928-944.

Samudrala R, Moult J (1998). **An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction.** *J. Mol. Biol.* **275**: 895-916.

Skolnick J, Kolinski A, Ortiz A (2000). **Derivation of protein-specific pair potentials based on weak sequence fragment similarity.** *Proteins.* **38**: 3-16.

Simons KT, Kooperberg C, Huang E, Baker D (1997). **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.** *J. Mol. Biol.* **268**: 209-225.

Shen MY, Sali A (2006). **Statistical potential for assessment and prediction of protein structures.** *Protein Sci.* **15**: 2507-2524.

Samudrala R, Levitt M (2000). **Decoys 'R' Us: a database of incorrect conformation to improve protein structure prediction.** *Protein Sci.* **9**: 1399-1401.

Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D (2003). **An improved protein decoy set for testing energy functions for protein structure prediction.** *Proteins*. **53**: 76-87.

Thompson JD, Higgins DG, Gibson TJ (1994). **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice.** *Nuc. Acids. Res.* **22**: 4673-4680.

Word JM, Lovell SC, Richardson JS, Richardson DC (1999). **Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation.** *J. Mol. Biol.* **285**: 1735-1747.

Zhang C, Liu S, Zhou H, Zhou Y (2004). **An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state.** *Protein Sci.* **13**: 400-411.

The Complementarity Plot: a novel tool for protein structure validation

1. Introduction

The last chapter described the utility of the combined use of surface and electrostatic complementarity measures in fold recognition. This was based on scores which were essentially averages over amino acid residues spanning the whole polypeptide chain enabling a global assessment of the packing and electrostatics in the context of the protein's three dimensional structure. In this chapter we examine whether both surface and electrostatic complementarities can be combined to detect local regions of suboptimal packing and electrostatics (arising due to coordinate errors) so that the methodology could serve as an effective tool for validation of either modeled or experimentally determined protein structures.

In the last decade, there has been an explosion in the number of protein crystal structures deposited in the Protein Data Bank (PDB) currently exceeding 75000 **(Berman et al., 2003)**. This exceptional growth in the available structural data could require a further refinement of existing validation tools to efficiently detect (local / global) structural errors and provide a just estimate of the overall reliability of the reported atomic coordinates **(Read et al., 2001**). The retraction of quite a few high profile structures **(Chang et al., 2006; Hanson and Stevens, 2000; Janssen et al., 2007)** indicates the possibility for erroneous atomic models seeping through the peer review process. Homology modeling, threading techniques and *de novo* structure prediction **(Bradley et al., 2005; Rohl et al., 2004)** should profit from effective validation protocols in assessing the confidence level associated with the final model. Thus, discerning validation procedures could find a wide range of applications in computational and experimental structural studies.

Currently, all 'state of the art' protein structure validation methods includes deviations in covalent bond lengths, bond angles and peptide planarity from ideal values which have been estimated from statistical analyses of either small molecules **(Engh and Huber, 1991; Engh and Huber, 2001)** from the CSD or high resolution protein crystal

structures from the PDB **(Jaskolski et al., 2007)**. Generally, deviations less than 3σ from (unimodal) ideal values are considered to be within the normal range **(Lovell et al., 2003)**. The Ramachandran Plot **(Ramachandran et al., 1963)** continues to be one of the most simple and effective indicators of error where the amino acid φ, ψ's in the disfavored regions of the plot could point to undue geometric strain (and thus possible errors) due to steric overlap between atoms constituting the two contiguous peptide planes, including a methyl group attached to the centrally located C^{α} atom. Subsequently, improvements have been made with regard to the delineation of allowed / disallowed regions based on the distribution of amino acid residues in protein structures and also in the scores used to estimate the quality of the plot **(Laskowski et al., 1993; Kleywegt and Jones, 1996; Davis et al., 2007)**. Combinations of side-chain torsion angles (χ) from a correctly determined structure are also expected to be in agreement with statistical distributions tabulated in rotamer libraries **(Dunbrack and Karplus., 1993)**. In addition, other validation measures have been proposed based on the dense packing of side-chain atoms within proteins **(Pontius et al., 1996)**, avoidance of non-local steric clashes **(Davis et al., 2007)** and the segregation of hydrophobic / hydrophilic residues **(Kleywegt, 2000)**, the former clustering to form cores whereas the latter either tending to interact exclusively with each other or projecting into the bulk solvent. Of these, one of the most successful is the 'Clash score' **(Davis et al., 2007)** involving the contacts of hydrogen atoms, which were generally not included in the refinement of protein crystal structures. Other scores attempt to identify packing defects within proteins **(Willard et al., 2003)**, satisfaction of hydrogen bonding potential **(McDonald and Thornton, 1995; Hooft et al., 1996)** and distortions in their geometry or the 'disharmony' between buried amino acid residues and their immediate atomic environments **(Vriend and Sander., 1993)**.

Some of these structural features involved in validation are encapsulated in the concept of complementarity which has been used in docking algorithms **(Mandell et al., 2001; Heifetz et al., 2002)**, prediction of side-chain conformations **(Liang and Grishin, 2002; Krivov et al., 2009)** and protein quaternary structures **(Caravella, 2002)**. Elevated values for surface (S_m) and electrostatic complementarity (E_m) measures found for

residues within native protein interiors arise naturally due to the stereo-specific interlocking of side-chains (avoiding short contacts and packing defects) **(Banerjee et al., 2003)** and the exquisite balancing of charges (inclusive of hydrogen bonds) **(Basu et al., 2012)** to stabilize the protein fold. Thus, the strain experienced by buried residues, consistent with the short and long range forces sustaining the native fold can be estimated by the combined use of the two complementarity measures (***S_m, E_m***). This chapter demonstrates the construction and application of a novel graphical tool namely the 'Complementarity Plot' (CP) which conveniently identifies residues with suboptimal packing and electrostatics and can also be used to judge the overall quality of a protein crystal structure in terms of packing and electrostatics.

The chapter further describes the design of a set of scores (tested on several databases) which describe the quality of the plot in several ensuing applications. Based on these scores, the ability of the plot to detect errors in side-chain rotamers, geometrical parameters and disqualify obsoleted, retracted structures has been tested. Possible applications of the plot in homology modeling and protein design have been surveyed. An attempt has also been made to probe (using the methodology of CP) the relationship of deviations in geometrical parameters to fold integrity.

2. Materials and Methods

2.1. Databases

The database **DB2** described in chapter 2 (containing 400 structures, R-factor ≤ 20 %, resolution ≤ 2 Á and homologues removed at greater than equal to 30% sequence identity, polypeptide chain-length: 75 to 500 residues) was used as a training set in the design of the complementarity and accessibility scores (***CS_l, rGb***) which were then independently tested on three datasets **UDB, MDB, LDB** spanning resolution ranges ≤ 1 Á, 2-2.5 Á, > 3 Á respectively **(**see **Supplementary Information** in CD enclosed**)**. For structure validation in the case of real data, 110 pairs of obsolete structures and their upgraded partners were collected **(OUDB :** see **Supplementary Information in the CD**

enclosed) from the PDB (ftp://ftp.wwpdb.org/pub/pdb/data/status/obsolete.dat). In order to ensure that an upgraded structure was genuinely better over its obsolete counterpart only those pairs were selected wherein the improvement in resolution and R-factor were better than 0.2 Á and 0.02 respectively. For calculations involving synthetic data a composite database consisting of 143 high-resolution structures was assembled (**SDB**) and subsets there from were used for detection of diffused errors (**SDB-1**), idealization (see below) (**SDB-2, SDB-3**) and protein design (**SDB-3, SDB-4**) (**Table 1**).

Table 1. The Datasets used in the calculations. Except for the pairs of obsolete and upgraded structures in OUDB, no protein with R-factor > 20 % were included in any of the databases. For oligomeric proteins, only the largest polypeptide chain was retained for calculations. In case of multiple occupancies, atoms with the highest occupancy were selected and the first conformer for equal occupancies. For all the databases, homologues were removed at sequence identity of 30% or more. Criteria for successful validation in Procheck: greater than -1.0 for all G-factor scores and 'INSIDE' or 'BETTER' recorded for bad contacts. Criteria for successful validation in Molprobity: Ramachandran favored: > 98%, Ramachandran outliers: < 0.05%, Poor Rotamers: < 1%, Bad backbone bonds: 0%, Bad backbone angles: < 0.1%, Clash-score ≤ 20.

Database	Resolution range	Chain length (aa)	Number of proteins	Additional Criteria	Usage
DB2	≤ 2 Å	75-500	400	No proteins with deeply embedded prosthetic groups, No missing atoms	Training, Parameterization of $CS_{l,,}$ *rGb*
UDB	≤ 1 Å	38 - 670	113	-	Computation of $CS_{l,,}$ *rGb*
MDB	> 2 Å , ≤ 2.5 Å	59 - 185	92	-	Same as UDB
LDB	≥ 3 Å	45 - 500	164	-	Same as UDB
OUDB	1.1-3.4Å	65-900	110 pairs of obsolete and corresponding upgraded structures	Difference in resolution, R-factor between obsolete and upgraded pair: 0.2 Å, 0.02 respectively	Pair-wise Comparison, Detection of errors in Rotamer, Regularization
SDB-1	≤ 2 Å	56-363	20	divided equally among the four major protein classes	Idealization
SDB-2	≤ 2 Å	56-387	30	satisfying all validation filters implemented in Procheck	Detection of low-intensity diffused synthetic errors in main-chain parameters
SDB-3	≤ 1 Å	38 – 670	68	No missing atoms	Idealization, Detection of synthetic errors in rotamer, Design
SDB-4	≤ 2 Å	57-363	25	satisfying all validation filters implemented in Molprobity	Design, Detection of single point mutations (Val ↔ Thr)

2.2. The Complementarity Plot

The individual (S_m^{sc}, E_m^{sc}) values of completely / partially buried residues were plotted in a Complementarity Plot (CP) spanning -1.0 to 1.0 in both the X (S_m^{sc}) and Y (E_m^{sc}) axes. Given the fact that for residues in correctly folded proteins both S_m^{sc}, E_m^{sc} are largely constrained to a limited range of values (as a function of their burial, see chapter 4), regions in CP encompassing points corresponding to such amino acids could be clearly delineated. From the database **DB2**, S_m^{sc}, E_m^{sc} values of all (target) residues irrespective of the amino acid type were plotted separately based on their burial bins accounting for 23850, 10624, 13255 residues in bins 1, 2 and 3 respectively. Thus in all, three plots (CP1, CP2, CP3) were obtained. To start with, all the buried residues from the database **DB2** were plotted in the CPs, which had been divided into square-grids (of width 0.05 × 0.05), and the center of every square grid was assigned an initial probability (P_{grid}) equal to the number of points in the grid divided by the total number of points in the plot. The probability of a residue to occupy a specific position in the plot was then estimated by bilinear interpolation from the probability values of its four nearest neighboring voxels. The plots were then contoured based on their probability values $P_{grid} \geq 0.005$ for the first contour level and ≥ 0.002 for the second **(Fig.1)**.

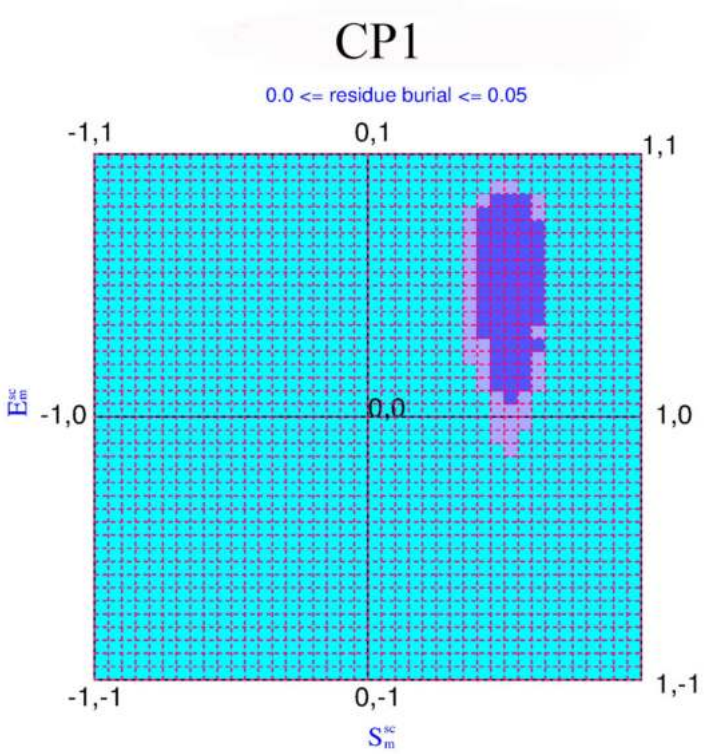


Fig.1. CP1: The Complementarity Plot for the 1st burial bin. 'Probable', 'Less probable' and 'Improbable' regions of the plot are colored in purple, mauve and sky-blue respectively.

The cumulative probability of locating a point within the second (outer) contour for the three plots were 91%, 90%, 88% respectively whereas for the first (inner) contour, the probability gradually dropped with increasing solvent exposure (82%, 76%, 71%). Inspired by the Ramachandran Plot **(Ramachandran et al., 1963)**, the region within the first contour was termed 'Probable', between the first and second contours 'Less Probable' and outside the second contour 'Improbable' **(Fig.2)** individually for all three plots (CP1, CP2, CP3). In such a plot residues with low S_m^{sc} and E_m^{sc} (< 0.2 for both) are easily identified.

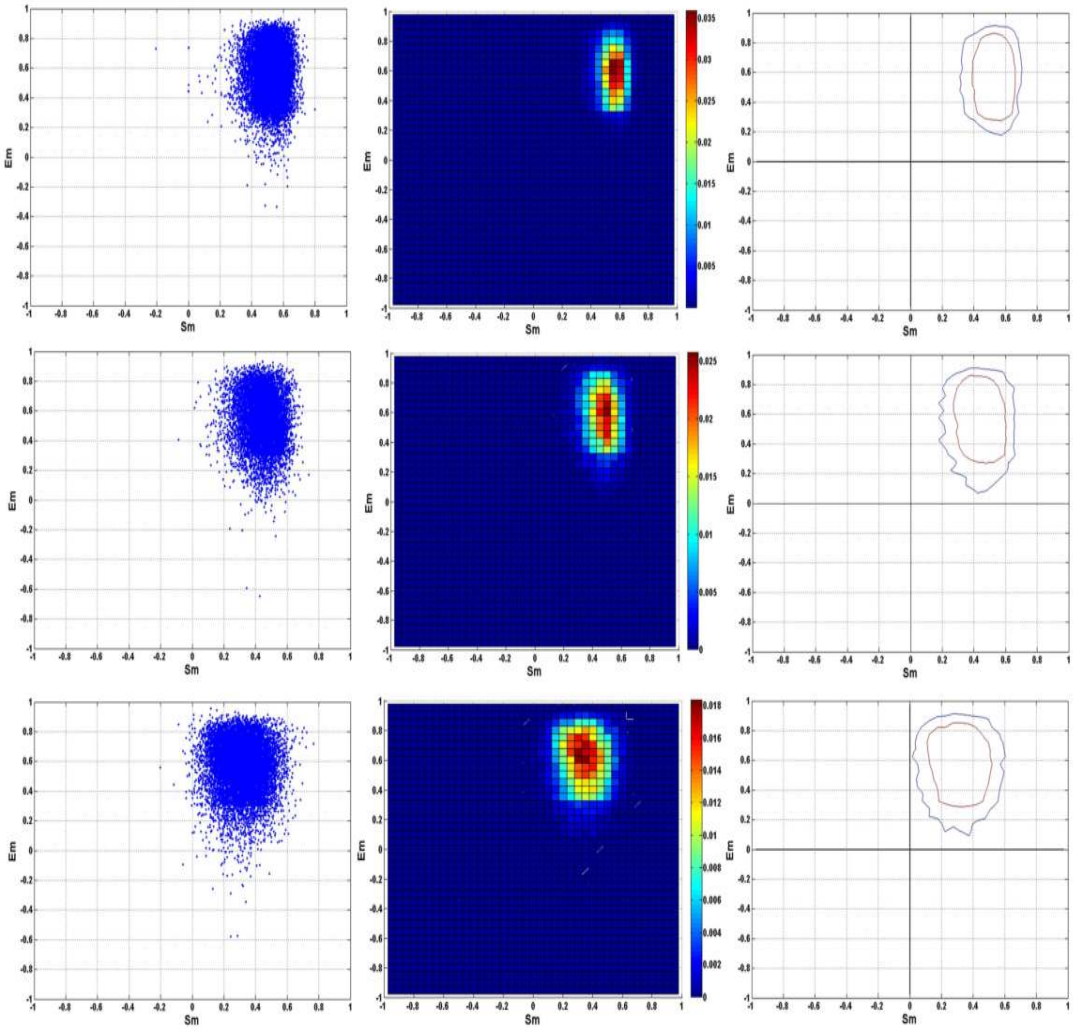


Fig 2. Construction of the Complementarity Plot. The left panel plots the distribution of (S_m^{sc}, E_m^{sc}) values for all buried or partially buried residues from **DB2** according to their burial (From top to bottom: CP1, CP2, CP3). The middle panel shows initial grid probabilities assigned to each two-dimensional grid of width 0.05 × 0.05. The right panel shows the (inner and outer) contour demarcating the 'probable', 'less probable' and 'improbable' regions.

2.3. Complementarity and Accessibility Scores

In order to quantify the quality of the plots, a Complementarity Score was designed wherein all points in each plot were first partitioned into two sets, those with zero and non-zero probabilities. Occurrence of any point with zero probability (essentially in the improbable region) implies that the corresponding residue exhibits suboptimal packing and/or electrostatics with respect to the rest of the protein and therefore should be penalized. The score thus consists of two terms, the first essentially the average of the non-zero log probabilities and the second, the fraction of residues with zero-probability multiplied by a penalty (*Pen*). Thus the score would be expected to decrease with increase in the points in the improbable regions of the plot. For a particular plot (say CP1) the score can be defined as:

$$Sl = \left[\frac{1}{N}\sum_{i=1}^{N}\log_{10}(P_i)\right] - Pen\cdot\left(\frac{N_{zero}}{N_{tot}}\right) \qquad (1)$$
$$= Sl_{non\text{-}zero} + Sl_{zero}$$

where N_{tot} is the total number of points in the plot which can be partitioned into points which fall in square grids of non-zero probability (N) with grid probabilities P_i and those located in grids of zero probability (N_{zero}). For the first term it was assumed that the probability assigned to one point (P_i) is independent of the others, leading to a multiplication of probabilities (P_1, P_2, ...) and converted into a summation by taking log ($\sum_{i=1}^{N}\log_{10}(P_i)$). There is some measure of arbitrariness in assigning the value for *Pen* which was computationally optimized. Even for accurately determined structures from **DB2**, generally 10% of the residues (per chain) would be located in the improbable regions of the plots. It was thus decided that for correctly folded proteins (of the kind found in **DB2**), the ratio of the two terms ($R_{Sl} = Sl_{zero} / Sl_{non\text{-}zero}$) should optimally be in the range 0.30, greater than which, it would unjustifiably begin to dominate the overall

score whereas too low a value (say less than 0.10) would compromise the sensitivity of the score to structural errors. Several values of *Pen* were tested on **DB2** where the two terms (Sl_{zero} & $Sl_{non\text{-}zero}$) were estimated for each polypeptide chain in the database; initially applying the same *Pen* for all the three plots (CP1, CP2, CP3; **Table 2**). For uniform penalties applied to all the three plots it was observed that R_{Sl} tended to increase from CP1 to CP3 as relaxation in packing constraints (with corresponding increase in solvent exposure) increased the relative fraction of points in the zero probability grids from CP1 to CP3 (N_{zero}/N_{tot} for CP1: 0.026 (± 0.029), CP2: 0.037 (± 0.048), CP3: 0.045 (± 0.043)). Thus, to introduce some measure of uniformity, *Pen* was modulated (CP1: 25; CP2: 20; CP3: 15) such that R_{Sl} was in the range 0.30 – 0.35 for all the three plots. Understandably, the ratios of the penalties (*Pen*) in the three plots (CP1/CP2: 25/20 = 1.25; CP1/CP3: 25/15 = 1.67) were correlated to the corresponding ratios of N_{zero}/N_{tot} (CP2/CP1: 0.037/0.026 = 1.42; CP3/CP1: 0.045/0.026 = 1.73).

Table 2. Sensitivity of CS_l to different values of penalty (*Pen*). The quantum of penalty *(Pen)* applied to CP1, CP2, CP3 is indicated in the first column of the table and R_{Sl} = Sl_{zero} / $Sl_{non\text{-}zero}$.

Pen			**R_{Sl}**			***CS_l***
CP1	**CP2**	**CP3**	**CP1**	**CP2**	**CP3**	
100	100	100	1.31 (±1.44)	1.75 (±2.22)	2.02 (±1.93)	-0.54 (± 2.33)
75	75	75	0.98 (± 1.08)	1.31 (± 1.66)	1.52 (± 1.45)	0.33 (± 1.75)
50	50	50	0.66 (± 0.72)	0.88 (± 1.11)	1.01 (± 0.96)	1.19 (± 1.17)
30	30	30	0.39 (± 0.43)	0.53 (± 0.66)	0.61 (± 0.58)	1.89 (± 0.71)
25	25	25	0.33 (± 0.36)	0.44 (± 0.55)	0.51 (± 0.48)	2.06 (± 0.59)
20	20	20	0.26 (± 0.29)	0.35 (± 0.44)	0.41 (± 0.39)	2.23 (± 0.48)
15	15	15	0.20 (± 0.22)	0.26 (± 0.33)	0.31 (± 0.29)	2.40 (± 0.36)
10	10	10	0.13 (± 0.14)	0.18 (± 0.22)	0.20 (± 0.19)	2.58 (± 0.25)
5	5	5	0.07 (± 0.07)	0.09 (± 0.11)	0.10 (± 0.10)	2.75 (± 0.14)
30	25	20	0.39 (± 0.43)	0.44 (± 0.55)	0.41 (± 0.39)	2.06 (± 0.60)
25	20	15	0.33 (± 0.36)	0.35 (± 0.44)	0.31 (± 0.29)	2.24 (± 0.48)
20	15	10	0.26 (± 0.29)	0.26 (± 0.33)	0.20 (± 0.19)	2.41 (± 0.37)

Finally,

$$CS_l = K + \sum_{j=1}^{3} wb_j \cdot Sl_j \qquad (2)$$

As has been mentioned, scores for deviant structures are expected to decrease in value. So for convenience of interpretation, K was empirically set to 5.0 so as to obtain an overall positive score from 0 to 5 in case of a favorable distribution spanning the three plots. It follows that such a constant merely acts as a scale factor universally applied to all CS_l scores. wb_j is the number of points in the j^{th} plot divided by the total number of points in the three plots and the (weighted) summation is over CP1, CP2 and CP3.

The sensitivity of ***CS**l* was also tested **(Table 2)** for different combinations of penalties by computing its mean and standard deviations for all chains in **DB2**. Standard deviations were especially high (1.17 to 2.33) for uniform penalties 100, 75, 50 whereas for different combinations of penalties in the range of 5 to 30, ***CS**l* was found to be fairly stable with standard deviations falling in range of 0.14 to 0.60 **(Table 2)**, and ***CS**l* was confirmed to be well behaved for the selected penalty values (*Pen* = 25, 20, 15 for CP1, CP2, CP3 respectively).

