STRUCTURAL INSIGHTS INTO FOLDED, UNFOLDED AND NASCENT PROTEIN STATES USING ENSEMBLE SAMPLING AND CLUSTER EXPANSION

ARUN CHANDRAMOHAN

(*M.Sc. University of Madras, India*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

IN

COMPUATIONAL AND SYSTEMS BIOLOGY (CSB)

SINGAPORE-MIT ALLIANCE

NATIONAL UNIVERSITY OF SINGAPORE

2014

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Arun Chandramohan

24 Jan 2014

Acknowledgements

I'm grateful to my four supervisors, Dr. Chris Hogue, Prof. Bruce Tidor, Prof. Greg Tucker-Kellogg and Prof. Paul Matsudaira for guiding me at different stages of my thesis. First, I would like to express my heartfelt gratitude to Dr. Chris Hogue for giving me the opportunity to choose my project myself and letting me to work on it in his lab. The complete freedom he provided me during the course of my research in his laboratory helped me immensely. His thought process and ability to look at things differently have inspired me. I would like to thank my co-supervisor Prof. Greg Tucker-Kellogg for academically adopting me to his lab and for his excellent guidance ever since. I feel lucky that I have always enjoyed informal and friendly guidance rather than strict supervision from the both of them and I'm deeply grateful for that. The most productive time of my thesis was at MIT with Prof. Bruce Tidor. His systematic approach has greatly improved me as a researcher and I would like to thank him for his guidance and supervision during my time at MIT. I'm also grateful to Prof. Paul Matsudaira for taking me under his wing after Chris's departure and helping me get through the last semester of my thesis.

None of this would be remotely possible if it were not for my parents, Chandramohan and Usha. Their constant support has helped me battle homesickness and has managed to keep me motivated throughout the course of my study. I cannot for a moment think that this thesis would be possible without the both of them and my sister, Uma. I cannot begin to thank them for their love and sacrifice. I would also like to thank my uncles Satish and Ganesh and their families for making me feel at home in Singapore.

Another equally important achievement during my time here is that I have made a lot of great friends who have made my time here thoroughly enjoyable. I'm truly blessed to have had amazing friends: Suhas, Vasanth, Parakalan, Sriram and Karthik as roommates for four years along with Asfa and Madhu. Sachin has been a close friend

right from school and I would like to specially thank him for great company and a lot of support during my time here. Special thanks to Hari and Arun who have extended great friendship and support during tough times.

I would like to thank Srinath for help with experiments and discussions. I would also like to thank Soumya for helping me out with wet-lab experiments, Yin-Ru for her help with DSMs, Nate at MIT for his help with codes and theory and Jeremy for proofreading my thesis.

I would like to thank my lab mates Liu Chengcheng, Zhang Bo, Zhao Chen, Sowmya KP for their support and Sihan for being a great housemate and friend at Boston.

Finally, I would also like to thank Prof. Gong Zhiyuan, Singapore-MIT Alliance, Mechanobiology Institute and the Department of Biological Sciences for their financial support over these years.

Table of Contents

Summary

The unfolded and nascent states of proteins are incompletely understood due to both methodological and conceptual challenges. Deciphering energetic interaction maps within protein structures also suffers from methodological limitations, since experimentally characterising them is cumbersome. In this work, these challenges are addressed using a strategy that leverages the computational tools of ensemble modelling and cluster expansion.

Ensemble modelling is used to generate the denatured state ensemble (DSE) of proteins from their sequences by sampling specific regions of protein conformational space. The ensemble model demonstrates excellent agreement with experimental small-angle X-ray scattering (SAXS) data on the radius of gyration of denatured proteins. The DSEs have been used to study native-like contiguous residual structures in set of proteins by comparison to their respective crystal structures. The residual structures contain secondary structural elements such as β-turns and very short motifs that could act as nucleating sites during protein refolding. This provides the first all-atom model of a set of DSEs with details of native-like structures and their implications on protein refolding.

Nascent polypeptides traverse a 100 Å long exit tunnel before they emerge from the ribosome. The dynamics of the polypeptide and its conformations inside the tunnel are unknown. Here, we study the polypeptide conformations and dynamics in the upper to central peptide tunnel with Delaunay triangulation and molecular dynamic simulations to investigate joint tunnel and polypeptide dynamics. Ensemble sampling and dock by superposition are used to describe the complete conformational space of the peptide in the tunnel. We find a decrease in volume and increase in surface area of the tunnel when the nascent chain is present, indicating a collapse of the tunnel and an increase in its surface convolution. Finally, *de novo* sampling and dock by

superposition detail the complete conformational space accessible to peptides at every segment of the tunnel. This shows increasing spatial freedom towards the exit which lets the peptides access both helical and extended conformations.