In order to check the expected distribution of amino acid residues w.r.t. burial, the following score was defined.

$$rGb = \frac{1}{N_{res}} \sum_{i=1}^{N_{res}} \log_{10}(\mathrm{Pr}_i) \qquad (3)$$

where N_{res} is the total number of residues in the polypeptide chain and Pr_i is the propensity of a particular amino acid (Val, Leu etc) to acquire a particular degree of solvent exposure (corresponding to buried residues in the three burial bins and a 4^{th} bin composed of exposed residues (***Bur*** > 0.30)).

$$\mathrm{Pr}_j = \frac{P(\mathrm{Re}s(j) \mid Bur(j))}{\left(\frac{N(\mathrm{Re}s(j))}{N_{DB}}\right)} \tag{4}$$

where *P(Res(j)|Bur(j))* is the conditional probability of *Res(j)* (say Val) to acquire a given burial, *Bur(j)*. *N(Res(j))* is the number of residues of identity *Res(j)* found in the database and N_{DB} is the total number of residues in the database (**DB2**). Glycines were disregarded in all the scores due to the lack of any non-hydrogen side-chain atoms.

To quantify the individual contributions of S_m^{sc} and E_m^{sc}, two additional (global) scores P_{Sm} and P_{Em} were further defined. The normalized frequency distribution separately for each burial bin was used to assign discrete probabilities (P (x < S_m^{sc} < (x+0.05))) to S_m^{sc} divided into intervals of 0.05. Three such probability distributions were computed one for each burial bin and a similar procedure was adopted for E_m^{sc}. Then, for each polypeptide chain, the individual probabilities were averaged over all buried or partially buried residues, giving rise to the two following measures:

$$P_{Sm} = \frac{\sum_{i=1}^{N_b} \log_{10}(P_i(S_m^{sc}))}{N_b}; \qquad P_i(S_m^{sc}) \neq 0 \tag{5}$$

$$P_{Em} = \frac{\sum_{i=1}^{N_b} \log_{10}(P_i(E_m^{sc}))}{N_b}; \qquad P_i(E_m^{sc}) \neq 0 \tag{6}$$

where N_b is the total number of buried or partially buried residues in a given polypeptide chain.

In addition, a local score (P_{count}) was also defined simply as the number of points in the improbable regions divided by the total number of points spanning the three plots.

2.4. Building Idealized structures

Idealization refers to the reversal of all main chain bond lengths, angles along with torsion angle ω to their corresponding ideal values. A locally developed algorithm was utilized to build idealized structures from the native coordinates which was cross checked using the 'Build and Edit protein' module in the Accelrys **(Studio, D., 2.5 Guide, Accelrys Inc., San Diego. 2009)** suite of programs. Both methods gave nearly identical results, as an RMSD of 0.035 Á (for 2HAQ) was obtained upon superposing **(Holm and Rosenstrom, 2010)** the two structures which had been built by an identical set of (idealized) geometrical parameters. For the in-house program, a single peptide plane consisting of atoms C^{α}_{i-1}, C_{i-1}, O_{i-1}, N_i, H_i and C^{α}_{i} was initially constructed based on ideal values for bond lengths, bond angles **(Engh and Huber, 2001)** and ω **(Vriend, 1990)**. Atomic coordinates of C_i, N_{i+1} and C^{α}_{i+1} of the successive peptide plane were then determined by the repeated application of the 'fourth atom fixing' procedure **(Ramachandran and Sasisekharan, 1968)**, in the course of which, the native values of φ, ψ were retained. The positions of the remaining atoms (O_i, H_{i+1}: second plane) were then generated by superposing the initially obtained idealized peptide plane onto the predetermined atoms C_i, N_{i+1} and C^{α}_{i+1}. Finally, the side chain atoms (extracted from the native coordinates) were threaded onto the idealized main-chain by superposing N, C^{α}, C coordinates of every residue onto their main-chain counterparts. When native values for all geometrical parameters were fed into the program a C^{α}-RMSD of 0.035 Á (side-chain RMSD: 0.5 Á, 2HAQ) was obtained between the rebuilt structure and the native coordinates upon superposition **(Holm and Rosenstrom, 2010),** which also confirmed the correctness of the idealization protocol. Idealized structures using conformation dependent ideal values (for bond angles) from a library (CDL: http://dunbrack.fccc.edu/nmhrcm/) were built by suitably adapting the algorithm given

above, where the ideal values were now dependent on residue identities and the relative orientation of contiguous peptide planes (φ, ψ) **(Berkholz et al., 2009)**. Hydrogen atoms were then removed and geometrically rebuilt by REDUCE **(Word et al., 1999)**. The idealized structures were then energy minimized by CHARMM **(Brooks et al., 1983)** with either hard (constant harmonic force parameter set to 250.0 for N, C^{α}, C, O atoms and 10.0 for C^{β}) or soft (5.0, 2.5: flexible backbone) harmonic restraints on main-chain atoms and C^{β}.

For the obsolete and upgraded pairs of crystal structures **(OUDB)** another procedure was adopted to 'regularize' the coordinates. The 'REFI' routine in 'O' **(Jones et al., 1991)** was used to regularize the geometry of the coordinates to convergence (RMSD shift 0.000 Á) and then used for subsequent calculations.

2.5. Incorporation of low-intensity diffused errors into native coordinates

A predetermined quantum of small random errors in pre-selected geometrical parameters (±0.5σ **; Engh and Huber, 2001)** approximately ranging from 1.5 - 2.5° for main-chain bond angles and ±1° for (φ, ψ)) was incorporated into native crystal structures, by perturbing the specified parameter on randomly chosen residues. The protein structure was then rebuilt using computational procedures described above.

2.6. Single residue swapping

In case of single residue swapping, completely buried ($0.00 \leq \boldsymbol{Bur} \leq 0.05$) valines and threonines initially located in the probable region of CP1 were identified (**SDB-4)** and their identities interchanged. Thus, each altered file contained a single transition (Val↔Thr) w.r.t. the native. While swapping, the native χ_1 torsion of the original residue was retained while other side-chain parameters (Val: C^{β}-$C^{\gamma 1}$ bond length, $C^{\gamma 1}$-C^{β}-$C^{\gamma 2}$ angle; Thr: C^{β}-$O^{\gamma 1}$, $O^{\gamma 1}$-C^{β}-$C^{\gamma 2}$: average values obtained from **DB2**) were altered according to the identity of the mutated residue.

2.7. Building Homology models

To test the performance of CP on homology models, 20 structures representing a fairly wide cross section of folds were selected as templates from the SCOP database **(Murzin et al., 1995)**. For each template structure 5 other sequences with varying identities (ranging from 13% to 90%) were chosen by a BLAST **(Johnson et al., 2008)** search (using the DELTA-BLAST algorithm) against the PDB. Sequence similarities and identities were calculated using the 'Align sequence profiles' module (scoring matrix: BLOSUM62) as implemented in the Accelrys **(Studio, D., 2.5 Guide, Accelrys Inc., San Diego. 2009)** suite of programs. The resultant alignment profile along with the template-backbone coordinates were fed to the 'Build homology model' module of Accelrys with 'High' optimization. The top most model with lowest total energy and physical energy were then selected. All models were finally energy minimized with flexible backbone and subjected to validation by the Complementarity Plot.

3. Results and Discussion:

As will be evident from the definition of the complementarity functions (see **Materials and Methods**), perfect fit between two surfaces (for example identical surfaces) will return a value of 1.00 for S_m^{sc}. Likewise E_m^{sc} will be 1.00 for perfect anti-correlation between two sets of electrostatic potential values on a given surface. Generally for completely buried ($0.00 \leq \boldsymbol{Bur} \leq 0.05$) residues in correctly folded proteins, both S_m^{sc} and E_m^{sc} lie in the narrow range of ~ 0.50 – 0.55 and ~ 0.50 – 0.70 respectively, regardless of their identity, thereby satisfying fairly stringent constraints in both packing and electrostatics **(Basu et al., 2012).** For higher solvent exposure ($0.05 < \boldsymbol{Bur} \leq 0.30$) there is some measure of relaxation in the constraints. The CP consists in plotting the surface (S_m^{sc}: X-axis) and electrostatic (E_m^{sc}: Y-axis) complementarity values of individual residues. The term 'Complementarity Plot' (CP) is perhaps a misnomer as there are actually three plots, each serving a given range of solvent exposure of the plotted residues (CP1, CP2, CP3 for burial bins 1, 2, 3: see **Materials and**

Methods). The constraints both in terms of packing and electrostatics are reflected in the dense population of points in a localized region of the CPs **(Basu et al., 2012)**. Thus points straying into the improbable regions of the plot denote either defective packing of side chain atoms and/or imbalance in the distribution of partial charges within the protein interior likely to be symptomatic of fold instability.

To quantify the character of a distribution of points spanning all the three plots, proportional to their net probability of occurrence, a complementarity score has been designed (***CS$_l$***). In addition, a second score (accessibility score: ***rGb***) essentially estimates the propensity of a particular amino acid residue (Leu, Val etc) to acquire a specified degree of solvent accessibility (see **Materials and Methods**).

For other specific applications and also to benchmark CP relative to other (local) measures, a local score (P_{count}) was also defined which simply consists in counting the number of points in the improbable regions divided by the total number of points in the three plots. Generally, validation criteria are distinguished in terms of whether they apply to the entire three dimensional structure of proteins (global) such as R-factor, Procheck G-factor scores or apply to individual residues (local), for example steric clashes between atoms (clash score). As both scores (***CS$_l$, rGb***) are computed on entire polypeptide chains (or a collection of points in the plots) they could be treated as 'global'. Thus both local (P_{count}) and global scores (***CS$_l$, rGb***) have been incorporated into CP, the principal difference between them being, the former is a count of individual points imbued with some particular attribute whereas the latter is an average of some sort over a collection of points. The standalone suite (see **Program availability**) also lists those residues which are in the improbable region of the plots so that the possible errors can actually be pinpointed by looking at the original structure. Thus, the CP-scores can be applied to each polypeptide chain (in turn) in a database or simply to a distribution of points in the plot without any reference to individual proteins. Here, the primary emphasis is on the performance of the global CP scores (***CS$_l$, rGb***) and P_{count} has been defined merely to compare the CP methodology with other existing local validation scores. However, much

should not be made of the distinction between local and global scores as in the final analysis the primary interest is whether a structure passed a particular validation test, or not, regardless of the specific nature of the score.

3.1. Testing the scores in different resolution ranges

The global scores (***CS$_l$, rGb***) were initially optimized on the training database, **DB2** (see **Materials and Methods**) to yield values of 2.24 (σ: ±0.48), and 0.055 (±0.022) respectively. They were then tested on 3 independent datasets consisting of ultrahigh (**UDB**), medium (**MDB**) and low resolution structures (**LDB**). Both the scores from **UDB** and **MDB** were in good agreement with values observed for the training set (**DB2**), in contrast to **LDB** which exhibited significant decrease (**Fig. 3**). The discriminating power of ***CS$_l$***, ***rGb*** consistent with the visually recognizable features in the distribution of points in the CPs was thus fairly well established.

As has been previously mentioned, deviations of less than 3σ in geometrical parameters (bond lengths, bond angles etc.) from ideal values are considered to be within the normal range. Thus, for all the scores a cut-off of 3σ from the mean was decided as the threshold (with the sole exception of ***rGb***) for successful validation. Thus, the threshold value for ***CS$_l$*** was set to 0.80 (μ-3σ from **DB2**). Similarly, the average values for P_{Sm} and P_{Em} for all chains in **DB2** were -0.855 ±0.054 and -1.492 ±0.099 respectively and their threshold values were set to -1.017 and -1.789 (μ-3σ). Again, considering all the polypeptide chains in **DB2**, an average of 8.75% (4.10), 9.25% (5.05) and 11.14% (6.00) of the points (per chain) were found to be in the improbable regions of CP1, CP2, CP3 respectively. Thus, any polypeptide chain was considered to have successfully passed the validation test for the 'local score', P_{count} when less than 15% (3σ; average σ from the three plots: 5.05) of its residues / points were located in the improbable regions taking into consideration all three plots. Only in case of ***rGb*** was the cutoff reduced to μ-2σ : 0.011, as the standard deviation was fairly high (σ = 0.4μ) and 3σ actually exceeds the mean. It was also confirmed by visual inspection that for structures with ***rGb*** approximately ~0.000, the three dimensional distribution of residues w.r.t. solvent

accessibility was non-native. Throughout this work, the two global scores CS_I and *rGb* have been used in conjunction, that is successful validation required the simultaneous satisfaction of their individual criteria.

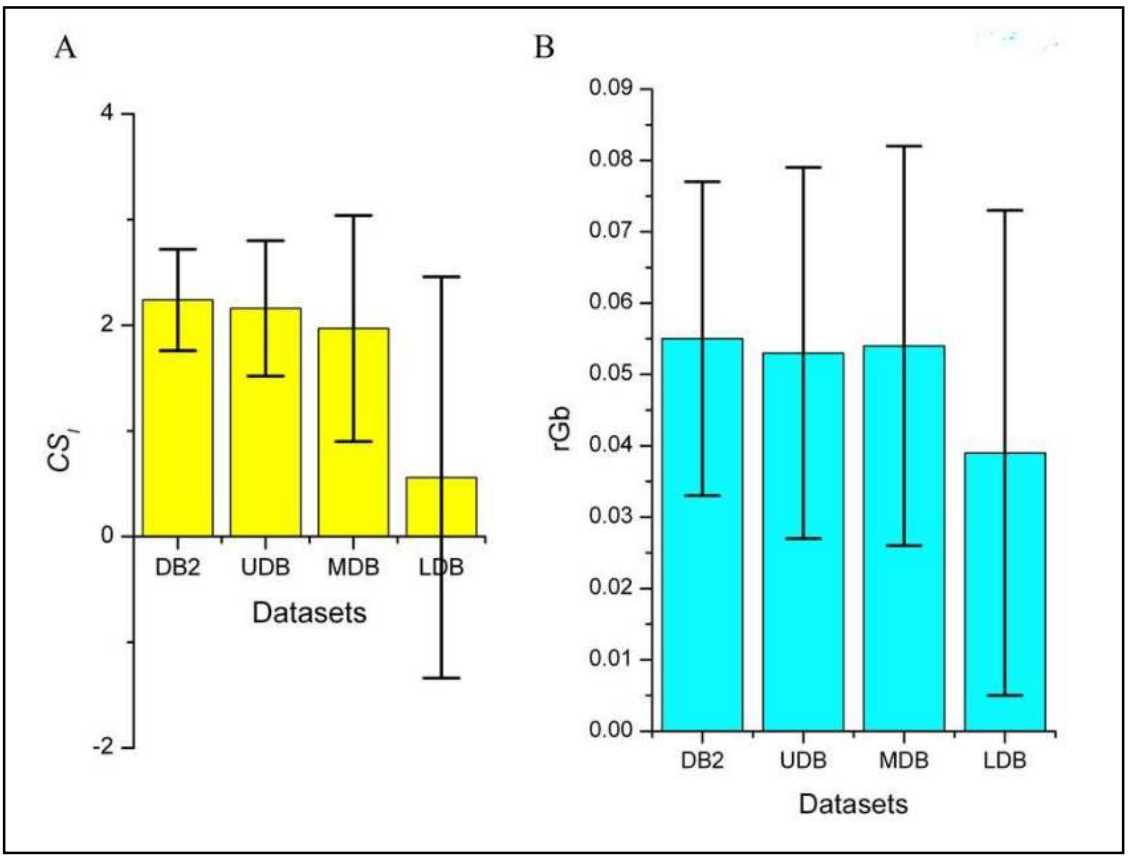


Fig.3. Training and testing of the Complementarity and Accessibility Scores (*$CS_{I,}$ $CS_{f,}$ rGb*) in DB2 and datasets of different resolution ranges (UDB, MDB, LDB). The average (colored filled bars) and standard deviations (error bars) for the three scores (A) *$CS_{I,}$* (B) ***rGb***.

Similarly, the average values for P_{Sm} and P_{Em} for all chains in **DB2** were 0.169 ± 0.012 and 0.061 ± 0.008 respectively and their threshold values were set to 0.133 and 0.037 (μ-3σ). Again, 9, 10 and 12% of the points were found in the improbable region of the plots, for the three burial bins respectively (from chains in **DB2**). Any polypeptide

chain was considered to have successfully passed the validation test for the 'local score', P_{count} when less than 12.5% of its corresponding points were located in the improbable regions of the plots.

3.2. Discriminating obsolete structures from their upgraded counterparts

In order to test the performance of CP for real data, a database **(OUDB)** consisting of 110 pairs of obsolete and upgraded structures were compiled. For each pair, the upgraded structure was better refined relative to its obsolete counterpart indicated by improvements in their corresponding resolutions and R-factors (see **Materials and Methods**). Firstly, the complementarity scores were computed for all the chains in the database and compared pairwise **(Fig. 4)**. On applying the validation criteria (for CP) mentioned above, 69, 97 structures passed the test for the obsolete and upgraded sets respectively. Based on the 'local' score (P_{count}), the corresponding numbers were 44, 72. For benchmarking, the packing and hydrogen bonding parameters were calculated by Whatcheck **(Hooft et al., 1996)** for each chain in the two sets and the number of residues with '*abnormal new packing environment*' and '*unfulfilled buried hydrogen bond donor or acceptor*' were summed. In case a residue appeared in both lists it was considered only once. Finally, the number of such anomalous residues divided by the chain length was used as a criterion for validation **(Fig. 4)**. Since no criteria for rejection is given in the Whatcheck manual, a variety of cutoffs were tried. A cutoff of 5% led to 25 and 53 successful validations in both sets and similar numbers obtained for cut offs of 10 and 15% were 89, 102 and 106, 110 respectively.

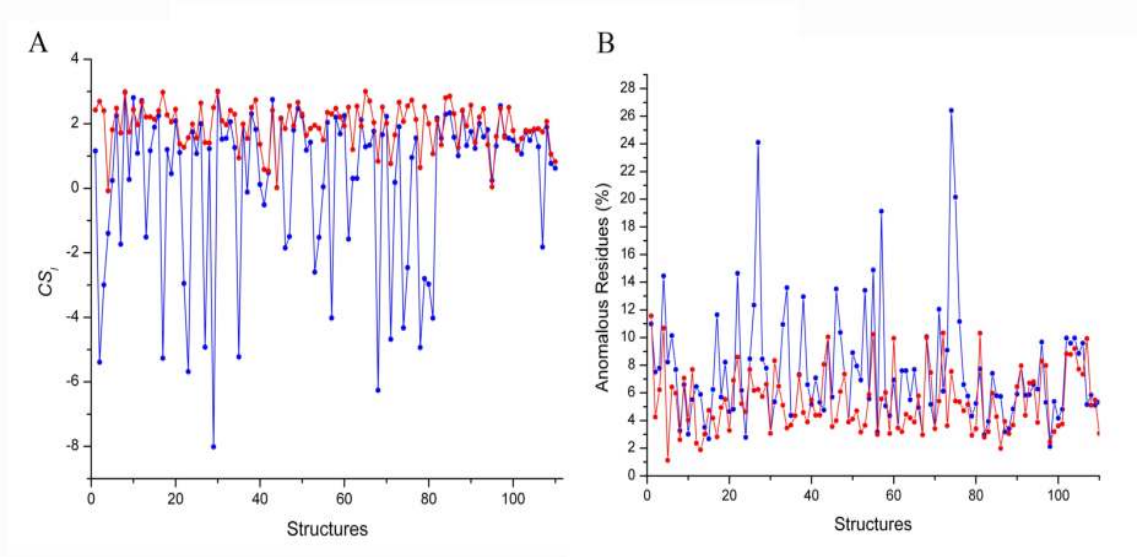


Fig.4. Comparison of CP and What-check packing parameters in case of obsolete and upgraded structures. (A) ***CS$_I$*** and (B) **Fraction of anomalous residues (%)** are plotted in red and blue for the obsolete and upgraded structures respectively.

3.3. Detection of errors in rotamers

As would be evident from the description of the CPs, the scores primarily concern the subjection of side-chain atoms to short and long range forces in the protein. Thus it would be expected that wrong assignments in side-chain rotamers due to low resolution data or some other reason, should evoke a sensitive response from these measures. To test this hypothesis those side-chains from the set of obsolete structures were compiled (1061 residues in all) which differed by more than 40° from their corresponding residues in their upgraded counterpart (involving χ_1 and χ_2) and yet were within 40° of another valid rotamer combination **(Berkholz et al., 2009)**. These two sets of residues (Obsolete, Upgraded) were plotted in the CPs and the partitioning of points in the probable, less probable and improbable regions **(Fig. 5)** compared against the standard distribution in **DB2** (CP1: 82.1%, 9.2%, 8.7%; CP2: 76.1%, 13.9%, 10.0%; CP3: 70.7%, 16.8%, 12.5%). For completely buried residues (CP1) in the obsolete set, the proportion of residues in the three regions (39.7%, 21.7%, 38.6%) significantly differed from that found in **DB2**, in contrast to the upgraded set which was found to be in fairly good

agreement (73.7%, 15.6%, 10.8%). Significant differences in the two distributions were also found for CP2 (obsolete: 42.8%, 20.5%, 36.7%; upgraded: 64.9%, 21.2%, 13.9%) and CP3 (obsolete: 47.7%, 29.3%, 22.9%, upgraded: 60.9%, 25.7%, 13.4%). Deviations from the expected distributions (**DB2**) were estimated by means of χ^2 (df=3-1, probable, less probable, improbable; $\chi^2_{0.05} = 5.991$) for each of the two sets separately for all the three CPs. For obsolete and upgraded structures, χ^2 were found to be 509.8, 21.55 (CP1), 191.8, 14.53 (CP2) and 67.82, 15.93 (CP3) respectively. As the points have been plotted without any reference to the rest of the polypeptide chains the χ^2 could be considered an adaptation of the 'local' score. The relative decrease in χ^2 for obsolete structures from CP1 to CP3 is obviously due to the relaxation in packing with increase in solvent exposure. The two sets could also be clearly discriminated by the global CP-scores applied to the entire distribution: $\boldsymbol{CS_I}$: -1.73, ***rGb*** : 0.027 (obsolete); $\boldsymbol{CS_I}$: 1.97, ***rGb*** : 0.031 (upgraded).

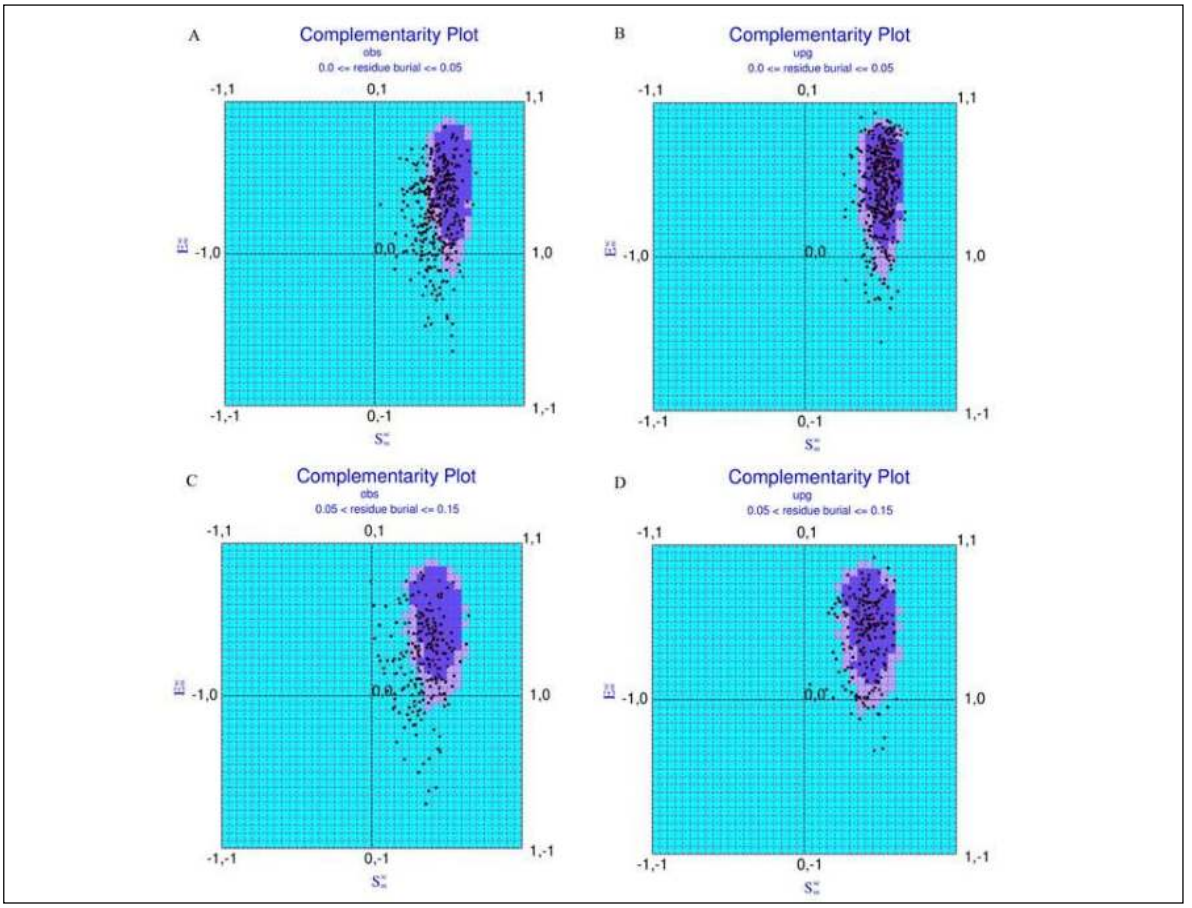


Fig. 5. Distribution of residues from obsolete structures that have a different (yet valid) side-chain rotamer than their upgraded counterparts. (A). CP1 for obsolete, (B). CP1 for upgraded, (C). CP2 for obsolete and (D). CP2 for upgraded structures.

In another test, a subset consisting of 222 deeply buried residues ($0.0 \leq$ ***Bur*** $\leq$ 0.05) from upgraded structures were identified which were originally found to be located in the probable region of CP1. They were then replaced by their corresponding counterparts from the obsolete structures. Subsequent to the replacement, 45% of the points were relocated in the improbable region of the plot, 16% were found in the less probable region whereas 39% were retained in the probable region **(Fig. 6).** The overall χ^2 for the distribution of points was 397.63. Thus, CP could have applications when

dealing with low-resolution data where automated side-chain rebuilding methods generally do not work very efficiently.

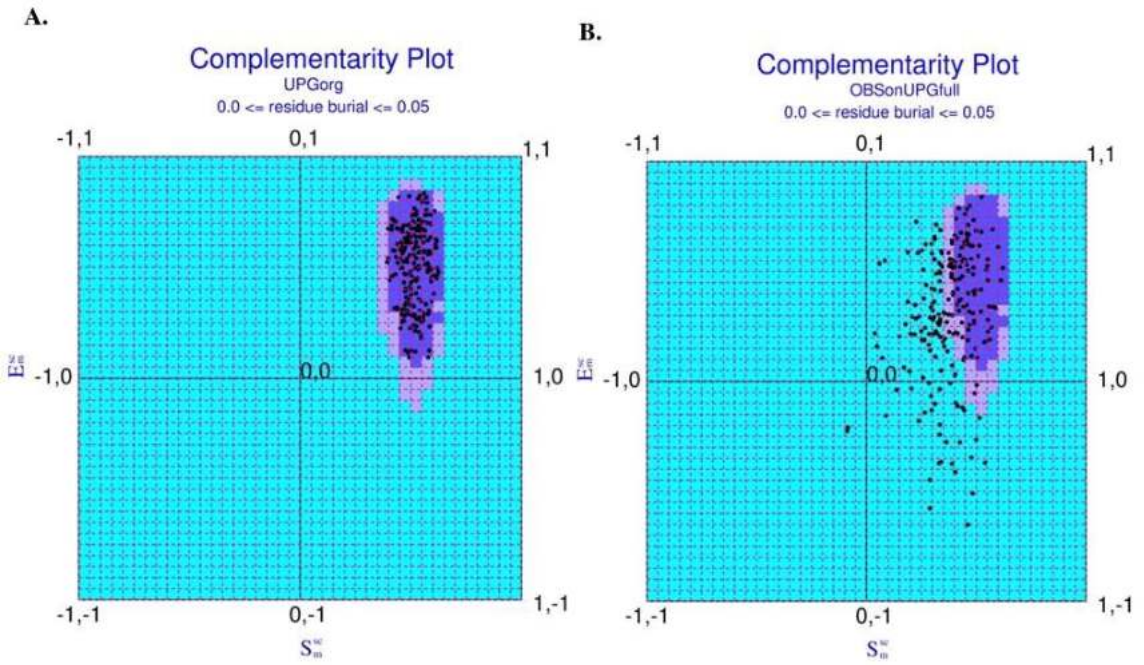


Fig.6. Distributions (in CP1) for residues with native side-chain conformers from the upgraded structures and replaced by rotamers from corresponding obsolete counterparts. (A) Distribution of residues with native side-chains all falling into the probable regions of CP1 and (B) distribution subsequent to the replacement.

Another calculation involving synthetic data, assigned a random rotamer combination **(Berkholz et al., 2009)** to a single specific buried residue (in PDB files from the database **SDB-3**) such that the native χ_1 was altered by more than 40°. Similar to the above situation, successful detection of error would be the transition of residues from the probable to the improbable regions of the CPs. Out of 1388 residues which were originally located in the probable region, 75.1% were relocated in the improbable region subsequent to the introduction of error followed by 10.3% in the less probable region

with the remaining 14.6% retained in the probable region as false positives **(Fig. 7)**. The CP-scores were CS_l : 3.15, ***rGb*** : 0.096 before and CS_l : -54.30, ***rGb*** : 0.1017 after the introduction of error whereas the χ^2 was found to be 6564.55 for the latter case.

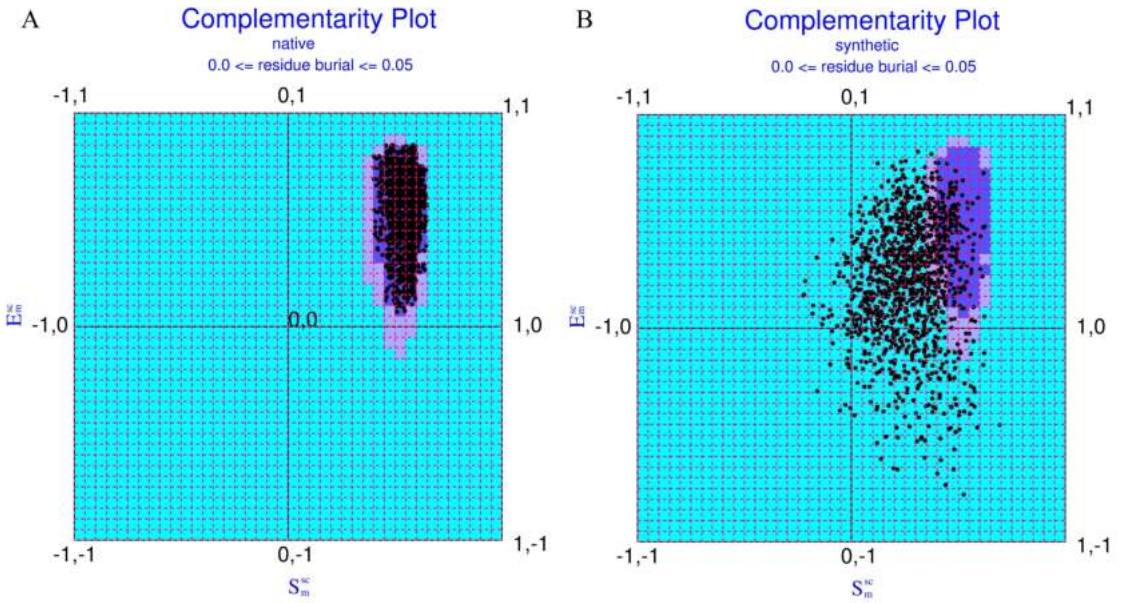


Fig.7. Distributions (in CP1) for residues with native side-chain conformers and replaced by random rotamers. (A) Distribution of residues with native side-chains all falling into the probable regions of CP1 and (B) distribution subsequent to the replacement.

3.4. Disqualifying retracted structures

A set of 28 retracted or suspected (obsolete without being superseded) crystal structures were subjected to a selection of validation protocols (Procheck, Clash-score from Molprobity and Whatcheck packing parameters) including CP. Structures which were either suspect in (complexed) ligand or contained embedded cofactors were not included in the calculation and for oligomeric proteins the largest polypeptide chain was retained. Procheck was used as an initial general filter and the remaining structures were specifically tested for packing defects by the other validation measures. A structure was considered to have passed the filters implemented in Procheck when all G-factor scores

were greater than -1.0 and 'INSIDE' recorded for bad contacts. The criteria for successful validation in the case of CP both with respect to the local (P_{count}) and global (***CS_l, rGb***) measures have already been mentioned and structures were considered to have passed the validation filter for Whatcheck when there was '*No series of residues with abnormal new packing environment'* and '*No stretches of four or more residues each having a packing Z-score worse than -1.75*' (Whatcheck output for packing parameter). A Clash-score (Molprobity) less than 20.0 was considered to be within the normal range. A total of 5 structures (1G40, 1G44, 2A01, 2ADH, 3KJ5) failed in all tests, whereas 15, 14, 4, 6, 5 were found to satisfy the validation criteria in Procheck, Whatcheck, Clash-score, P_{count} and (***CS_l, rGb***) respectively **(Table 3)**. Of the 15 structures (passing Procheck), 6, 11, 10, 11 failed in Whatcheck, Clash-score, P_{count} and (***CS_l, rGb***) respectively. Surface complementarity alone was also considered (P_{sm}) separately in order to specifically test for packing defects (by CP) in these structures. A total of 11 structures managed to exceed the threshold in P_{sm}. Again, 6 structures passed procheck and failed in P_{sm}. More importantly, there were 9 structures (1BEF, 1DF9, 2QID, 1RID, 1Y8E, 1S7B, 2F2M, 2CK9, 2MT2) which passed Whatcheck packing parameters, however failed to meet the threshold in P_{sm} indicative of packing defects which was also reflected in their high clash-scores. Thus, the performance of CP to detect anomalous packing in these retracted structures seem to be somewhat better than Whatcheck packing parameters and comparable to the Clash-score of Molprobity.

Table 3. Comparison of the different validation measures for retracted / suspected structures.

PDB ID	Resolution, R-facor	Procheck	Whatcheck-packing	Clash-score	P_{count}	(CS_l, rGb)	P_{sm}
1BEF	2.10, 0.186	+	+	-	-	-	-
1CMW	2.60, 0.192	+	-	-	-	-	+
1DF9	2.10, 0.199	-	+	-	-	-	-
2QID	2.10, 0.199	-	+	-	-	-	-
1G40	2.20, 0.198	-	-	-	-	-	-
1G44	2.60, 0.234	-	-	-	-	-	-
1L6L	2.30, 0.198	+	+	-	-	-	+
1RID	2.10, 0.206	-	+	-	-	-	-
1Y8E	2.20, 0.195	-	+	-	-	-	-
2A01	2.40, 0.228	-	-	-	-	-	-
2HR0	2.26, 0.180	+	-	-	-	-	+
1PF4	3.80, 0.240	+	-	-	-	-	-
1S7B	3.80, 0.320	+	+	-	-	-	-
1Z2R	4.20, 0.280	+	-	-	-	-	-
2F2M	3.70, 0.282	+	+	-	-	-	-
2A73	3.30, 0.233	+	+	-	-	-	+
2ADH	2.4, NULL	-	-	-	-	-	-
2CK9	2.85, 0.187	+	+	-	-	-	-
2MT2	2.30, NULL	-	+	-	-	-	-
2PZ3	2.42, 0.314	-	-	-	-	+	+
2QNS	3.00, 0.238	-	-	-	-	-	+
2RA7	1.99, 0.242	+	+	+	+	+	+
3A00	1.80, 0.222	+	+	+	+	+	+
3K78	2.80, 0.274	+	+	-	+	-	+
3KJ5	3.00, 0.366	-	-	-	-	-	-
3O7Y	2.41, 0.180	+	-	+	+	+	+
3O7Z	2.55, 0.183	+	-	+	+	+	+
3O8K	2.70, 0.268	-	-	-	+	-	-

Success or failure to meet the validation criteria (see **Text**) for all the measures is indicated by '+' and '-' respectively. Information regarding these retracted or suspected structures were obtained from http://main.uab.edu/Sites/reporter/articles/71570/**, Read et al., 2011** and ftp://ftp.wwpdb.org/pub/pdb/data/status/obsolete.dat.

3.5. Detection of low-intensity diffused errors

Since CPs are probabilistic in nature and are most effective when the entire polypeptide chain is taken into account, they should be able to detect an overall decline in the accuracy of the coordinates due to low-quantum random errors in geometrical parameters diffused over the entire structure. To probe the performance of CPs in such

circumstances, random errors were incorporated throughout the fold in preselected geometrical parameters: (i) approximately 1.5 - 2.5° for main-chain bond angles (±0.5σ **(Engh and Huber, 2001)**) and (ii) ±1° for (φ, ψ). 30 high-resolution structures from **SDB-2** were used for these calculations and 20 erroneous models generated per native structure for each of the geometrical parameters leading to 600 models per set. From this set, 142, 152 files (main-chain bond angles, (φ, ψ)) passed the validation filters (criteria stated in the previous section) in Procheck. The average all-atom RMS deviations of these models with respect to their corresponding native structures were 1.89Å ±0.71 and 1.67 Å ±0.56 respectively. Of these 108, 109 files failed to meet the criteria for successful validation in CP with 78, 77 registering negative values for at least one of the two **(*CS$_I$*, *rGb*)** scores.

3.6. Probing the role of deviations in maintaining structural integrity

One of the questions addressed in this work is the contribution of deviations (in geometrical parameters) in maintaining structural integrity of the native fold. For this purpose, 20 high resolution structures (**SDB-1**)**,** spanning the four major protein classes and ranging from 56 to 363 residues in chain length were selected and the structures rebuilt (see **Materials and Methods**) by reverting all main-chain bond lengths, angles and ω-torsions to their corresponding unimodal ideal values (as tabulated in Procheck **(Laskowski et al., 1993)**, ω: Whatif **(Vriend, 1990)**), while retaining native values for all other dihedral angles (φ, ψ, χ). This led to such large-scale distortions in the idealized structures (with respect to the original native model) that often their (C^{α}) RMSDs exceeded 10 Å **(Fig. 8)**.

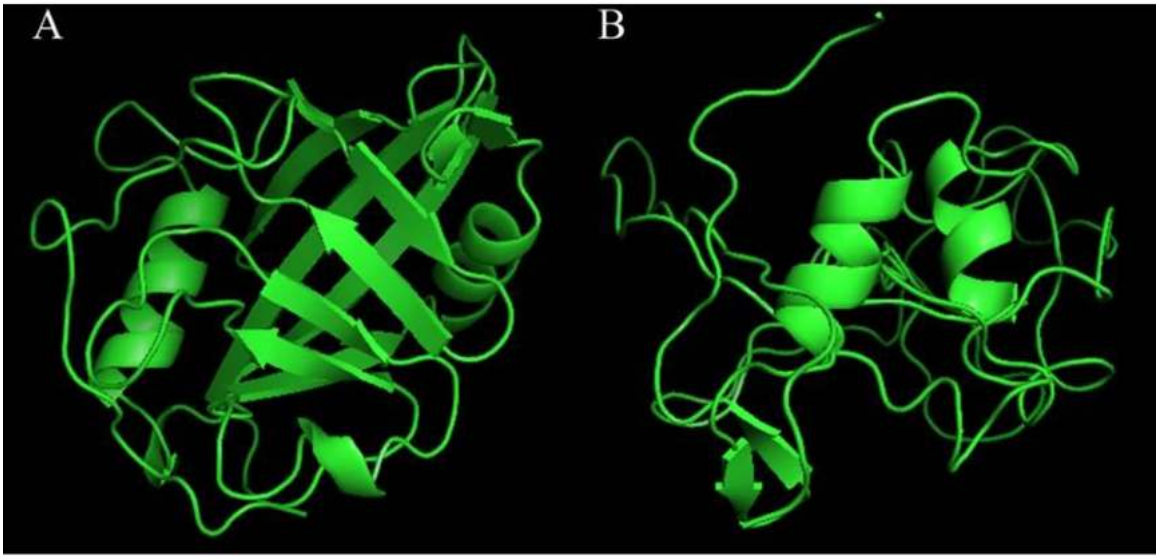


Fig. 8. Distortions in the native fold due to the reversal of all main-chain bond lengths, angles and ω-torsions to their corresponding (unimodal) ideal values. (A) the native structure of cyclophilin from *L. donovani* (2HAQ) and **(B)** its corresponding idealized structure (C^{α}-RMSD: 12.86 Å, calculated at one-to-one atomic correspondence). Figure constructed by **PyMol** [http://www.pymol.org/].

Although the degree of structural distortions is estimated by the RMSDs, its effect on packing and electrostatics can be conveniently assessed using the CP measures. The distortions were more pronounced for larger polypeptide chains (~100 residues or more in length) due to the accumulation of a higher number of angular idealizations. Also, proteins containing greater β-sheet content had more severe deformations most probably rationalized **(Hooft et al., 1996)** by the distribution in N-C^{α}-C (τ) angle with respect to secondary structure. The procedure also led to a sharp decline in $\boldsymbol{CS_I}$ (-10.54, σ = ±3.48)

averaged over all 20 structures, relative to their corresponding native values (2.47 ±0.41 respectively: **Table 4**).

Table 4. Complementarity and Accessibility scores for idealized structures. Average scores (***CS$_I$***, ***rGb***) and standard deviations (in parentheses) obtained for different forms of idealization on the database **SDB-1**. The same scores have also been tabulated for the native proteins in the original databases **DB2** and **SDB-1.** Ideal values for pre-selected geometrical parameters were obtained from **Engh and Huber, 2001;** Whatif **(Vriend, 1990)** or a Conformation Dependent Library **(CDL) (Berkholz et al., 2009)**

Idealization protocol	***CS$_I$***	***rGb***
DB2 (≤ 2 Å, 400)	2.24 (0.48)	0.055 (0.022)
SDB-1 (≤ 2 Å, 20)	2.47 (0.41)	0.060 (0.020)
Main-chain bond-lengths [a], angles [a] and ω [b] idealized	-10.54 (3.48)	0.000 (0.031)
Main-chain bond-lengths [a], angles [a] and ω [b] idealized and energy-minimized with flexible backbone	-2.58 (2.61)	0.004 (0.030)
Main-chain bond-lengths [a], angles [a] idealized (with native ω)	-10.52 (3.80)	0.009 (0.03)
Main-chain bond-angles [c] idealized (with native ω)	-8.98 (3.87)	0.017 (0.028)
Main-chain bond-angles [c] idealized (with native ω), energy-minimized with rigid backbone	-1.42 (2.59)	0.019 (0.025)
Main-chain bond-lengths [a] idealized	2.45 (0.36)	0.060 (0.020)
Main-chain bond-angles [a] idealized	-10.56 (3.75)	0.010 (0.030)
ω idealized [b]	-7.80 (3.80)	0.022 (0.030)
Main-chain bond-angle: N-C^{α}-C (τ) [a] idealized	-7.80 (3.95)	0.031 (0.027)
Main-chain bond-angle: C^{α}_i-C_i-N_{i+1} [a] idealized	-4.98 (4.73)	0.047 (0.026)
Main-chain bond-angle: C_{i-1}-N_i-C^{α}_i [a] idealized	-3.95 (3.36)	0.037 (0.030)

Little or no improvement was observed in the quality of the rebuilt structures by either retaining native ω values or utilizing ideal values (for bond angles) derived from a conformation dependent library (CDL) **(Berkholz et al., 2009) (Fig. 9)**.

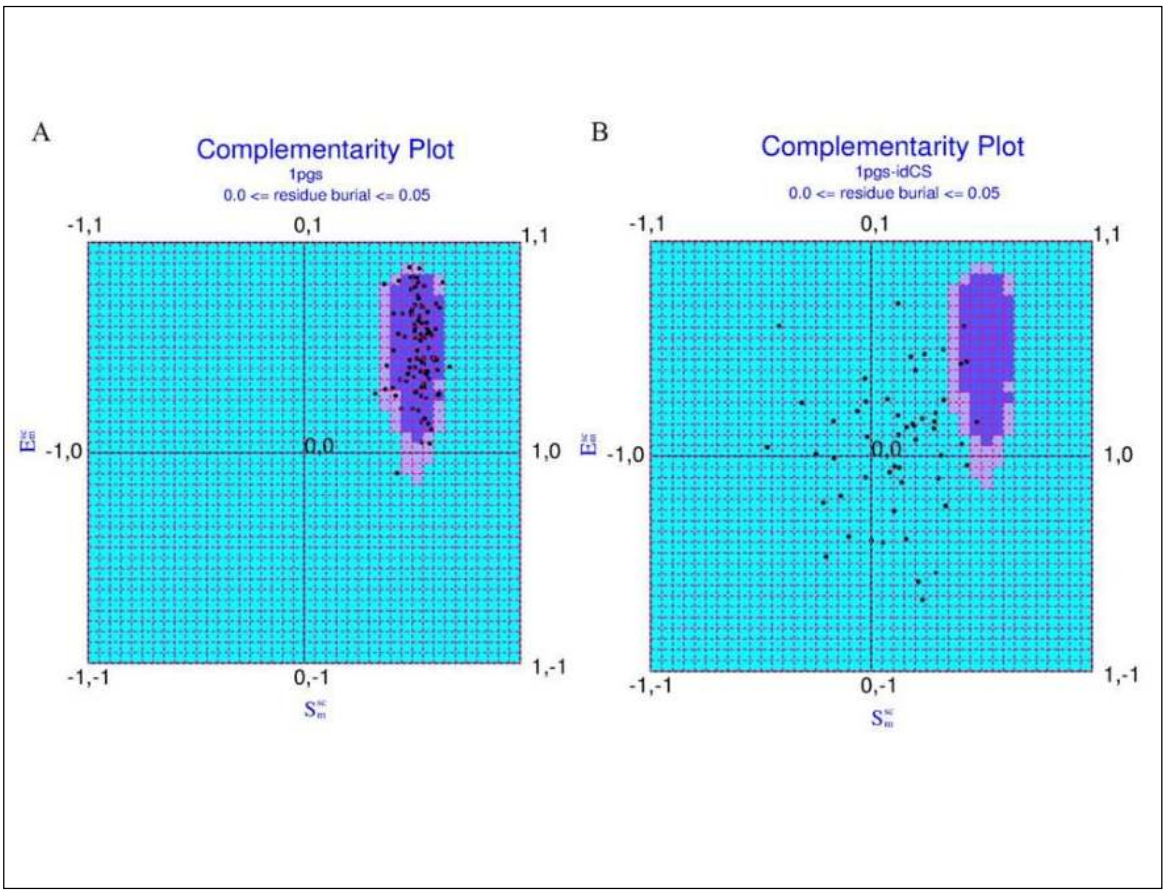


Fig.9. Effect of CDL-idealization probed by CP. Distribution for **(A)** the native polypeptide chain (1PGS) and **(B)** its corresponding idealized structure generated utilizing CDL ideal values.