Disordered proteins have flat energy landscapes and sampling them is precarious. TraDES is a MonteCarlo-based structure sampling tool which can provide structural representations of disordered proteins based on random-coil sampling. These ensembles can be used to describe the disordered protein's structures and help in understanding the order and disorder in the structures. These ensembles provide excellent starting points for including experimental data as constraints to gain structural insights on the mechanism of disordered proteins.

Deciphering energetic maps within proteins is essential to understanding allosteric communication. Energetic coupling between different residue positions in proteins is identified by representing the energy of the protein using cluster expansion.

Cluster expansion breaks down complex interactions into single and pair-wise coupling functions. This energetic coupling is used to describe and predict three interacting clusters of positions in the WW-domain of the peptidyl-prolyl isomerase protein Pin1. Experimental evidence verifies that these positions identified in the network play important roles in protein stability and function.

Key words: denatured proteins, disordered proteins, allostery, ribosome, nascent polypeptide, energetic interactions, cluster expansion, molecular dynamics, ensemble sampling.

List of Tables

List of Figures

Abbreviations

Gd-HCl	Guanidinium hydrochloride
Rgyr	Radius of gyration
ACBP	Acyl-CoA-binding protein
DSE	Denatured state ensemble
SAXS	Small-angle X-ray scattering
FTIR	Fourier transform infrared
NMR	Nuclear magnetic resonance
RDC	Residual dipolar coupling
PRE	Paramagnetic relaxation enhancements
NOE	Nuclear Overhauser effect
NP	Nascent polypeptide
MD	Molecular dynamics
TraDES	Trajectory Directed Ensemble Sampling
RNase	Ribonuclease
IDP	Intrinsically disordered protein
CD	Circular dichroism
FRET	Förster resonance energy transfer
SM-FRET	Single molecule-FRET
MC	Monte Carlo
PDB	Protein Data Bank
PDF	Probability density function
PPII	Polyproline II helix

TRAP	TraDES-R-Analysis Package
NC	N-terminal to C-terminal (end-to-end)
CC	C-α to C-α
100%E	100% Extended
100%C	100% Random-coil
AAD	Average Absolute Difference
AR	Average Ratio
FM	Flexible Meccano
PTC	Peptidyl transferase centre
RNA	Ribonucleic acid
DT	Delaunay triangulation
GB	Generalised Born
HCAP	Human cancer-associated proteins
PFG-NMR	Pulsed field gradient NMR
NCBI	National Center for Biological Information
MSA	Multiple Sequence Alignment
RMSE	Rootmeansquared error
CV	Cross-validation

1. Introduction

"The extreme rapidity of the refolding makes it essential that the process take place along a limited number of "pathways", even when the statistics are severely restricted by the kinds of stereochemical ground rules that are implicit in a so-called Ramachandran plot. It becomes necessary to postulate the existence of a limited number of allowable initiating events in the folding process. Such events, generally referred to as nucleations, are most likely to occur in parts of the polypeptide chain that can participate in conformational equilibria between random and cooperatively stabilized arrangements."

This theory was postulated in 1972 by Christian Anfinsen during his Nobel lecture and has not yet been convincingly proven after all these years. Only recently have experimental evidences been pointing towards these nucleating structures and their effects on protein folding. Anfinsen's original work on ribonuclease had provided the first model of protein folding in which the entire process is driven by its free energy gain during the process of attaining the native structure. These energetics have established in detail how proteins always favour the narrow low-energy well of the native structure.

Although Anfinsen had postulated the existence of nucleating structures, the denatured state was long thought to lack any ordered structure and was projected as a completely random model. Evidence of residual structures has changed this notion but no correlation has been drawn between such residual structures and the originally proposed nucleating structures. Native-like residual structures have been identified using NMR in the denatured state and provide force to the theory of them being important initiators during protein folding. But these studies have been scarce.

Current understanding of protein folding models is derived from studies of folding in individual simple globular proteins. It is extremely exhaustive to develop a general model of protein folding from these individual studies due to the large heterogeneity in protein structure. An alternative approach would be to model the denatured state of proteins, which can be used to study the mechanism of an array of proteins.