The values for ***rGb*** (0.000 ±0.031) for the idealized structures were also substantially reduced as structural distortions often led to the exposure of hydrophobic residues to the solvent. Energy minimization subsequent to idealization improved the complementarity scores (***CS$_I$*** : -2.58 ±2.61) even though they were still significantly less than their corresponding native values, with a surge in their standard deviations **(Fig. 10)**.

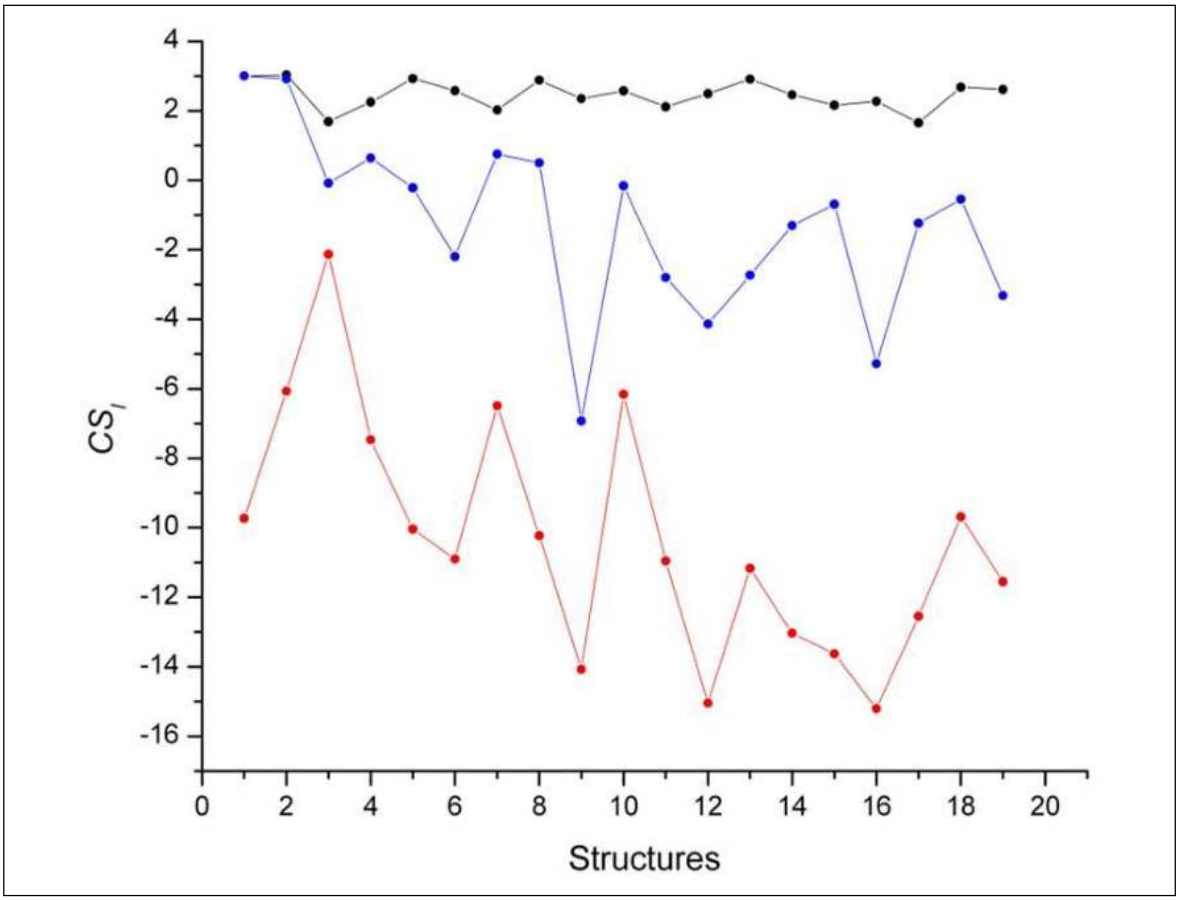


Fig.10. CS_I scores for structures before and after idealization. 20 structures (from **SDB-1**) with increasing chain-length along the X-axis. Black : native; red : idealized; and blue : idealized + energy-minimized.

The substantially low values for ***rGb*** remained unaltered even after energy minimization, indicative of hydrophobic residues still remaining exposed to the solvent. Minimization also did not improve the (C^{α}) RMSDs (calculated at a one-to-one atomic correspondence subsequent to superposition by Dali server **(Holm and Rosenstrom, 2010)** between native and idealized coordinates, which in some instances could not even be superposed onto each other (**Table 5**).

Table 5. Structural distortions due to idealization as reflected in the RMSDs. RMSDs calculated between C^{α} atoms of idealized (all main-chain bond lengths, bond angles and ω) and the native coordinates (calculated at a one-to-one atomic correspondence) subsequent to superposition by Dali server. The same calculation was repeated for energy minimized coordinates subsequent to idealization. '-' stands for non-superposable structures.

PDB ID	RMSD (Å) [a]	
	Idealized vs. native	Idealized & Energy Minimized vs. native [b]
1AKO	13.98	-[c]
1BGF	7.30	6.77
1CEM	16.61	17.35
1CHD	22.42	22.18
1CKA	3.05	3.10
1ERZ	24.44	22.78
1HBQ	22.60	-
1IFC	11.30	11.48
1LMB	4.56	4.56
1MKB	-	-
1MLA	23.33	22.05
1PDO	4.56	5.29
1PGS	-	-
1SFP	-	-
1SRV	13.82	-
1STN	18.02	18.10
1UBI	4.31	4.14
2CPL	12.52	12.58
2END	7.64	7.41
2LIS	9.09	9.11

Thus, in summary, in no case could the original structure be reconstituted by any form of energy minimization of the idealized coordinates. Calculations using both unimodal and CDL ideal values were repeated on a larger dataset of 68 ultrahigh resolution (≤ 1 Å) structures (**SDB-3**), which gave similar pattern of results (**Table 6**).

Table 6. Complementarity and accessibility scores for idealized structures of ultra-high resolution. Average scores (CS_I, *rGb*) standard deviations (in parentheses) for the idealized structures. Structures idealized by different methods from a database of 68 ultra-high resolution structures (**SDB-3**). Unimodal and CDL ideal values obtained from **Engh and Huber, 2001** and a Conformation Dependent Library **(CDL) (Berkholz et al., 2009).**

Parameters used for Idealization	CS_I	*rGb*
Unimodal ideal values	-9.82 (3.75)	-0.009 (0.032)
CDL ideal values	-6.64 (4.12)	0.024 (0.003)

To determine the relative contribution of each geometrical parameter in the distortions of the reconstituted (idealized) polypeptide chains, calculations (from **SDB-1**) were repeated by individually idealizing bond lengths, angles and ω in turn, while retaining native values for all other parameters. Idealizing bond lengths were found to cause no significant distortions while all the angular parameters played an influential causal role in giving rise to structural deformations. Idealizing either τ or ω was found to have a more pronounced effect on the distortions amongst all other angular parameters **(Table 4)**.

Another method adopted for idealization involved regularizing the whole protein structure using the 'REFI' routine in the software 'O' (see **Materials and Methods**). 110 pairs of obsolete and upgraded structures **(OUDB)** were used for this calculation and their side-chain RMS deviations computed along with the Complementarity scores before and after the regularization. For both the sets (Obsolete, Upgraded), the procedure did not lead to any substantial structural alterations w.r.t the original coordinates borne out by both side-chain RMS deviations (Obsolete: 0.252 Å ± 0.153, Upgraded: 0.205 Å ± 0.147) and the scores (CS_I) before (Obsolete: 0.229 ± 2.49, Upgraded: 1.95 ± 0.67) and after (Obsolete: -0.09 ± 2.77, Upgraded: 1.65 ± 0.84) the regularization.

3.7. Detection of unbalanced charges in the protein interior

CP takes into account long range electrostatics of the whole protein molecule as part of its validation protocol. In order to examine the additional efficacy of this feature in error detection (involving misidentification of side-chains) native sequences of 93 structures (**SDB-3 & SDB-4**) were redesigned by switching polar or charged to hydrophobic residues and vice versa. All deeply or partially buried residues from a chosen set of amino acid identities (***Bur*** $\leq$ 0.30) were changed to those of an altered hydrophobic character, though similar in size and shape in most of the cases: Ala → Ser, Ser ↔ Cys, Thr ↔ Val, Phe ↔ Tyr, Leu → Asn (transition probability : 0.5), Leu → Asp (0.5), Ile → Met, Met → Ile (0.5), Met → Arg (0.5), Glu → Arg (0.5), Glu → Gln (0.5), Asp ↔ Asn, Arg → Met (0.5), Arg → Glu (0.5). Side-chains of these designed sequences were then threaded onto the native backbone using SCWRL4.0 and the resulting structures were energy minimized with flexible backbones, subsequent to hydrogen fixation by REDUCE **(Word et al., 1999)**. Molprobity was used to ensure that the redesigned models were devoid of errors / outliers in the other validation parameters. All 93 redesigned structures passed all the validation filters in Molprobity with minimum Clash scores (1.26 ±0.64; <percentile>: 98.23 ±1.55) and satisfying all other validation filters, reflected in the overall Molprobity scores (1.00 ±0.27; <percentile>: 99.38 ±1.21). Although, $\boldsymbol{CS_l}$ dropped to 0.36 ±1.23; w.r.t. native ($\boldsymbol{CS_l}$**:** 2.21 ±0.62), the polar to hydrophobic transitions (or vice versa) were naturally captured in the poor ***rGb*** scores (0.005 ±0.026) reflecting non-native like distribution of amino acids (native: 0.054 ±0.026) with regard to burial and also in the distribution of suboptimal points primarily with regard to E_m^{sc} **(Fig. 11)**.

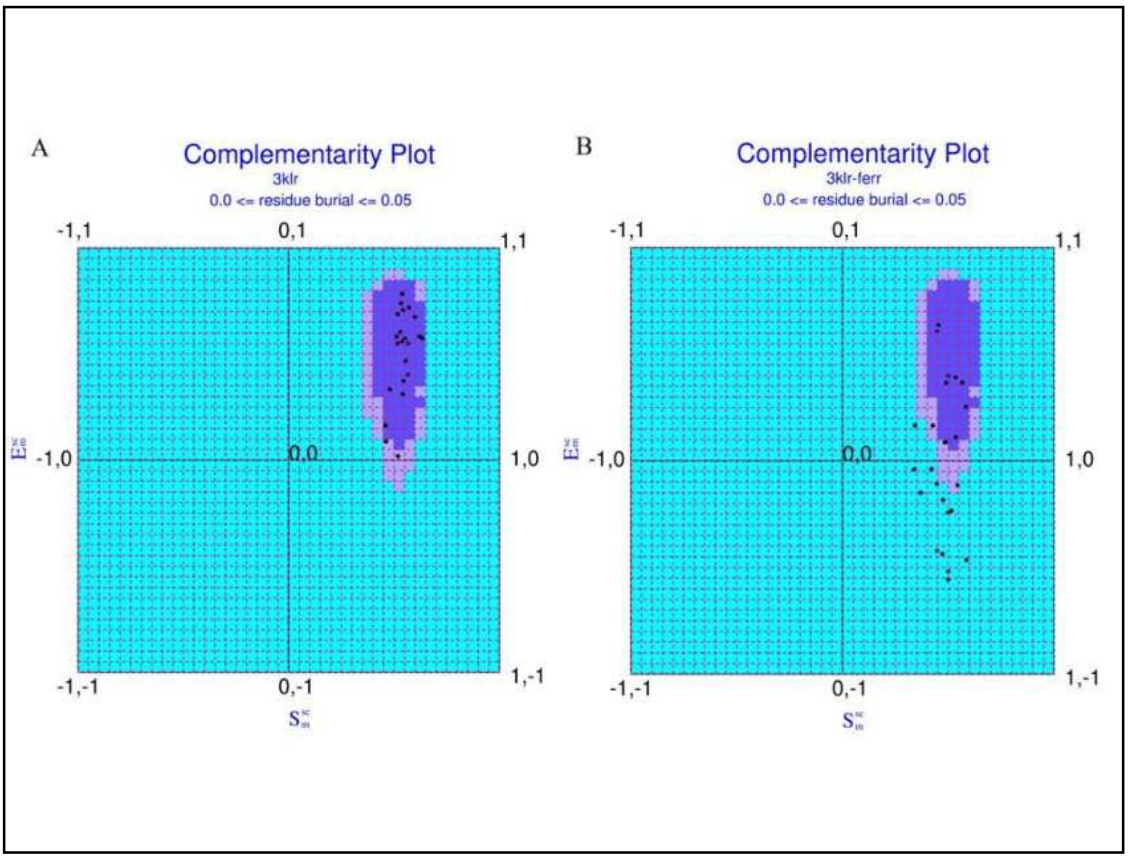


Fig.11. Ability of CP to detect residues with unbalanced charges in the protein interior. (A) Native distribution of 3KLR in CP1 and **(B)** subsequent to the 'polar to hydrophobic' transitions. All buried residues have been included in the plot. As can seen from the plot, the mutated residues have a tendency to be found in the improbable region (suboptimal for E_m^{sc}).

74 redesigned models failed to meet the criteria for successful validation (in CP) whereas 58 registered negative values in at least one of the two (***CS$_l$, rGb***) scores (**Fig. 12**). On the other hand, consideration of the 'local' score (P_{count}) led to the rejection of 77 structures. By considering electrostatic complementarity alone, (P_{Em}) 66 structures failed to meet the threshold criteria. 198 unfulfilled hydrogen bonds (for buried residues) were detected by Whatcheck in the native structures which increased to 1160 for the redesigned models demonstrating a comparable ability of Whatcheck and CP to detect such errors. 82 redesigned models had more than 2 (average obtained from native)

unfulfilled hydrogen bonds over and above the native. Thus, the local electrostatic parameters of CP and Whatcheck appear to perform comparably.

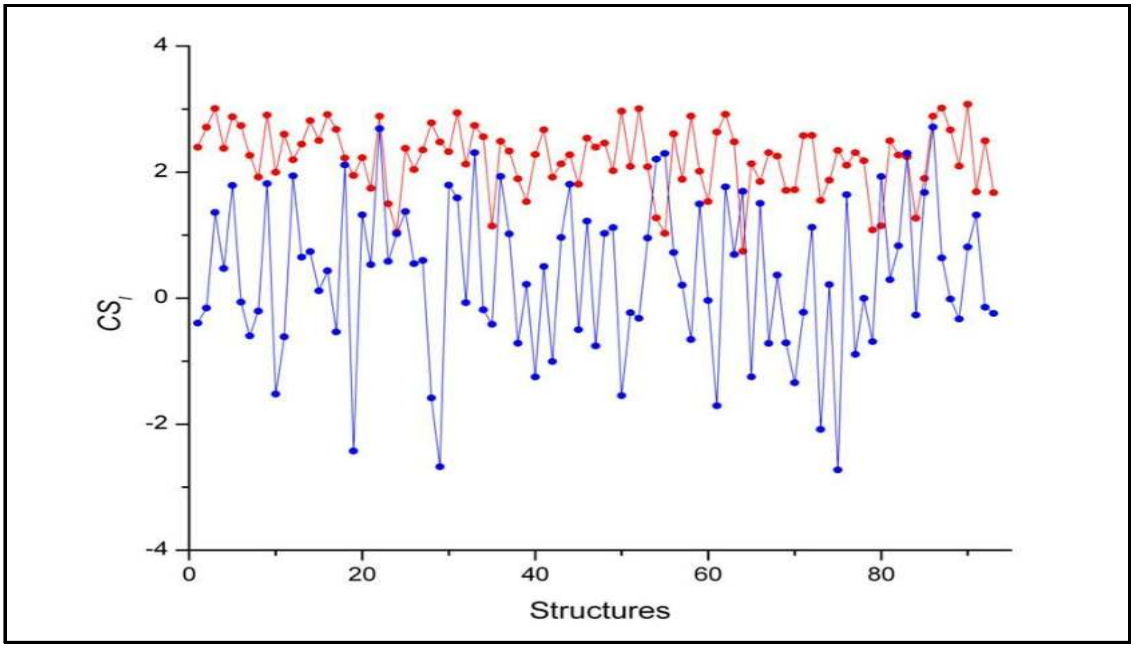


Fig.12. CS_l scores for native and corresponding redesigned structures. The native CS_l scores for 93 structures (from **SDB-3** and **SDB-4**) plotted in red along with those subsequent to the 'hydrophobic to hydrophilic' transitions plotted in blue.

An attempt was also made to detect single point mutations Val → Thr and Thr → Val (which could be interpreted as a sequencing error or accidental mutant) conserving all other parameters, in 25 high resolution structures (**SDB-4**). Molprobity was used to ensure that the redesigned models were validated in the other parameters before and after the *in-silico* mutation (see **Materials and Methods**). Initially, all these deeply buried residues (Val: 227, Thr: 136 from burial bin 1) were in the probable regions of the plot which relocated to the improbable regions in 31.3% and 23.5% (less probable: 19.6% & 26.5%) of the cases respectively, upon mutation. 26.4% of the altered threonines (Val → Thr) were also detected to have unfulfilled hydrogen bonds in their side-chains by Whatcheck.

3.8. Quality assessment of homology models

Finally, the method was tested on homology models (20 folds) with templates of varying sequence identity (w.r.t. the modeled sequence; ranging from 13.5% to 90.3%). Both ***CS$_{I,,}$ rGb*** correlated fairly well with sequence identities and somewhat better with sequence similarities of the modeled sequences **(**see **Table S1** in **Supplementary Information** in CD enclosed**)**. The (non-linear) correlation of ***CS$_I$*** with both sequence identity and similarity were best fitted to cubic-polynomial curves with R^2 of 0.69 and 0.72 respectively (**Fig. 13**).

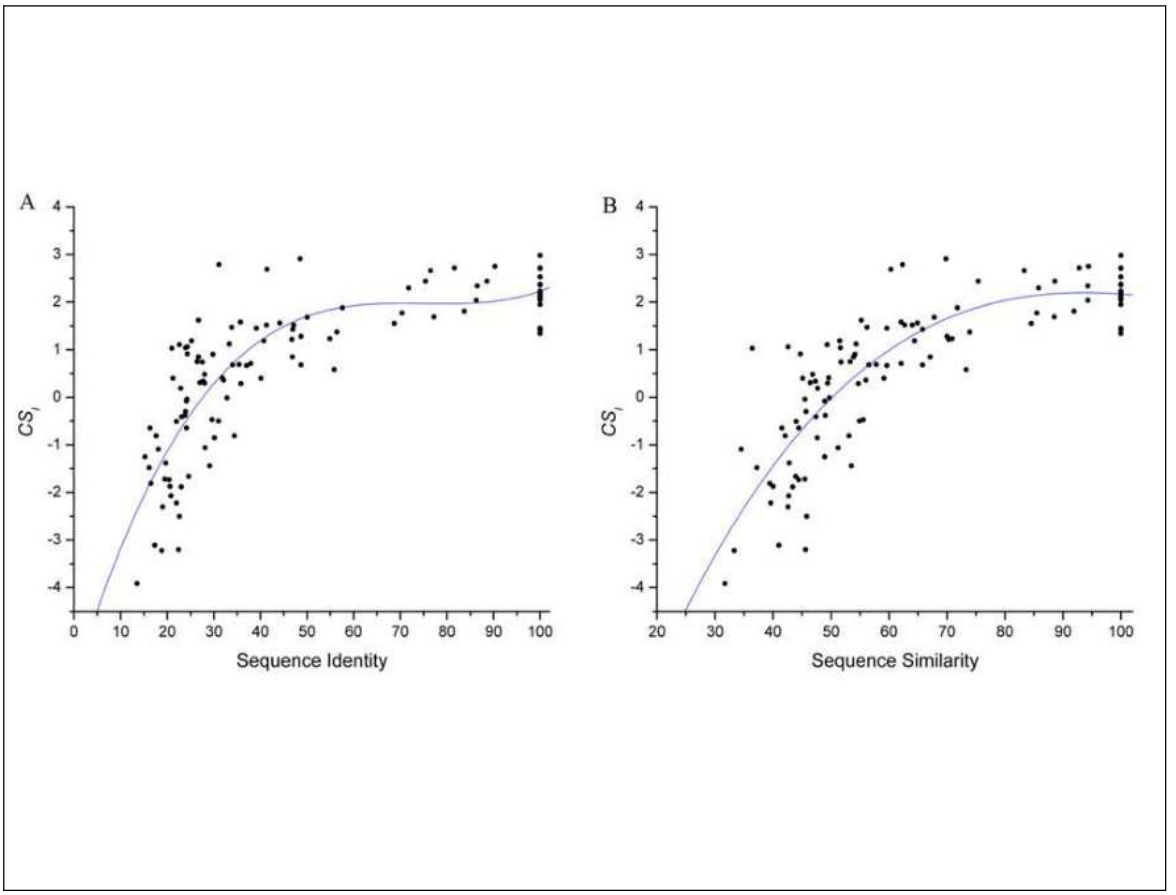


Fig.13. *CS$_I$* scores for homology models as a function of sequence identity and similarity. Both distributions are best fitted to cubic polynomial curves with R^2 of (A) 0.69 for identity and (B) 0.72 similarity respectively.

Interestingly, there was a significant improvement in the scores upon energy minimization of the models obtained from Accelerys (Modeller). On an average, there was an increase of ~150 to 175% in the CS_I scores, for the models before and after energy minimization. Generally, a fairly steep decline in CS_I was noted below 30% sequence identity, even though 8 out of 47 such models were found above the CS_I cutoff (0.80) for successful validation. Thus, the scores could definitely be used as measures, either to judge the overall quality of the models or the appropriate choice of the template. CP was then compared with the Modeller-DOPE score which also provides a measure of complementarity in the interior of protein structural models. 22 homologous structures of 2HAQ (Cyclophilin-like-fold) were assembled ranging in sequence identity from 17 to 74%. Homology models were built using these sequences with 2HAQ as a template in Accelerys (Modeller), which provided their DOPE scores. Both the scores gave a significant correlation with sequence identities w.r.t. the template (CS_I: 0.79; DOPE: -0.66, **Fig. 14**), their mutual correlation being -0.51. However, unlike CS_I which is normalized over the entire polypeptide chain, the DOPE score gave almost zero correlation (-0.12, calculated on 50 models) when estimated over a collection of folds. The methodology was also compared with QMEAN **(Benkert et al., 2009)** which is reportedly sensitive for estimating the proximity of models to a native target. QMEAN global scores were estimated for the native folds and their corresponding homology models for each set and compared with CS_I. There was appreciable agreement between the two sets of scores exhibited by high correlation coefficient (0.866 calculated over 120 models) between them.

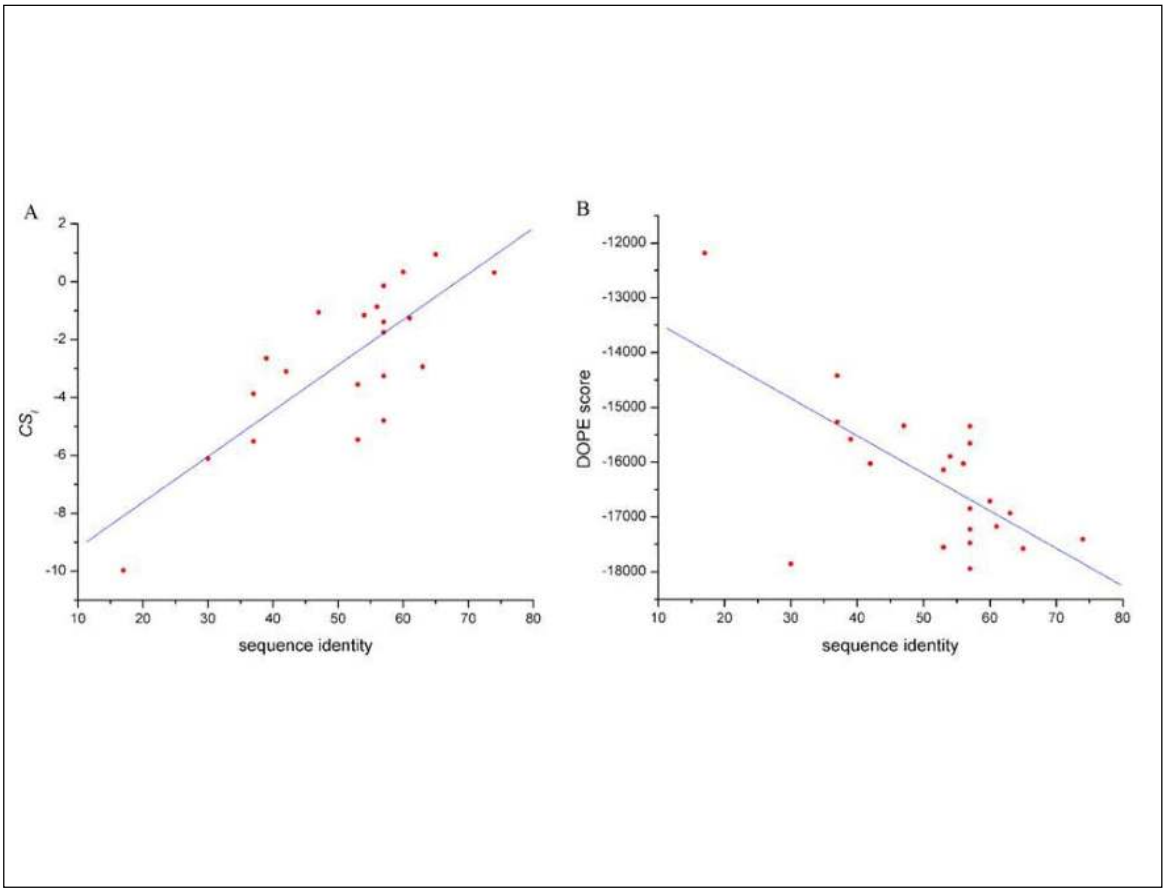


Fig.14. CS_l and DOPE score (Modeller) for homology models (built on the template 2HAQ) as a function of sequence identity. The Pearson's correlation with sequence identity for (A) CS_l and (B) DOPE-score are 0.79 and -0.66 respectively.

4. Conclusions:

The Complementarity Plot as a validation technique is probabilistic in nature and can be utilized either over the full chain, or on any distribution of points. In addition, the quality of packing and electrostatics of a local region in a protein can also be assessed. This user defined choice of either 'global' or 'local' measures gives considerable flexibility in the use of the plot. With regard to the final validation results, good agreement was established between the local ($\mathbf{P_{count}}$) and global (***CS$_l$, rGb***) scores in the

CP methodology. Further, this is the only validation procedure which combines both packing and electrostatics in a single unified measure and displays graphically (apart from actually listing) residues with faulty packing or electrostatics. Since the plot essentially has to do with the packing and electrostatics of side-chain atoms, it performs fairly well in the detection of errors involving side-chain torsion angles. Further, it finds its application in the detection of packing anomalies in retracted structures. It is also especially sensitive to low-intensity errors in main-chain geometrical parameters diffused over the entire polypeptide chain, which could arise either due to low resolution, sub-standard data, model-bias or a host of other factors. The current work clearly indicates that over and above the commonly used validation techniques, the quality of packing within proteins and the global electrostatics should be included separately in any validation package. Calculations involving residue swapping from hydrophilic to hydrophobic or vice versa indicates that the methods developed could be particularly effective in protein (full sequence / core) design, wherein multiple mutations could accompany the design process. The CP could also be a sensitive indicator of the correct choice of template in homology modeling.

5. Program availability:

The standalone suite of programs (**Sarama**) for the Complementarity Plot (Linux Platform) with detailed features and documentation is available at: http://www.saha.ac.in/biop/www/sarama.html

References

Banerjee R, Sen M, Bhattacharyya D, Saha P. (2003). **The Jigsaw Puzzle Model: Search for Conformational Specificity in Protein Interiors.** *J. Mol. Biol*. **333:** 211–226.

Basu S, Bhattacharyya D, Banerjee R. (2012). **Self-Complementarity within Proteins: Bridging the Gap between Binding and Folding.** *Biophys. J.* **102:** 2605-2614.

Berman HM, Henrick K, Nakamura H, (2003). **The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.** *Nature Struct. Biol* **10**: 98.

Berkholz DS, Shapovalov MV, Dunbrack RL, Karplus PA. (2009). **Conformation Dependence of Backbone Geometry in Proteins.** *Structure* **17:** 1316-1325.

Benkert P, Kunzli M, Schwede1 T (2009). **QMEAN server for protein model quality estimation.** *Nucl. Acids. Res.* **37:** W510–W514.

Bradley P, Misura KMS, Baker D (2005). **Toward high-resolution de novo structure prediction for small proteins.** *Science* **309:** 1868-1871.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. (1983). **CHARMm: A program for macromolecular energy, minimization, and dynamics calculations.** *J. Comput. Chem.* **4:** 187-217.

Caravella JA. (2002). **Electrostatics and packing in biomolecules: accounting for conformational change in protein folding and binding.** PhD thesis. Massachusetts Institute of Technology, Cambridge.

Chang G, Roth CB, Reyes CL, Pornillos O, Chen Y-J, and Chen AP (2006). **Structure of MsbA from E. coli: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters.** Retraction. *Science* **314**: 1875.

Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, III, Snoeyink J, Richardson JS, Richardson DC, (2007). **MolProbity: all-atom contacts and structure validation for proteins and nucleic acids.** *Nucl. Acids. Res.* **35:** W375–W383.

Dunbrack RL, Jr., and Karplus M. (1993). **A backbone dependent rotamer library for proteins: application to sidechain prediction.** *J. Mol. Biol.* **230**: 543-571.

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

Engh RA, and Huber R (2001). **International Tables for Crystallography. In International Tables for Crystallography,** M.G. Rossmann and E. Arnold, eds. (Dordrecht, The Netherlands: Kluwer Academic Publishers), pp. 382-392.

Hanson MA, and Stevens RC (2000) **Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 A resolution.** Retraction. *Nat. Struct. Biol.* **7**: 687-692.

Heifetz A, Katchalski-katzir E, and Eisenstein M. (2002). **Electrostatics in protein–protein docking.** *Protein Sci,* **11**: 571–587.

Holm L, and Rosenstrom P. (2010). **Dali server: conservation mapping in 3D.** *Nucl. Acids. Res.* **38:** W545–549.

Hooft RWW, Sander C, and Vriend G, (1996). **Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures.** *Proteins* **26**: 363-376.

Hooft RWW, Vriend G, Sander C, and Abola EE, (1996) **Errors in protein structures.** *Nature* **381**: 272

Janssen BJC, Read RJ, Brunger AT, Gros P (2007). **Crystallographic evidence for deviating C3b structure?** *Nature* **448**: E1-E2, discussion E2-E3.

Jaskolski M, Gilski M, Dauter Z, Wlodawer A, (2007). **Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them?** *Acta Cryst D* **63:** 611-620.

Jones TA, Zou JY, Cowan SW, Kjeldgaard M. (1991). **Improved methods for building protein models in electron density maps and the location of errors in these models.** *Acta Cryst A*47: 110-119.

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL (2008). **NCBI BLAST: a better web interface.** *Nucl. Acids. Res.* **36:** *W5-W9.*

Kleywegt GJ, Jones TA, (1996). **Phi/Psi-chology: Ramachandran revisited.** *Structure* **4:** 1395-1400.

Kleywegt GJ. (2000). **Validation of biomacromolecular structures – lessons learned from X-ray crystallography.** *Acta Crystallogr* D **56**: 249-265.

Krivov GG, Shapovalov MV, Dunbrack RL. Jr. (2009). **Improved prediction of protein side-chain conformations with SCWRL4.** *Proteins.* **77**: 778-795.

Laskowski RA, MacArthur MW, Moss DS, Thornton JM. (1993). **PROCHECK: a program to check the stereochemical quality of protein structures.** *J. Appl. Crystallogr.* **26**: 283-291.

Lawrence MC, and Colman PM. (1993). **Shape complementarity at protein/protein interfaces.** *J Mol Biol.* **234**: 946–950.

Lee B, and Richards FM. (1971). **The interpretation of protein structures: Estimation of static accessibility.** *J. Mol. Biol.* **55:** 379-400.

Liang S, Grishin NV. (2002). **Side-chain modeling with an optimized scoring function.** *Protein Sci.***11**: 322-331.

Lovell SC, Davis IW, Arendall WB III., de Bakker PIW, Word, JM, et al. (2003). **Structure Validation by Cα Geometry: φ,ψ and Cβ Deviation.** *Proteins: Struct. Funct. Genet.* **50**: 437-450.

Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, and Ten Eyck LF. (2001). **Protein docking using continuum electrostatics and geometric fit.** *Protein Eng.* **14**: 105–113.

McCoy AJ, Epa VC, and Colman PM. (1997). **Electrostatic complementarity at protein/protein interfaces.** *J Mol Biol.* **268**: 570-584.

McDonald IK, and Thornton JM (1995). **The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains.** *Protein Eng* **8:** 217-224.

Murzin AG, Brenner SE, Hubbard T, Chothia C. (1995). **SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures.** *J. Mol. Biol.* **247:** 536-540.

Pontius J, Richelle J, Wodak SJ. (1996). **Deviations from standard atomic volumes as a quality measure for protein crystal structures.** *J. Mol. Biol.* **264**: 121-136.

Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963). **Stereochemistry of polypeptide chain configurations.** *J. Mol. Biol.* **7**: 95-99.

Ramachandran GN, Sasisekharan V (1968). **Conformation of polypeptides and proteins.** *Adv. Protein. Chem.* **23**: 283-437.

Read RJ, Adams PD, Arendall WB, III, Brunger AT, Emsley P, et al. (2011) **A New Generation of Crystallographic Validation Tools for the Protein Data Bank.** *Structure* **19**: 1395-1412.

Rohl CA, Strauss CEM, Misura KMS, Baker D (2004). **Protein structure prediction using Rosetta.** *Method. Enzymol.* **383**: 66-93.

Rocchia WS, Sridharan A, Nicholls E, Alexov, A. Chiabrera and B. Honig. (2002). **Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects.** *J. Comput. Chem*. **23**: 128-137.

Shapovalov MS, and Dunbrack RL, Jr. (2011). **A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions.** *Structure*, **19**: 844-858.

Touw WG, Vriend G (2010). **On the complexity of Engh and Huber refinement restraints: the angle tau as example.** Acta Cryst D **66:** 1341-1350.

Vriend G, Sander C (1993). **Quality control of protein models: Directional atomic contact analysis.** *J. Appl. Crystallogr.* **26:** 47-60.

Vriend G. (1990). **WHAT IF a molecular modeling and drug design program.** *J. Mol. Graph.* **8:** 52-55.

Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, Wishart DS (2003). **VADAR: A web server for quantitative evaluation of protein structure quality.** *Nucl. Acids. Res.* **31:** 3316-3319.

Word JM, Lovell SC, Richardson JS, Richardson DC (1999). **Asparagine and Glutamine: Using Hydrogen Atom contacts in the choice of side-chain amide orientation.** *J. Mol. Biol.* **285:** 1735–1747.

Computational design of the hydrophobic core of a beta-barrel protein

1. Introduction.

Protein design is a growing field of research in computational and experimental biophysics particularly related to the 'inverse protein folding problem'. The problem concerns identifying protein primary sequences consistent with and supportive of a given fold, an idea which has found considerable application in the *de novo* design of targeted protein structures. Classically, from a purely structural point of view, design could be either of the hydrophobic core **(Lazar et al., 1997; Tsai et al., 1997; Johansson et al., 1998; Munson et al., 1996; Kashiwada et al., 2000)** or of the full protein sequence **(Shah et al., 2007; Fung et al., 2008)**. In recent years it has been somewhat guided by strong biomedical and industrial interest leading to the engineering of protein hormones and enzymes to perform existing functions under a wide range of conditions or to perform entirely new functions. Protein engineers are also attempting for the possible construction of a range of self-organizing macromolecules **(Street and Mayo, 1999)** which might come out successful in the far-future. However, the current state-of-the-art is to redesign portions of globular proteins to insert particular motifs, increase thermal stability or to modify functions. Successful applications in the field include engineering metal-binding centers **(Lu and Valentinet, 1997)** and the introduction of disulfide bonds **(Chakraborty et al., 2005; Das et al., 2007; Indu et al., 2010)**. As far as full sequence design is concerned, earlier attempts typically led to poorly defined states or molten globules, instead of a single target fold. However, over the past two decades considerable success has been achieved.

The theoretical and computational aspects of protein design concern the involvement of two major steps 1) sampling methods and 2) fitness functions. Genetic algorithms subsequent to random sampling is generally used to generate a wide range of sequences whereas monte-carlo methods have been used extensively for side-chain conformer sampling. Fitness functions are generally integral and subsequent to the sampling procedure making the overall computational pipe-line to run in cycles until some pre-decided threshold of convergence is attained. Traditionally, statistical potentials

or pseudo-energy functions have been used to rank the desirability of each amino acid sequence for a particular backbone structure in atomistic protein design. The filtered sequences then require experimental validations. Thus in a 'protein design cycle' **(Street and Mayo, 1999)**, an energy expression is used to determine plausible sequences which are subsequently synthesized and tested in the laboratory. Depending upon the correlation between the computed and experimentally determined properties of the designed sequences, terms are either added to or eliminated from the pre-existing energy functions to generate new sequences completing the cycle. The current chapter describes a computational method to re-design the hydrophobic core of a beta barrel single domain protein (Cyclophilin A of *Lieshmania donovani* : 2HAQ) where conventional random sampling methods have been used along with novel fitness functions based on complementarity and network analysis.

2. Materials and Methods

2.1. Random Sampling

The hydrophobic core of cyclophilin was identified by calculation of residue solvent accessibilities, contacts and visual inspection. The core was found to be constituted of 18 residues in all (see **Results**). Random sampling was then performed from these shortlisted hydrophobic residues at each core position. A total of 10^5 sequences were sampled wherein all short listed residues were sampled nearly equally at each position.

2.2. Threading and Minimization

After the first round of screening at the sequence level (by applying cutoffs in van der Waals envelope volume, sequence entropy, secondary structural propensity: see **Results**) the full chain of the filtered sequences contained 1 to 18 alterations in the core w.r.t. the native. These were then threaded onto the native backbone with their side-chain torsions being optimized by SCWRL4.0 **(Krivov et al., 2009)**. Hydrogen atoms were

then removed and rebuild by REDUCE. Subsequent to another round of screening at this stage (by applying cutoffs in packing densities, short contacts and presence of probable disulfide linkages: see **Results**), each of the filtered structures were energy minimized in CHARMM **(Brooks et al., 1983)** by 500 steps of Steepest Descents followed by 20000 steps of Adopted Basis Newton-Raphson method with a gradient tolerance of 0.001 and a distance dependent dielectric multiplied by 4.0 using the CHARMM-22 forcefield **(MacKerell, 1998)**. Backbone-flexibility in a design protocol has been given considerable importance in the literature **(Desjarlais and Handel, 1999)** which was suitably taken care of, during the minimization by applying soft harmonic restraints on main-chain atoms and C^{β} (the constant harmonic force parameter being set to 5.0 for N, C^{α}, C and O atoms and 2.5 for C^{β}). However, the energy minimized structures registered a C^{α} RMSD of as small as 0.022 (± 0.001) Å w.r.t. the native backbone.

2.3. Packing Density

Static Van der Waals envelope volume **(Gerstein and Richards, 2012)** were summed up for each residue in a given core sequence , which was one of the filters applied in the initial stages. For calculating packing density from atomic coordinates a standalone software (http://bioinfo.mbb.yale.edu/hyper/mbg/SurfaceVolumes/code-mbg/bin-alpha/) utilizing the voronoi method was used). Packing density (P_d) of a residue in a folded chain was then computed by dividing the van der Waals envelope volume (V_{vdw}) of the residue by its voronoi volume (V_{vor}).

$$Pd = \frac{V_{vdw}}{V_{vor}}$$

2.4. Sequence Entropy

Sequence heterogeneity was examined by computing shannon entropy of each core sequence by the following standard expression:

$$S_e = -\sum_{i=1}^{N_c} P_i \log_2(P_i)$$

where P_i is the discrete probability of occurrence of the i^{th} residue and N_c is the total number of residues in the core.

2.5. Secondary Structural Propensity score

For each residue in a randomly generated core sequence, its (Chou-Fashman) propensity to reside on the corresponding secondary structural element (helix / sheet) in the native structure was determined and summed up to give the propensity score (SC_{prop}) for the sequence.

2.6. Short Contact

For short-listing of hydrophobic residues at each position in the native core, a short contact was defined when any non-hydrogen atom (side / main chain) of a threaded conformer was found within a distance of 2.5 Å (or less) of any other non-hydrogen backbone atom contributed by the rest of the polypeptide chain.

For the second round of screening, short contacts (between two non-hydrogen side-chain atoms contributed by two different residues) were defined based on the particular atomic pair with their van der Waals radii being sampled from the AMBER94 all atom molecular mechanics forcefield **(Cornell et al., 1995)**. The van der Waals radii of the two atoms (in contact) were summed up and a constant value of 1.3 Å was subtracted to set the cutoff for short contact for each of pair. The choice of 1.3 Å was optimized such that the C-C short contact distance becomes 2.5 Å.

2.7. Complementarity Scores

Surface and electrostatic complementarities were calculated for each buried or partially buried residues in a designed structure as described in Chapter 3. Complementarity Scores (***CS_{gl}, CS_{cp}, CS_{l}***) were computed as detailed in Chapter 4 and Chapter 5 for the whole chain as well as for the core.

2.8. Network Similarity and Distance

Initially the surface contact network was identified from the native core. The criteria to assign a link between two interacting side-chains have been described in Chapter 2. The network contained 12 links between 18 core residues **(Table 1, Figure 1)** with an embedded triplet clique (59-PHE, 71-TYR, 134-PHE) with linear branching (8 links in all), two disjoint standalone links (each connecting 2 nodes) and a 3 residue open linear chain (see Chapter 2).

Table 1. Surface Contact networks of the Cyclophilin core. Node1 is connected to Node2 by a non-covalent link.

Node1	Node2
29-VAL	47-LEU
29-VAL	63-CYS
31-PHE	45-ILE
43-ILE	160-VAL
47-LEU	59-PHE
59-PHE	71-TYR
59-PHE	134-PHE
59-PHE	151-PHE
71-TYR	134-PHE
76-PHE	85-ILE
85-ILE	164-ILE
120-LEU	151-PHE

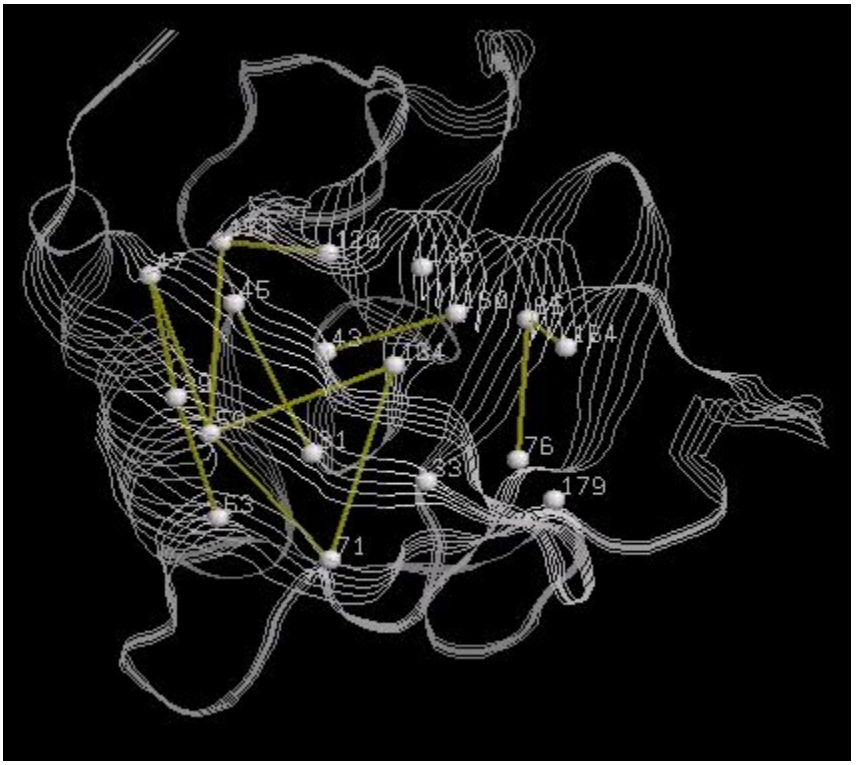


Figure 1. Links present in the surface contact network of the 2HAQ core. Figure constructed in RASMOL **(Sayle et al., 1995)**.

Based on this (template) network, corresponding links were identified from other designed structures and the similarity between the two networks were quantified by the following measure where $N_{designed}$ is the number of equivalent links (w.r.t. the template network) present in a given structure and N_{native} is the total number of links in the template.

$$s_{net} = \frac{N_{designed}}{N_{native}}$$

Another abstract distance measure (d_{net}) was formulated which quantifies the dissimilarity between two adjacency matrices (A and A`) corresponding to the template network and the corresponding network from the threaded structure. Thus, A(i,j) and

A`(i,j) represents the adjacency between the same $i^{th}$ and $j^{th}$ nodes in graphs A and A` since their residue positions are identical in both adjacency matrices. Distance between two such (undirected) graphs of identical size could be determined by counting the number of links that are present in one and absent in the other and then dividing by the number of links present in either of the two graphs.

$$d_{net}(A, A`) = \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} | A(i,j) - A`(i,j) |}{nL}$$

where A(i,j) and A`(i,j) are the matrix elements of adjacency matrices A and A` based on 2HAQ and the threaded structure respectively and nL is the number of elements in the set E U E` where E and E` are the sets of links corresponding to graphs A and A`. It can be shown that d_{net} is formally a metric in a vector space.

3. Results and Discussion

3.1. Identification of the cyclophilin-core

2HAQ (CyclophilinA of *Lieshmania donovani)* was chosen as the target fold and its hydrophobic core was characterized prior to the design process. 2HAQ is a single domain globular protein constituted of a beta-barrel and two helices on either side of the barrel. The molecule has a single cysteine and therefore no disulphide bonds which makes it a good candidate to study folding and design. The cluster of residues composing the only hydrophobic core of cyclophilin within the barrel, also interconnect the secondary structural elements constituting the molecule. The details with regard to the construction of surface contact networks have been described in detail in Chapter 2. Since these residues constituting the core have low solvent accessibility their side chains have high surface and electrostatic complementarity w.r.t. to their immediate atomic environments and rest of the protein respectively. The contact network of 2HAQ was

thoroughly examined visually in RasMol **(Sayle et al., 1995)**. 18 completely buried hydrophobic residues composed the core within the barrel located on helices and sheets, with the exception of 71-Tyr which was found to reside on a loop.

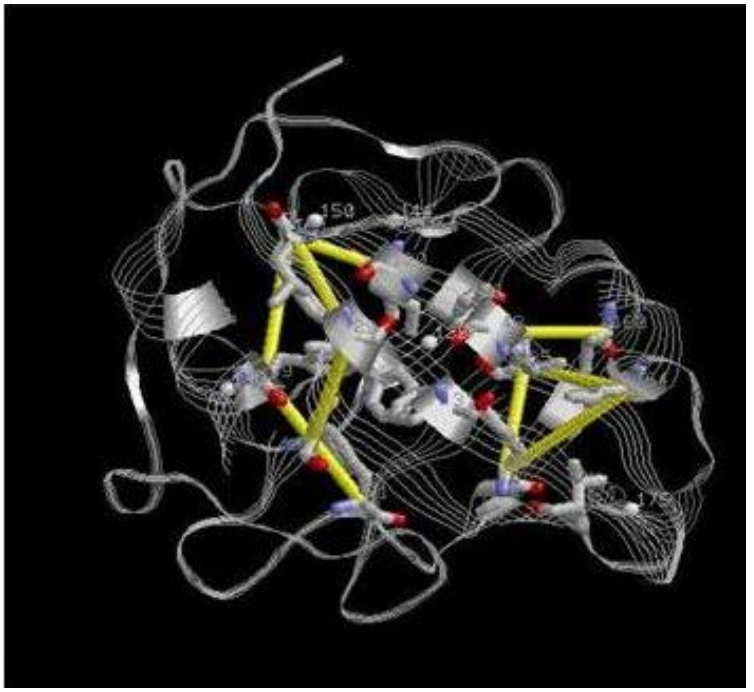


Figure 2. The Cyclophilin core. A set of 18 completely buried in-faced hydrophobic residues spatially connecting all the major secondary structural elements (helices and sheets). Figure constructed in RASMOL **(Sayle et al., 1995)**.

Table 2. Residues sustaining the Cyclophilin core, their burial, secondary structural location and side-chain shape and electrostatic complementarity.

Residue	Burial	Secondary Structural Location	S_m^{sc}	E_m^{sc}
29-VAL	0.00	Sheet	0.608	0.605
31-PHE	0.00	Sheet	0.623	0.306
33-VAL	0.01	Sheet	0.579	0.589
43-ILE	0.00	Sheet	0.501	0.631
45-ILE	0.00	Sheet	0.542	0.681
47-LEU	0.00	Sheet	0.562	0.436
59-PHE	0.00	Helix	0.577	0.415
63-CYS	0.01	Helix	0.630	0.412
71-TYR	0.00	Loop	0.619	0.210
76-PHE	0.00	Sheet	0.554	0.447
85-ILE	0.00	Sheet	0.520	0.515
120-LEU	0.00	Sheet	0.562	0.384
134-PHE	0.00	Sheet	0.572	0.594
136-ILE	0.00	Sheet	0.565	0.699
151-PHE	0.00	Sheet	0.601	0.369
160-VAL	0.03	Helix	0.450	0.550
164-ILE	0.00	Helix	0.555	0.560
179-VAL	0.00	Sheet	0.484	0.613

3.2. Short-listing of hydrophobic residues at each core position based on complementarity and clashes w.r.t. the main chain atoms alone : 1st filter

To start the design process, all side-chains were initially removed from the native backbone of the protein and each position from the selected list of 18 residues were sequentially mutated to conformers of hydrophobic residues (ALA, VAL, LEU, ILE, PHE, TYR, TRP, CYS, MET: 71 conformers in total) in turn, sequentially selected from Dunbrack's Rotamer library. The purpose of this procedure was to shortlist residues at

each position and more importantly to eliminate those cases where side-chain conformers are either involved in short contacts (two non-hydrogen atoms within 2.5 Å or less) with main chain atoms or fail to meet the threshold values in surface or electrostatic complementarity w.r.t to their immediate environment constituted of main chain atoms alone. Prior to this calculation surface and electrostatic complementary values of buried side-chains (w.r.t native main chain atoms) had been estimated from polypeptide chains in the database **DB2** (S_m^{mc}, E_m^{mc}: see Chapter 3) and based on their statistics, the threshold values for allowed conformers were set to 0.25 and 0.30 for surface and electrostatic complementarity respectively. If at least one conformer of a particular hydrophobic residue passed the cutoff, then the amino acid was considered to be a potential candidate at that position. This method led to the elimination of specific hydrophobic amino acids at only two residue-positions (native: 63-CYS, 160-VAL; out of 18) where bulky residues were clearly involved in extensive steric clashes with the main chain atoms. However, the total number of possible combinations (reduced from (9^{18}=) 1.5009e+17 to 1.3066e+14) was still astronomically high.

Table 3. Short-listed hydrophobic residues at each position in the native backbone of Cyclophilin.

Residue Position (Native)	Short-listed hydrophobic residues
29-VAL	MET, ALA, LEU, VAL, ILE
31-PHE	CYS, TRP, MET, PHE, ALA, VAL, ILE
33-VAL	LEU, VAL, ILE, TYR, TRP, CYS, MET, ALA, PHE
43-ILE	LEU, VAL, ILE, ALA
45-ILE	LEU, VAL, ILE, TYR, CYS, TRP, MET, PHE, ALA
47-LEU	LEU, VAL, TYR, CYS, PHE, ALA
59-PHE	LEU, VAL, ILE, TYR, TRP, MET, PHE
63-CYS	CYS, ALA
71-TYR	LEU, VAL, ILE, TYR, CYS, TRP, MET, PHE, ALA
76-PHE	LEU, VAL, ILE, TYR, CYS, TRP, MET, PHE
85-ILE	LEU, VAL, ILE, CYS, MET, ALA
120-LEU	LEU, VAL, ILE, TYR, CYS, TRP, ALA
134-PHE	LEU, VAL, ILE, TRP, MET, PHE, ALA
136-ILE	LEU, VAL, ILE, TYR, CYS, TRP, MET, PHE, ALA
151-PHE	LEU, VAL, ILE, CYS, TRP, MET, PHE, ALA
160-VAL	VAL, CYS
164-ILE	LEU, VAL, ILE, CYS, MET
179-VAL	LEU, VAL, ILE, TYR, CYS, TRP, MET, PHE, ALA

3.3. Random Sampling

Random sampling was then carried out from the pool of short listed residues as probable candidates at each residue position. To restrict the number of trials, random sampling was restricted to 10^5 sequences, though it was ensured that the each short listed residue was sampled nearly equally at each position in the polypeptide chain. Sequence identities w.r.t. the native sequence at the core was calculated for all the designed sequences and sequences with 15% sequence identity (w.r.t.) native were found to be the

most frequent. The probability of retaining the native residue at each position is 0.11 if all the 9 hydrophobic residues are considered. However for the short listed set (shown above) the probability of selecting the native (core) residue falls within 0.11 to 0.25 for 16 out of 18 core-positions. For only two positions (63-CYS and 160-VAL) the probability is 0.50. Thus the observation that most of the sequences exhibited a sequence identity between 10 – 20 % (w.r.t. native), could possibly be expected.

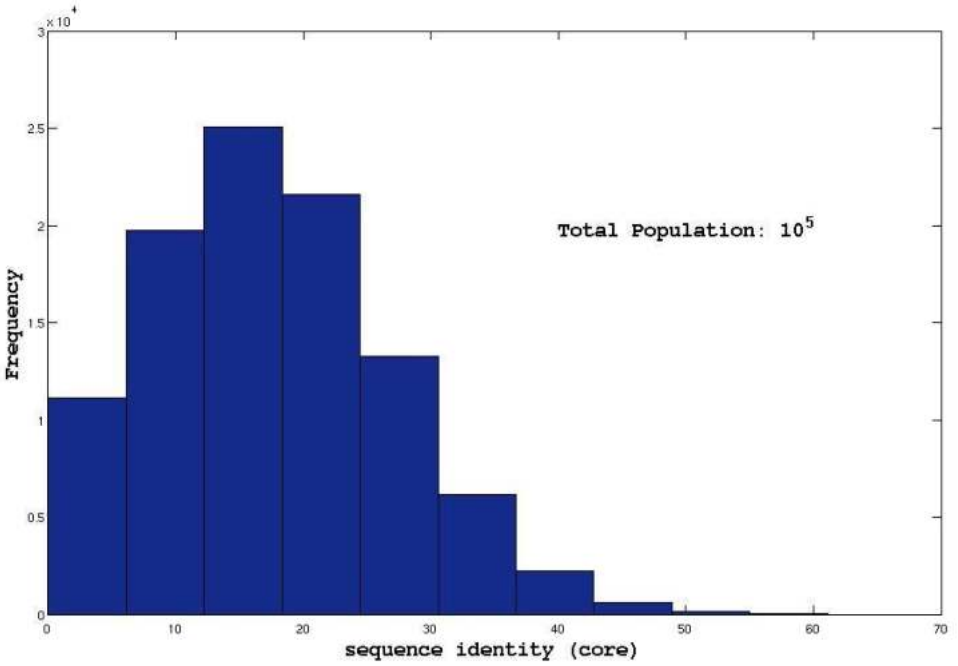


Figure 3. Sequence identities of the initial randomly sampled ensemble of sequences w.r.t. the native core.

3.4. Screening based on sequence level static informations: 2nd filter

From the shortlisted residues at each hydrophobic position, 10^5, 18 residue sequences were then constructed with random selection of allowed residues at each position, which were then subject to several filters based on: 1) the sum of the van der Waal's envelope volumes of the 18 residues 2) sequence entropy (to test for sequence heterogeneity) and 3) Chou-Fashman secondary structural propensity (see **Materials and Methods**). 2HAQ was structurally aligned with 17 homologues (by Dali server) and structurally equivalent core residues identified. Estimation of the van der Waals envelope volume of the core residues for these sequences gave an average of 1947.6 (± 28.59) $Å^3$. Similarly, sequence entropy and Chou-Fashman propensities of these core sequences were found to be 2.44 (±0.16) and 23.06 (±0.66) respectively.

Table 4. Van der Waals envelope volume, sequence entropy and Secondary structural propensity of core sequences from Cyclophilin homologues.

Cyclophilin Homologue	$V_{vdw}(Å^3)$	S_e	SC_{prop}
1A58	1922.98	2.525	23.36
1DYW	1935.02	2.405	22.11
1IHG	1930.01	2.224	22.91
1QOI	1952.89	2.525	23.04
1XO7	1908.24	2.169	24.79
1ZKC	1953.35	2.705	22.67
2CFE	2001.33	2.288	22.87
2CMT	1941.60	2.663	22.90
2FU0	1892.04	2.523	22.92
2GW2	1948.28	2.377	22.96
2HAQ	1955.19	2.324	23.59
2HE9	1986.46	2.330	22.76
2HQJ	1977.63	2.505	22.82
2PLU	1987.60	2.505	22.94
2R99	1923.27	2.510	21.90
2X25	1954.63	2.642	23.32
2ICH	1928.32	2.224	24.17
3K2C	1958.74	2.530	23.04

Based on these values, those sequences were selected which were found to have their summed van der Waals envelope volumes between 1900 to 2000 $Å^3$, sequence entropy between 2.0 to 3.0 and secondary structural propensity greater than 20. Since the same amino acid could have nearly equal Chou-Fasman propensities for different secondary structural elements (helices and sheets), the cut off for this parameter was

relaxed. On the application of these filters, 33116 sequences were finally obtained. These sequences were then threaded onto the native backbone, with their side-chain torsions being optimized by SCWRL4.0. Hydrogens were then removed from all structures and rebuilt by REDUCE. Most of these filtered sequences exhibited a sequence identity of about 20 % (w.r.t. the native core)

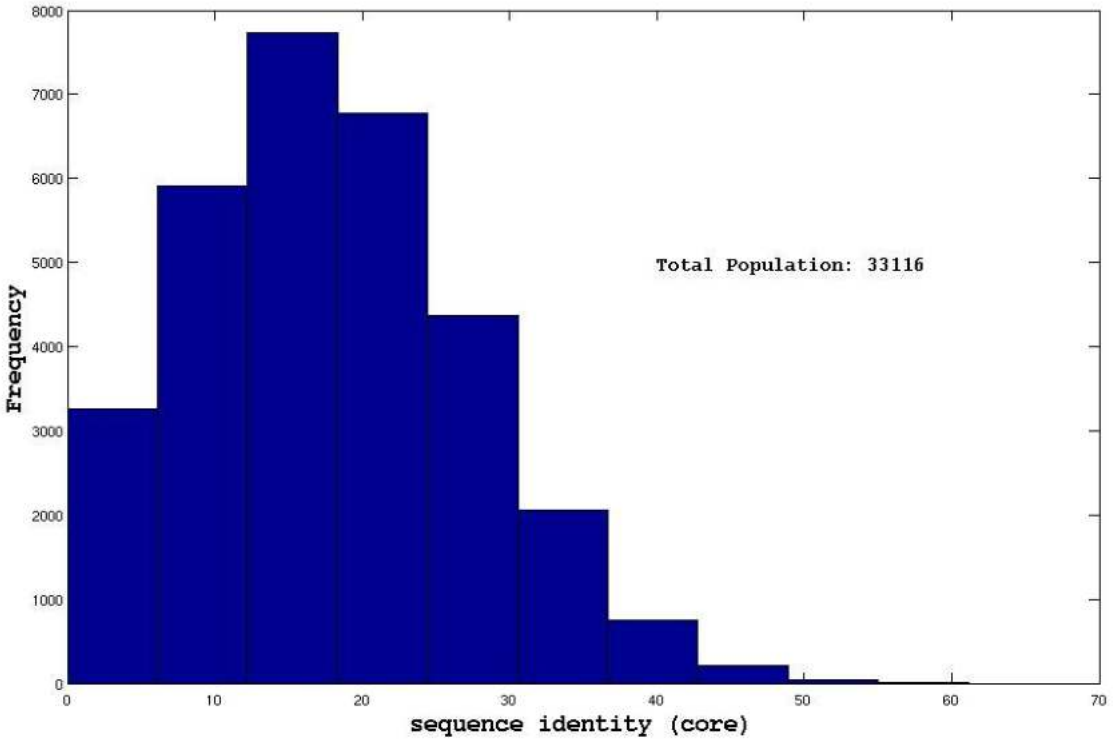


Figure 4. Sequence identities of the ensemble of sequences w.r.t. the native core subsequent to the first screening.

3.5. Cutoff on Packing density and short-contacts : 3rd filter

For this set of structures, packing density for individual core residues were estimated along with possibility of disulfide bridges (based on a CYS-SG – CYS-SG distance cutoff : 2.25 Å) and number of atomic short contacts (see **Materials and Methods**). Initially, average packing densities of residues (along with their standard deviations) distributed in different burial bins (bin1: $0.0 \leq \boldsymbol{Bur} \leq 0.05$, bin2: $0.05 \leq \boldsymbol{Bur} \leq 0.15$, bin3: $0.15 \leq \boldsymbol{Bur} \leq 0.30$; see definition of ***Bur*** in Chapter 3) were computed from the database **DB2**. For the 1st burial bin, the mean packing densities (μ) of residues were found within the range: 0.67 to 0.74 ($\pm$ 0.05) irrespective of the residue identity. Similarly, for the 2nd and 3rd bins, the values were found to be within the range: 0.59 to 0.68 ($\pm$ 0.011) and 0.61 to 0.73 ($\pm$ 0.16) respectively. Residues in the designed structures were distributed in two groups, 1) core residues with burial $\leq$ 0.30 and 2) non-core residues with burial $\leq$ 0.30. The structures (from the pool of 33116) were then filtered out which contained

1) three or more short contacts in the core,
2) eight or more short contacts in the overall structure,
3) no possible disulfide linkages,
4) 80% of the core residues having a packing density within the range of $\mu \pm 2\sigma$ (μ, σ obtained from **DB2**) and
5) 100% of the non-core buried or partially buried residues having a packing density within the range of $\mu \pm 3\sigma$ (**DB2**).

This led to a reduction in the number of structures from 33116 to 7158.

3.6. Complementarity Cutoff : 4th filter

These structures were then energy minimized by CHARMM and their surface and electrostatic complementarities calculated for all buried and partially buried residues. For all the structures in this set, complementarity scores ($\boldsymbol{CS_{gl}}$, $\boldsymbol{CS_{cp}}$ see Chapter 4) and ($\boldsymbol{CS_l}$

see Chapter 5) were computed for the set of all core residues and the full chain separately. For the native structure (2HAQ), CS_{gl}, CS_{cp}, CS_{l} were found to be 3.59, 0.0154, 2.54 for the full chain and 6.01, 0.024, 3.16 for the core. In addition, there were 16 (out of 18) core residues in the probable region of the plot (completely buried: CP1) and none in the improbable region.

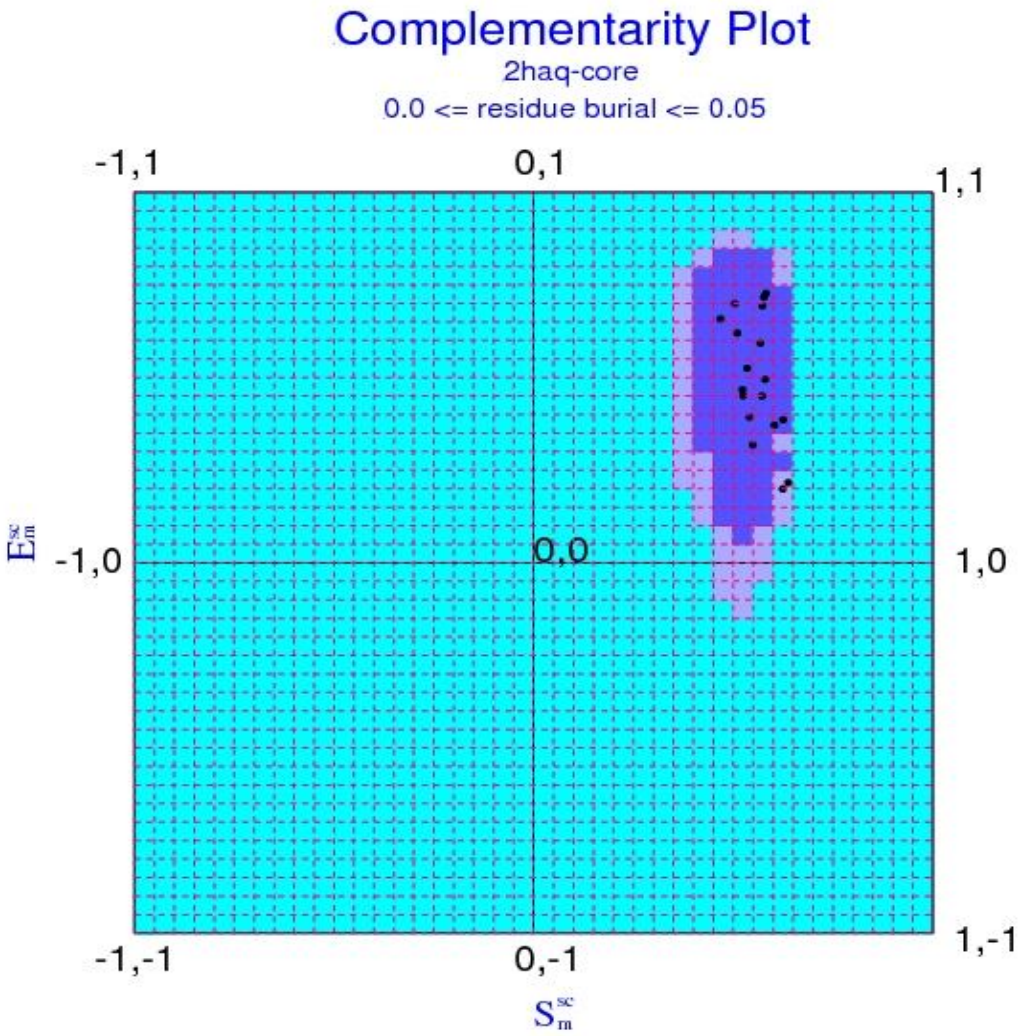


Figure 5. Distribution of points from the native 2HAQ core in the Complementarity Plot (CP1).

For the designed structures,the solvent accessibility of the core residues might be different when compared with the native core. Only those designed structures were hence

forth considered, whose designed core consisted of residues which were all buried or partially buried and at least 16 (out of 18) residues were found to be in the probable regions of the plots (CP1, CP2, CP3). This filter reduced the number of structures to 346. Compared to the native core the sequence identities of the designed core ranged from 0 – 55.6 % with a maximum observed between 22 – 28 %.

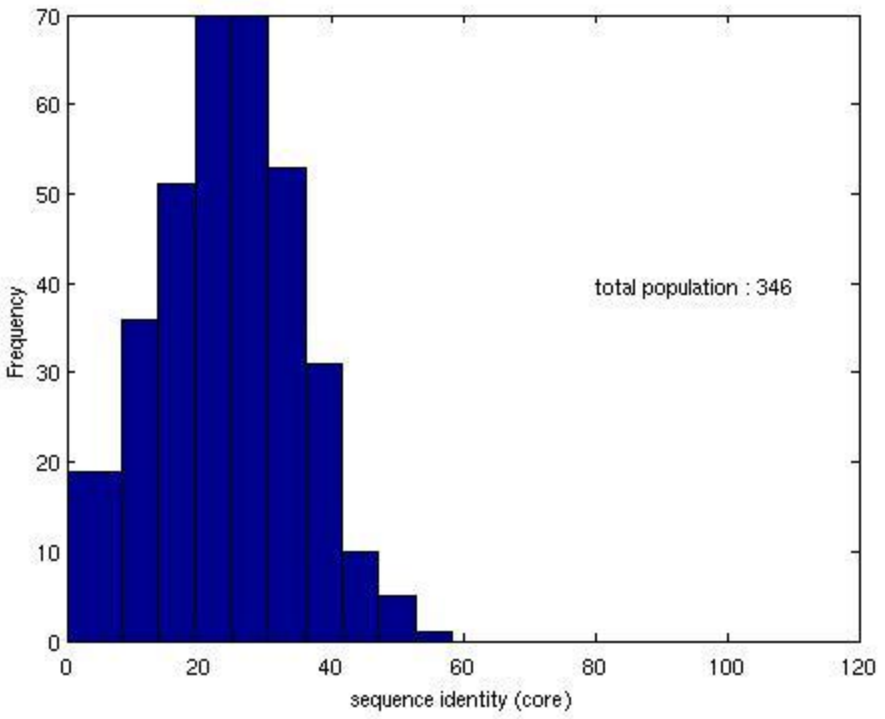


Figure 6. Sequence identities of the ensemble of sequences w.r.t. the native core subsequent to the second screening.

1000 structures were generated by threading randomly selected hydrophobic residues (with their side-chain torsions optimized by SCWRL4.0 **(Krivov et al., 2009)**) and their complementarity scores (***CS**$_{gl}$*, ***CS**$_{cp}$*, ***CS**$_{l}$*) were then calculated subsequent to energy minimization (by CHARMM **(Brooks et al., 1983)**). Interestingly, minimization seemed to raise the scores substantially. Scores were then averaged over the ensemble and found to be: ***CS**$_{gl}$*: 2.53 (± 0.20), ***CS**$_{cp}$*: 0.011 (± 0.0008), ***CS**$_{l}$*: 1.71 (± 0.32) (full

chain) and $\boldsymbol{CS_{gl}}$: 3.01 (± 0.79), $\boldsymbol{CS_{cp}}$: 0.0112 (±0.003), $\boldsymbol{CS_{l}}$: 1.70 (±1.30) (core). Since $\boldsymbol{CS_{l}}$ (core) had the highest standard deviation among all the complementarity scores, this was chosen as the initial parameter for further screening and from the set of 346 sequences (previously screened), only 28 (Table) were found to exceed the mean (1.70) in $\boldsymbol{CS_{l}}$ (core) obtained from the random structures.

Network parameters s_{net}, d_{net} (see **Materials and Methods**) along with the accessibility score (***rGb***) and link density ($\mathbf{L_d}$ see **Chapter 2**) were then computed for both the full chain and core for these 28 sequences. Since the native structure contained a triplet clique (59-PHE, 71-TYR, 134-PHE) in the core, this geometrically constrained packing motif was also exhaustively searched for in the cores of the short listed sequences. Both ***rGb*** (0.0544 ± 0.0029) and $\mathbf{L_d}$ (core: 0.1 ±0.018; full-chain: 0.049 ±0.007) were found to be uniform for all sequences and very close to their corresponding native values (***rGb*** **:** 0.059, $\mathbf{L_d}$ (core): 0.05, $\mathbf{L_d}$ (full-chain): 0.08).

Table 5. The selected list of 28 core sequences and their corresponding scores. N_{prob} is the number of residues found in the probable region of the CPs out of 16 core residues.

Ref	Sequence	Seq Id (core %)	CS_l (core)	N_{prob} (core)	CS_{gl} (core)	CS_{cp} (core)	Presence of a triplet clique
s1	AAMVFLMAYFLLWIAVLL	33.3	2.06	16	3.26	0.013	-
s2	AIFAVLFALIMYFWIVLA	22.2	2.04	16	3.65	0.012	-
s3	LFIACLMAFWVYFAAVMM	22.2	2.03	16	3.29	0.014	-
s4	LFFLCLYAIFMIMMCVCC	22.2	2.01	16	4.00	0.016	-
s5	VWVLILVAIVMVIWAVIC	33.3	1.99	16	4.04	0.014	-
s6	IVVLLLLAWICIVWCVCM	16.7	1.99	17	3.87	0.014	-
s7	MMFAIVYACLLCIYMVMV	16.7	1.97	17	4.12	0.017	-
s8	VFIAWFVAIFCYFAAVLV	33.3	1.96	16	4.08	0.015	-
s9	MFYAIVIAFLCAFMFVLI	27.8	1.96	16	3.75	0.014	-
s10	MWMLMLFAYMVALIAVCV	33.3	1.96	16	4.22	0.014	-
s11	IAVIMYIAVFLCFMFVML	33.3	1.95	16	5.08	0.020	-
s12	LFVIMFMAIVCAFMIVLV	33.3	1.94	16	4.31	0.015	-
s13	AAVVFCFCFFLVFMFVCC	38.9	1.94	17	4.34	0.014	-
s14	ACAIVYFAFYVWFMCVMI	22.2	1.94	16	5	0.017	-
s15	MIIIMLVAYIVCFIWVVC	33.3	1.93	16	4.70	0.015	-
s16	MMFAVVFAALCAIFWVIV	22.2	1.93	16	4.22	0.015	-
s17	VFVAIVVAAYMWFVFVIL	44.4	1.93	16	4.59	0.018	-
s18	ACVLCLYAYYMLFILVMC	38.9	1.92	16	4.07	0.016	+
s19	MCAVYYVAFYVVWMMVVA	5.6	1.92	17	4.30	0.017	-
s20	VVVIFCFALIVVLMLVVI	27.8	1.92	16	4.60	0.016	-
s21	VCYVICWACIVWMMCVVV	22.2	1.91	16	4.18	0.014	-
s22	VAFVFVWALVILIWVVLV	27.8	1.90	16	4.74	0.018	-
s23	LAFVMLYALWCIALFVLI	16.7	1.90	18	3.94	0.015	-
s24	LMLACYLAFFIWFIAVVV	33.3	1.86	17	5.02	0.019	-
s25	VMVVYVFCVWLLFVVVMA	38.9	1.85	17	5.18	0.018	-
s26	AILIFLWCYFCALVFVIV	50.0	1.85	17	4.58	0.018	-
s27	VMFAILFALLILFLLVLV	50.0	1.85	16	5.27	0.021	+
s28	VVFAFFIAYMMIIYAVLV	22.2	1.85	16	5.43	0.018	+

As can be seen from Table 5, all the designed structures gave much higher scores in CS_{gl} and CS_{cp} (core) than the estimate of the same measures from the randomly generated structures. However, CS_{gl} and CS_{cp} were clearly found to be more sensitive indicators of the overall compatibility of these core-sequences with the native fold and these scores for some of the designed structures exceeded the average obtained from random substitutions by a very good margin (CS_{gl}: 3.01 (± 0.79), CS_{cp}: 0.0112 (±0.003)). Thus, from these 28 sequences, 6 sequences with CS_{gl} score (core) greater than equal to 5.00 were finally selected (s28, s27, s25, s11, s24, s14) of which two (s27, s28) contained a triplet clique in the core.

4. Conclusion

The primary objective of the study was to apply the complementarity measures in the redesign of the hydrophobic core in cyclophilin from *L. donovani*. In addition, other measures based on sequence heterogeneity, volume, clashes and packing densities were also implemented to discard improbable sequences at different stages. Use of network parameters clearly shows that at least theoretically there can be multiple packing arrangements which can sustain the native core. Thus, along with the evolutionary conserved native core-packing arrangement, alternative cores could also satisfy the general packing constraints (shape complementarity, packing density, avoidance of steric clashes etc), an observation which definitely requires further experimental validation.

Reference

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983). **CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations.** *J Comp Chem* **4**: 187-217.

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995). **A Second Generation Force Field for**

the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc* **117**: 5179-5197.

Chakraborty K, Thakurela S, Prajapati RS, Indu S, Ali PS, Ramakrishnan C, Varadarajan R (2005). **Protein stabilization by introduction of cross-strand disulfides.** *Biochemistry* **44**: 14638-14646.

Das M, Kobayashi M, Yamada Y, Sreeramulu S, Ramakrishnan C, Wakatsuki S, Kato R, Varadarajan R (2007). **Design of Disulfide-linked Thioredoxin Dimers and Multimers Through Analysis of Crystal Contacts.** *J Mol Biol.* **372**: 1278-1292.

Desjarlais JR, Handel TM (1999). **Side-chain and Backbone Flexibility in Protein Core Design.** *J. Mol. Biol.* **289**: 305-318.

Fung HK, Floudas CA, Taylor MS, Zhang L, Morikis D (2008). **Toward Full-Sequence De Novo Protein Design with Flexible Templates for Human Beta-Defensin-2.** *Biophys J.* **94**: 584–599

Gerstein M, Richards FM (2012). **Protein Geometry: Volumes, Areas, and Distances.** The International Tables for Crystallography F, Chapter 22. Version: fr823.

Indu S, Kumar ST, Thakurela S, Gupta M, Bhaskara RM, Ramakrishnan C, Varadarajan R (2010). **Disulfide conformation and design at helix N-termini.** *Proteins* **78**: 1228-1242.

Johansson JS, Gibney BR, Rabanal F, Reddy KS, Dutton PL (1998). **A Designed Cavity in the Hydrophobic Core of a Four-R-Helix Bundle Improves Volatile Anesthetic Binding Affinity.** Biochemistry, **37**: 1421-1429.

Kashiwada A, Hiroaki H, Kohda D, Nango M, Tanaka T (2000). **Design of a Heterotrimeric R-Helical Bundle by Hydrophobic Core Engineering.** *J. Am. Chem. Soc.* **122**: 212-215.

Krivov GG, Shapovalov MV, Dunbrack RL (2009). **Improved prediction of protein side-chain conformations with SCWRL4.** *Proteins.* **77**: 778-795.

Lazar GA, Desjarlais JR, Handel TM (1997). **De novo design of the hydrophobic core of ubiquitin.** *Protein Science* **6**: 1167-1178.

Lu Y, Valentinet JS (1997). **Engineering metal-binding sites in proteins.** *Current Opinion in Structural Biology* **7**: 495-500

MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK,

Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998). **All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins.** *J Phys Chem* B **102**: 3586-3616.

Munson M, Balsubramanian S, Fleming KG, Nagi AD, O'brien R, Sturetevant JM, Regan L (1996). **What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties.** *Protein Science* **5**: 1584-1593.

Tsai J, Gerstein M, Levitt M (1997). **Simulating the minimum core for hydrophobic collapse in globular proteins.** *Protein Science* **6**: 2606-2616.

Sayle RA, Milner-White EJ (1995). **RASMOL: biomolecular graphics for all.** *Trends Biochem Sci* **20**: 374-376.

Shah PS, Hom GK, Ross SA, Lassila JK, Crowhurst KA, Mayo SL (2007). **Full-sequence Computational Design and Solution Structure of a Thermostable Protein Variant.** *J. Mol. Biol.* **372**: 1–6.

Street AG, Mayo SL (1999). **Computational protein design.** *Structute* **7**: R105-R109.

Appendix I

Persistence map of dynamic contact networks using shape complementarity : its evolutionary relationship

A 50 ns molecular dynamic simulation was carried out at room temperature (310 K) on cyclophilin from *Leishmania donovani* (2HAQ). The problem was approached from a network perspective of the protein interior and both static and dynamic features of such networks were analyzed in detail. 2HAQ was structurally aligned with 17 homologues and surface contact networks (see Chapter 2) were constructed. Following is a detailed description of the analysis carried out on the static structures.

The 17 structures belonging to the cyclophilin-like fold were chosen from the SCOP database **(Murzin et al., 1995)** which had greater than 40% sequence identity upon structural alignment with 2HAQ (PDB ID_Chain (RMSD (Å), sequence identity (%)): 1XO7_A (0.5, 74), 3ICH_A (0.8, 65), 2PLU_A (1.3, 63), 2X25_B (1.2, 61), 2CFE_A (1.2, 60), 1QOI_A (0.8, 57), 1A58_A (1.2, 57), 1IHG_A (1.2, 57), 2R99_A (1.3, 57), 1DYW_A (1.4, 57), 2HQJ_A (1.4, 57), 2CMT_A (1.2, 56), 3K2C_B (1.4, 54), 2GW2_A (0.8, 53), 2HE9_A (0.8, 53), 2FU0_A (1.3, 47), 1ZKC_A (1.2, 42)). Surface contact networks (at $S_m \geq 0.4$, $Ov \geq 0.08$ for both A→B and B→A in a A↔B link; see Chapter 2) were generated for all the 17 native structures along with 2HAQ. Unlike networks defined while describing packing motifs (see Chapter 2), these networks could contain unconnected disjoint components and even isolated binary links. Here the primary emphasis was to represent a fold as a unique subset of relevant links, highly conserved amongst members of that fold. Pairwise structural alignment (using Dali Server **(Holm and Rosenstrom, 2010)**) with 2HAQ (considered to be the template)

provided the mapping between the nodes of 2HAQ and each of the 17 homologous proteins. In case of insertions-deletions or non-alignment, the node was considered to be absent in the related protein. Every link in the contact network of 2HAQ was searched systematically in the 17 homologues and counted for the number of times the corresponding (mapped) nodes were found to be present and connected. Only those links from 2HAQ were retained which were present in at least 75% of the other 17 homologues. This led to a subgraph of 22 links which could be considered as the evolutionarily conserved network (CycnetEC) representative of the Cyclophilin-like fold **(Figure 1)**.

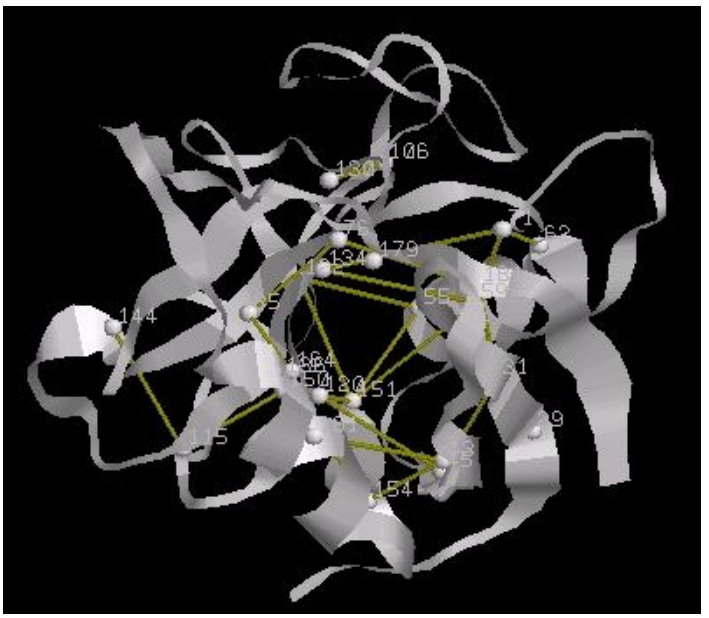


Figure 1. The evolutionarily conserved surface contact network constituting of 22 links in the Cyclophilin-like fold. Figure constructed in RASMOL **(Sayle et al., 1995)**.

2HAQ was selected as a representative member of the Cyclophilin-like fold and based on the initial crystal structure, a molecular dynamics (MD) simulation was carried out at 310 K (37°C). Initially the protein was solvated in a cuboidal box of dimensions 78.673 × 68.897 × 78.317 $Å^3$ and the overall charge of the system neutralized by the addition of a single Na^+ ion through the xleap module of AMBER **(Dejoux et al., 2001)**.

11088 waters were then added following the TIP3P model **(Jorgensen et al., 1983)**. The structure was then energy minimized initially for 200 steps of steepest descent followed by 19800 steps of ABNR incorporated in the SANDER module (AMBER), utilizing the force field and molecular topologies implemented in the AMBER 2002 force field. The energy-minimized structure was heated up to 310 K in NAMD **(James et al., 2005)**. The MD simulation run was then carried out for 50 ns in NAMD involving 50000 steps with a snapshot being collected at each picosecond interval. A NPT ensemble with the Langevin piston temperature set to 310 K was incorporated with the pressure being fixed at 1.0325 bar. SHAKE was incorporated to keep the bond-lengths constrained with a tolerance level of 0.005 Å. Visual molecular dynamics program (VMD) **(Humphrey et al., 1996)** was used to view the trajectories obtained from NAMD and to obtain the coordinates for the snapshots.

Root Mean Square deviations of the C^{α} atoms with respect to the initial crystal structure was calculated for the entire 50 ns trajectory and the system was found to reach equilibration within approximately 4ns and oscillated thereafter in the range of 1.38 Å (± 0.11). Thus, the initial 4ns data were discarded and the rest of the 46ns trajectory were subjected to further analysis.

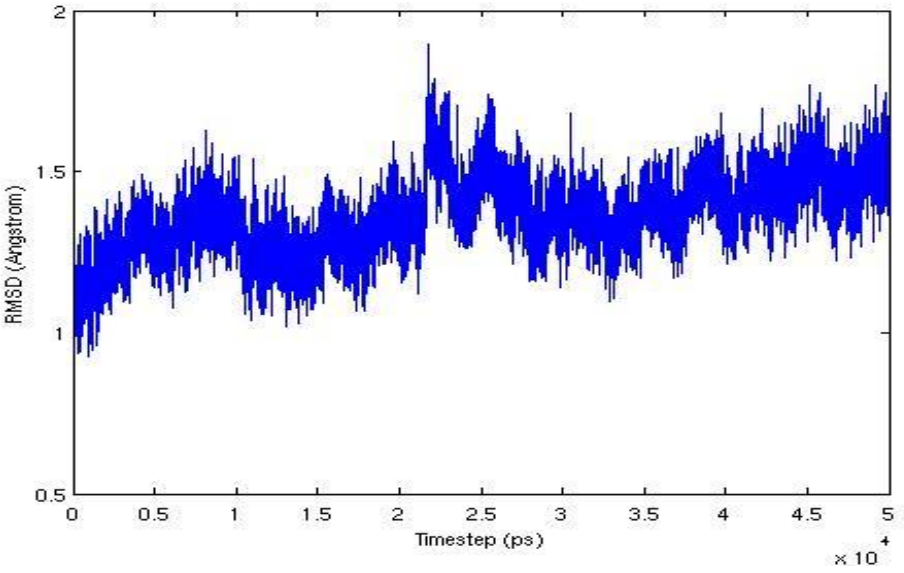


Figure 2. C^{α}-RMSD of the 50 ns molecular dynamic trajectory of 2HAQ.

A snapshot was selected at an interval of 20 ps (one per 20 snapshots) and surface contact networks were generated for each of them (at $S_m \geq 0.4$, $Ov \geq 0.08$ for both A→B and B→A in a A↔B link; see Chapter 2). Thus the analysis were carried out on 2300 snapshots. All possible links in a surface contact network obtained from each of the selected snapshot was then tested for adjacencies in a 166 × 166 symmetric matrix corresponding to 166 residues in 2HAQ. Each possible link was then counted for the number of times it appeared in a snapshot (i.e., satisfied the contact criteria) and divided by the total number of snapshots to obtain the persistence of the link in the dynamic trajectory. Links with high dynamic persistence (≥ 0.75) were accumulated and this subset of 28 links could be considered as the dynamically persistent network (CycnetDP) representative of the cyclophilin-like fold.

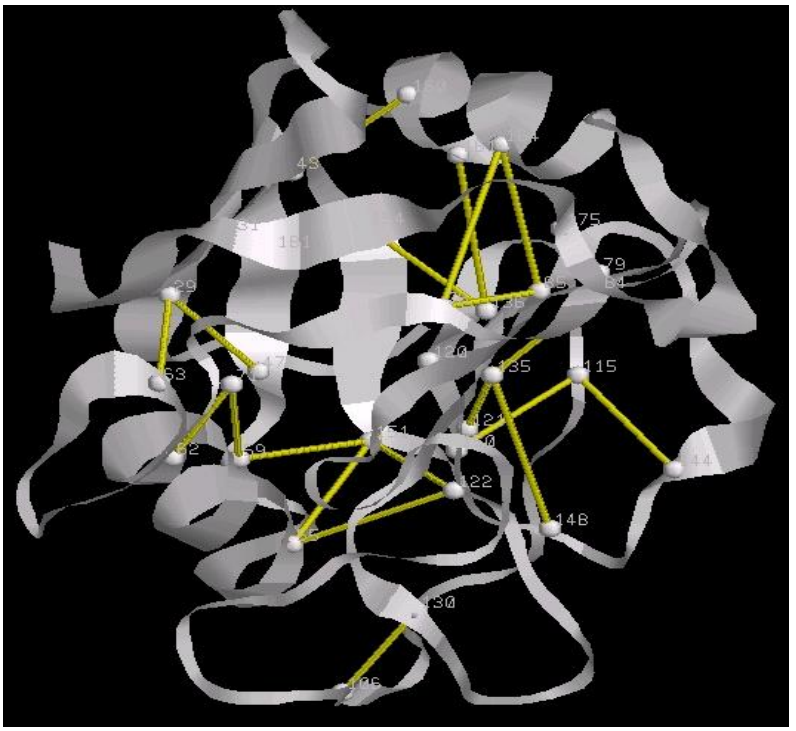


Figure 3. The dynamically persistent surface contact network constituting of 28 links generated from the 50 ns MD simulation of 2HAQ (representative of the Cyclophilin-like fold). Figure constructed in RASMOL **(Sayle et al., 1995)**.

A thorough comparison was then carried out between the evolutionarily conserved network (CycnetEC) and the dynamically persistent network (CycnetDP) representative of the Cyclophilin-like fold in terms of network parameters (s_{net}, d_{net}) described in Chapter 6. Constituent links of both these networks were found to connect all the crucial secondary structural elements (Helices and Sheets) and a total of 15 links were found to be common between them (which exceeded 0.75 in both evolutionary conservation and dynamic persistence). This gave an s_{net} and d_{net} values of 0.42 and 0.57 respectively between the two networks. All these 15 common links had both the dynamic persistence and evolutionary conservation greater than 0.80 (**Table 1**). This common network contained one triplet clique and several discrete open linear chains and traversed the entire three dimensional structure by connecting between the beta sheets and anchoring helices (both sheet-sheet, sheet-helix contacts). Thus, this could be considered as the evolutionarily conserved and dynamically stable optimal subgraph to hold the native cyclophilin-like fold.

Table1. The evolutionarily conserved and dynamically persistent links in the Cyclophilin-like fold.

Node1	Node2	Dynamic Persistence	Evolutionary Conservation
62-LEU	71-TYR	0.995	0.900
59-PHE	151-PHE	0.995	0.950
85-ILE	164-ILE	0.993	0.900
106-PHE	130-ASN	0.990	0.900
120-LEU	151-PHE	0.988	0.900
115-HIS	150-VAL	0.978	1.000
59-PHE	71-TYR	0.975	0.850
55-PHE	151-PHE	0.966	0.950
122-MET	151-PHE	0.950	0.950
55-THR	122-MET	0.939	0.950
31-PHE	181-ILE	0.930	0.950
136-ILE	161-VAL	0.928	0.850
76-PHE	85-ILE	0.856	0.850
136-ILE	154-VAL	0.842	0.810
115-HIS	144-LEU	0.807	0.900

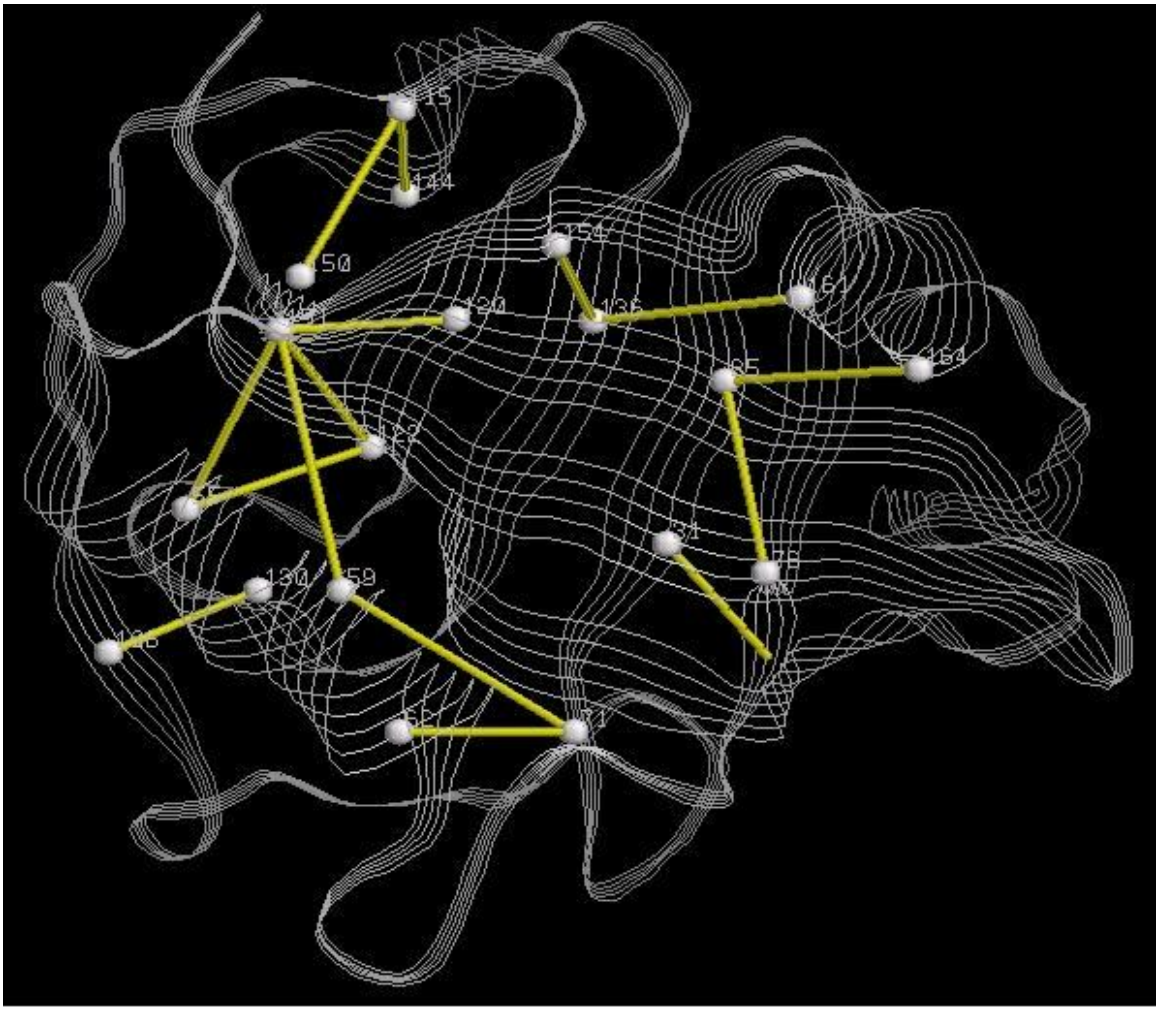


Figure 4. The evolutionarily conserved and dynamically persistent optimal subgraph holding the Cyclophilin-like fold. Figure constructed in RASMOL.

References

Dejoux A, P. Cieplak, N. Hannick, G. Moyna, & F.-Y. Dupradeau, AmberFFC (2001). **A Flexible Program to Convert AMBER and GLYCAM Force Fields for use with Commercial Molecular Modeling Packages.** *J. Mol. Model.*, **7**: 422-432.

Holm L, Rosenström P (2010). **Dali server: conservation mapping in 3D.** *Nucl Acids Res.* **38**: W545-549.

Humphrey W, Dalke A, Schulten K (1996). **VMD - Visual Molecular Dynamics.** *J. Molec. Graphics* **14.1**: 33-38.

James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kale, and Klaus Schulten (2005). **Scalable molecular dynamics with NAMD.** *Journal of Computational Chemistry*, **26**: 1781-1802.

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983). **Comparison of simple potential functions for simulating liquid water.** *J. Chem. Phys,* **79**: 926-935.

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995). **SCOP: a structural classification of protein database for the investigation of sequences and structures.** *J Mol Biol.* **247**:536-540.

Sayle RA, Milner-White EJ (1995). **RASMOL: biomolecular graphics for all.** *Trends Biochem Sci.* **20**: 374-376.

Appendix II

Geometry and electrostatics of Salt bridges within proteins

As has been demonstrated by McCoy et al., **(McCoy et al, 1997)** salt bridges are important at the interface in determining the magnitude of electrostatic complementarity, however, since all the other charges from the two interacting molecules contribute to the potentials, complementarity can still be significant even when the salt bridges are computationally neutralized (Chapter 1). There does not appear to be a universal rule regarding the role of salt bridges in stabilizing protein structures. Due to desolvation effects, they are in general considered to be destabilizing **(Honig and Yang, 1995)**, though instances have been observed where networks of ionic bonds contribute favorably to the thermal stabilization of the protein **(Bogan and Thorn, 1998; Torshin et al., 2002; Di Primo et al., 1997; Walker and Causgrove, 2009)**. In order to study the pattern of networks (constituted by ionic bonds) and their associated E_m values, a total of 3076 networks were extracted from the database, **DB2** and classified according to a topological scheme, described in detail in Chapter 2. Briefly, charged residues (represented as nodes) are connected by an edge when there exist an ionic bond (or salt bridge) between which was detected when a positively charged nitrogen atom of lysine (NZ), arginine (NH1, NH2) or positively charged histidine (HIP: ND1 NE2, both protonated) were found to be within 4.0 Å of a negatively charged oxygen atom of glutamate (OE1, OE2) or aspartate (OD1, OD2).

A unique network-topology is numerically represented by *n* concatenated strings of numbers separated by delimiters (where *n* is the number of nodes in the network). Each string begins with the degree of a node (from the highest degree node following a descending order in degrees), followed by the degrees of its linked nodes sorted in descending order.

The distribution of such networks was found to be dominated by isolated ionic bonds (**11-11:** 2445, **Figure 1**) followed by bifurcated salt bridges consisting of three nodes (**211-12-12:** 475). For networks, with number of nodes greater than three, the

overwhelming majority fell into the class of open linear chains (see **Chapter 2**) or their variants. The reason for this topological preference has obviously to do with the fact that no two adjacent nodes can carry like charge. A few examples of four membered closed rings either isolated or 'fused along an edge' (see **Chapter 2**) were also found. For closed rings, the topological constraint due to charge allows only an even number of nodes.

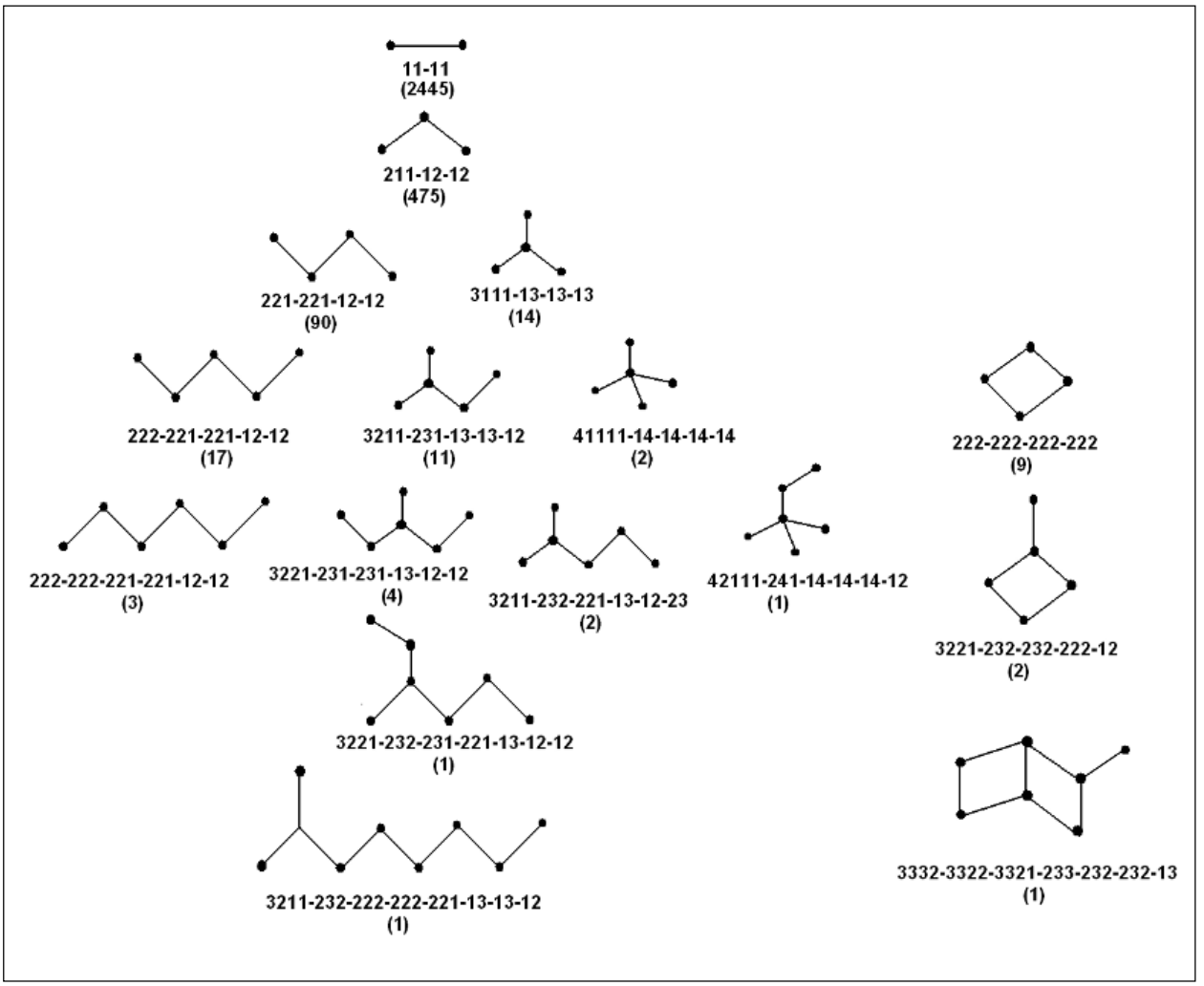


Figure 1. Statistical distribution of networks of ionic bonds. Each network topology is demonstrated by an identifier numerical string**.** The number of such networks found is given in parenthesis below the identifier.

Overall, a mild enhancement in E_m (see Chapter 3) was observed for charged residues, involved in salt bridges **(Figure 2),** with the exception of histidine which was found to prefer metal coordination sites more than salt bridges**.** The highest average value of E_m^{sc} was obtained for Glutamate (0.68) which also had the highest increment in $<E_m^{sc}>$ upon inclusion into a salt bridge. Arginine exhibited the highest propensity to form ionic bonds (5.83), compared to other charged residues (Glu: 4.77, Asp: 3.92, Lys: 3.43). The participation of histidine in such networks was by and large negligible (propensity: 0.22). The highest value in $<E_m^{sc}>$ was obtained for Glutamate (0.68) amongst salt bridge forming residues.

Propensity (*Pr(x,s)*) of a charged residue, *x* to go into a salt bridge was computed by the following expression:

$$\Pr(x,s) = \frac{\frac{N(x,s)}{N(t,s)}}{\frac{N(x,d)}{N(t,d)}}$$

where *N(x,s)* is the count of the residue *x* found in salt bridges, *N(t,s)* is the total number of residues involved in salt bridges, *N(x,d)* and *N(t,d)* is the count of residue *x* and the total number of residues in the database.

Several instances have been recorded where bifurcated salt bridges contribute more to the electrostatic stabilization within proteins than the isolated ionic bonds **(Torshin et al., 2002; Di Primo et al., 1997; Walker and Causgrove, 2009)**. Bifurcated salt bridges were further analyzed for compositional and geometrical bias **(Table 1)**. Compositional preferences were clearly distinguishable for arginine containing salt bridges (75.4% of the whole set) with Glu-Arg-Glu having the highest occupancy (17%). Similar preferences have been previously observed for arginine-glutamate salt bridges in case of helix stability **(Walker and Causgrove, 2009)**.

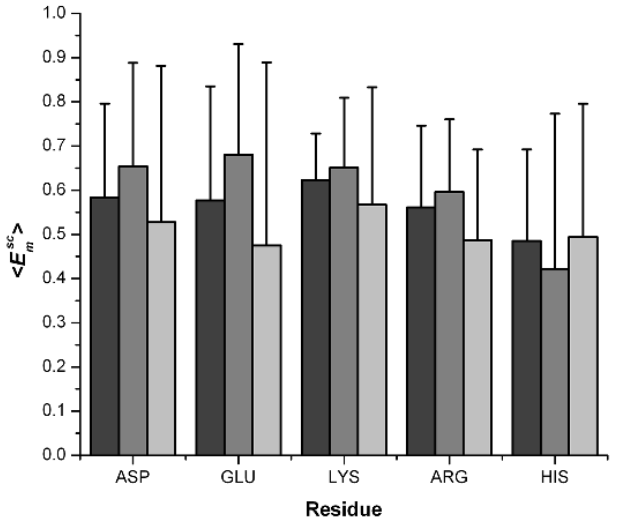


Figure 2. Charged residues involved in salt bridges lead to only mild enhancement in E_m^{sc}. The figure shows mean values of E_m^{sc} (filled thick bars) along with their standard deviations (thin error bars) for charged residues involved in salt bridges (gray), not involved in salt bridges (light gray) and pooled together (deep gray). As can be seen, histidine (positively charged) shows a reverse trend.

The angle subtended by three residues forming a bifurcated salt bridge was computed as follows: except for lysine, which has a unique charged nitrogen atom (NZ), the effective (or resultant) charge centers were determined as the midpoint of the two (degenerate) charged O (aspartate, glutamate) and N (arginine, positively charged histidine) atoms. The bifurcation angle (γ) between the two vectors connecting the three charge-centers was then computed. Geometry of the bifurcated salt bridges were analyzed in terms of the bifurcation angle, γ which was found to be obtuse and fairly well constrained (~110° ± 30°) irrespective of the residue composition.

Table 1. Composition and Geometry of the Bifurcation angle (γ)

Composition	Count	Percentage	<γ>
GLU-ARG-GLU	81	17.05	120.6 (28.2)
ASP-ARG-GLU	52	10.95	121.1 (26.1)
GLU-ARG-ASP	45	9.50	110.8 (29.1)
ASP-ARG-ASP	43	9.05	112.3 (27.3)
ARG-GLU-ARG	35	7.37	97.4 (37.4)
ARG-ASP-ARG	28	5.89	92.9 (30.7)
GLU-LYS-GLU	26	5.47	116.0 (21.7)
LYS-GLU-ARG	26	5.47	102.1 (27.7)
ARG-GLU-LYS	26	5.47	100.4 (40.7)
ASP-LYS-ASP	26	5.47	93.6 (26.1)
LYS-GLU-LYS	22	4.63	113.8 (26.0)
GLU-LYS-ASP	18	3.79	109.0 (14.3)
ASP-LYS-GLU	13	2.74	115.3 (22.9)
ARG-ASP-LYS	13	2.74	93.8 (33.9)
LYS-ASP-ARG	8	1.68	95.4 (20.9)
LYS-ASP-LYS	7	1.47	76.8 (8.4)
LYS-ASP-HIS	2	0.42	126.1 (15.4)
GLU-HIS-GLU	1	0.21	165.4
ARG-GLU-HIS	1	0.21	70.9
GLU-HIS-ASP	1	0.21	70.1
ASP-HIS-ASP	1	0.21	169.6

References

McCoy AJ, Epa VC, Colman PM (1997). **Electrostatic complementarity at protein/protein interfaces.** *J Mol Biol,* **268**: 570-584.

Honig B, Yang A-S (1995). **Free energy balance in protein folding.** *Adv Protein Chem* **46**: 27-58.

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

Torshin IY, Weber IT, Harrison RW (2002). **Geometric criteria of hydrogen bonds in proteins and identification of 'bifurcated' hydrogen bonds**. *Protein Eng* **15**: 359-363.

Di Primo C, Deprez E, Sligar SG, Hui Bon Hoa G (1997). **Origin of the photoacoustic signal in Cytochrome P-450cam: role of the Arg186−Asp251−Lys178 bifurcated salt bridge.** *Biochemistry* **36**: 112–118.

Walker KD, Causgrove TP (2009). **Contribution of arginine-glutamate salt bridges to helix stability** *J Mol Model* **15**: 1213-1219.

Publications

Basu et al. BMC Bioinformatics 2011, 12:195
http://www.biomedcentral.com/1471-2105/12/195

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Mapping the distribution of packing topologies within protein interiors shows predominant preference for specific packing motifs

Sankar Basu[1], Dhananjay Bhattacharyya[2] and Rahul Banerjee[1*]

Abstract

Background: Mapping protein primary sequences to their three dimensional folds referred to as the 'second genetic code' remains an unsolved scientific problem. A crucial part of the problem concerns the geometrical specificity in side chain association leading to densely packed protein cores, a hallmark of correctly folded native structures. Thus, any model of packing within proteins should constitute an indispensable component of protein folding and design.

Results: In this study an attempt has been made to find, characterize and classify recurring patterns in the packing of side chain atoms within a protein which sustains its native fold. The interaction of side chain atoms within the protein core has been represented as a contact network based on the surface complementarity and overlap between associating side chain surfaces. Some network topologies definitely appear to be preferred and they have been termed 'packing motifs', analogous to super secondary structures in proteins. Study of the distribution of these motifs reveals the ubiquitous presence of typical smaller graphs, which appear to get linked or coalesce to give larger graphs, reminiscent of the nucleation-condensation model in protein folding. One such frequently occurring motif, also envisaged as the unit of clustering, the three residue clique was invariably found in regions of dense packing. Finally, topological measures based on surface contact networks appeared to be effective in discriminating sequences native to a specific fold amongst a set of decoys.

Conclusions: Out of innumerable topological possibilities, only a finite number of specific packing motifs are actually realized in proteins. This small number of motifs could serve as a basis set in the construction of larger networks. Of these, the triplet clique exhibits distinct preference both in terms of composition and geometry.

Background

Despite several decades of arduous effort, mapping of protein primary sequences to their three dimensional folds, referred to as the second genetic code, remains an unsolved scientific problem. What appears to be lacking is a comprehensive theory, integrating two factors which definitely condition the isomorphism between sequence and fold, namely (1) the pattern of hydrophobicities embedded in the polypeptide chain [1] and (2) the packing of amino acid side chains to give densely packed [2] protein interiors. Under the present circumstances, the more tractable approach is the 'inverse protein folding problem' [3,4], that is to identify protein primary sequences [5] consistent with and supportive of a given fold, an idea which has found considerable application in the *de novo* design of targeted protein structures [6-9]. Yet even here, it was realized earlier on, that in *de novo* design, attainment of dense, well-packed protein cores (a hallmark of native, correctly folded proteins) was neither an automatic part of the design process nor acquired simply by chance [10,11]. Most often, it was observed (especially for longer sequences) that design led to molten globules or complete unraveling of the structure [12,13]. An instructive example was the repeated failure to design parallel $(\alpha/\beta)_8$ - TIM barrel [14,15], finally resolved successfully by Offredi *et al.* [16], where a term optimizing for side chain packing specificity was deliberately included in the

* Correspondence: rahul.banerjee@saha.ac.in
[1]Crystallography and Molecular Biology Division, Saha Institute of Nuclear Physics, 1/AF, Bidhannagar, Kolkata - 700 064, India
Full list of author information is available at the end of the article

Self-Complementarity within Proteins: Bridging the Gap between Binding and Folding

Sankar Basu,[†] Dhananjay Bhattacharyya,[‡] and Rahul Banerjee[†*]
[†]Crystallography and Molecular Biology Division and [‡]Biophysics Division, Saha Institute of Nuclear Physics, Kolkata, India

ABSTRACT Complementarity, in terms of both shape and electrostatic potential, has been quantitatively estimated at protein-protein interfaces and used extensively to predict the specific geometry of association between interacting proteins. In this work, we attempted to place both binding and folding on a common conceptual platform based on complementarity. To that end, we estimated (for the first time to our knowledge) electrostatic complementarity (E_m) for residues buried within proteins. E_m measures the correlation of surface electrostatic potential at protein interiors. The results show fairly uniform and significant values for all amino acids. Interestingly, hydrophobic side chains also attain appreciable complementarity primarily due to the trajectory of the main chain. Previous work from our laboratory characterized the surface (or shape) complementarity (S_m) of interior residues, and both of these measures have now been combined to derive two scoring functions to identify the native fold amid a set of decoys. These scoring functions are somewhat similar to functions that discriminate among multiple solutions in a protein-protein docking exercise. The performances of both of these functions on state-of-the-art databases were comparable if not better than most currently available scoring functions. Thus, analogously to interfacial residues of protein chains associated (docked) with specific geometry, amino acids found in the native interior have to satisfy fairly stringent constraints in terms of both S_m and E_m. The functions were also found to be useful for correctly identifying the same fold for two sequences with low sequence identity. Finally, inspired by the Ramachandran plot, we developed a plot of S_m versus E_m (referred to as the complementarity plot) that identifies residues with suboptimal packing and electrostatics which appear to be correlated to coordinate errors.

INTRODUCTION

All forms of biomolecular recognition are said to involve interaction between complementary molecular surfaces. This specific match between two interacting surfaces is primarily supposed to have a dual aspect: 1) surface (or shape) complementarity (1) arising out of the steric fit of closely packed interface atoms in van der Waals contact; and 2), electrostatic complementarity (2) mediated by long-range electric fields due to charged or partially charged atoms. For small-molecule ligands or cofactors binding to proteins, the above point of view appears to be only partially true. Not only can one ligand adopt a wide range of conformations upon binding to different proteins, the binding pocket also exhibits more variability in shape and physicochemical characteristics than can be accounted for by the multiple conformations adopted by the ligand (3–5). For protein-protein interfaces, however, the concept appears to have greater plausibility and wider appeal. Due to the relatively larger size of protein-protein interfaces (~1600 Å^2 on average) (6), the surfaces have to be carefully tailored so that extended areas buried upon association can move into close contact. A variety of shape correlation and electrostatic complementarity measures incorporated into docking algorithms have been shown to be effective in predicting the interfaces between interacting proteins (7,8). Electrostatic complementarity based on optimized charge distribution has also been used to match two halves of the same molecule (myoglobin) from a repertoire of homologous structures (9). On the other hand, surface complementarity has found application in determining native side-chain torsions within proteins (10,11) and has also served to rationalize the variability in the quaternary arrangements of legume lectins (12). Lawrence and Colman (1) and McCoy et al. (2) formulated and estimated shape correlation (S_c) and electrostatic complementarity (EC) measures for a wide range of proteins in quaternary association, protein-inhibitor, and antigen-antibody complexes. It thus appears reasonable that threshold values of geometric and electrostatic complementarities will have to be satisfied for the stereospecific association between two polypeptide chains. Within proteins, surface complementarity (S_m) has been used to enumerate specific modes of packing between amino acid side chains (13) and, somewhat analogously to protein interfaces, all residues upon burial achieve uniformly high measures of surface fit (14).

Although the notion of complementarity lends itself naturally to the characterization of interprotein association, it has been suggested that both binding and folding should be approached from a common conceptual platform (15,16). The native conformation adopted by the polypeptide chain leads to the stereospecific packing of its buried side chains and optimal electrostatic interactions due to the strategic three-dimensional placement of charges. Thus, folding can possibly be described as the self-recognition of the polypeptide chain as it collapses onto itself. However, one inherent problem in equating binding with folding lies

Submitted December 8, 2011, and accepted for publication April 17, 2012.

*Correspondence: rahul.banerjee@saha.ac.in

Editor: Bertrand Garcia-Moreno.

0006-3495/12/06/2605/10 $2.00

doi: 10.1016/j.bpj.2012.04.029