Another completely distinct pathway to protein folding has been co-translational folding, where proteins fold during synthesis with faster kinetics. The mechanism and its component members are poorly understood. An important partner is the ribosome and its peptide tunnel. Due to the ribosome being large and the tunnel being enclosed, the nascent polypeptide and its structure has not been studied but there are evidences on its importance in folding.

In this thesis, we will address the problem of understanding protein structural landscape in three realms of unfolded state: denatured, intrinsically disordered and nascent states using structure sampling.

In Chapter 2, we will use ensemble structure sampling to model the dimensions of a set of denatured proteins and describe the residual structures in them. In addition, we will also correlate the predicted structures with experimental data and derive meaningful insights into protein refolding.

In Chapter 3, we look at the nascent polypeptide structure inside the ribosome using Monte Carlo and molecular dynamics (MD) simulations. The effects of the polypeptide on the ribosome tunnel and its dynamics are studied using MD simulations. The geometry of the tunnel and the structures that can be accommodated are described by ensemble modelling and dock by superposition. Finally, the presence of any spatial thresholds in the tunnel is verified by generating different ensembles corresponding to every threshold and analysing their quality.

In Chapter 4, we deal with the large disordered landscape and use the TraDES package to develop representative structural ensembles that match random-coil sampling models.

In Chapter 5, we use statistical mechanics to calculate the energy of any sequence in a given structure. This is further broken down to understand individual interactions between different positions in the protein and amino acids. Finally, energetically interacting clusters in the WW-domain protein structure are identified.

Finally, Chapter 6 summarises the thesis and provides concluding remarks. Outlook and future directions are also briefly described.

1.1 Unfolded state

Proteins exist in multiple structural states which vary from the low-energy folded structure to the unstructured random-coil form. These various states are important parts along the protein folding/refolding pathway. The initial starting point is the denatured state and the end point is the folded structure with partially-folded, intermediate and disordered states in between. Until recently, proteins were believed to be functional only in their folded state and its native structure was the focus of most of the early structural studies while the unfolded state was considered to be non-functional. The unfolded proteins were considered to follow a random coil model [1] which follows polymer-like behaviour [2]. Initially, the unfolded realm of proteins comprised only of the denatured state. The denatured state had been observed and its properties were attributed to the loss of a protein's native structure [3] long before its structure was established [4]. Early studies confirmed the complete loss of protein structure in the presence of denaturants such as urea and guanidinium hydrochloride (Gd-HCl) [5]. Once the protein structure was solved, it remained the focus of all structural efforts. Proteins were also found to exist in an intermediate molten globule state [6-8] with native-like topology and tertiary structure during its folding. First hints of a complex unfolded structure came from FTIR and electron microscopy of tau protein [9] and 'non-A beta component of Alzheimer's disease amyloid plaque protein' NACP [10], which exhibited an absence of secondary structures even under physiological conditions and were labelled as "natively denatured/unfolded". Since then, a number of such intrinsically disordered proteins have been studied. First evidences of complex structures in denatured proteins were also soon established by NMR and site-directed mutagenesis in staphylococcal nuclease (SNase) [11] and NMR studies on the 434-repressor N-terminal [12]. Unlike the small window of possible folded states, the unfolded landscape is much larger in denatured and intrinsically disordered proteins. Thus, the disordered and denatured states behave

more similarly to random coils than ordered structures and are represented as complex structural ensembles. The disordered proteins are flexible, while the denatured proteins are extended. They have larger hydrodynamic and structural dimensions than folded proteins with similar residue lengths or molecular mass. Their dimensions are also distributed over a larger range of values making them harder to characterise by a single method. Every protein state has its implications on the protein functions. The denatured state ensemble (DSE) may hold important clues to decipher the folding mechanism and also aid the protein in its folding kinetics. The intrinsic disordered ensemble directly relates to its function, flexibility in structure and its ability to bind to diverse ligands and regulation. In summary, the protein unfolded space is vastly larger than its folded space, has important functions and is hard to characterise due to its heterogeneity.

1.1.1 Denatured state

Proteins generally exist in their native low-energy structure in physiological conditions. This structure can be disrupted by physical, chemical and biological agents (denaturants) [5,13]. This disruption of structure from its folded state [14] is due to the conformational changes that take place in the peptide chain. This disruption is evidenced by changes in the chemical and physical properties of the protein such as aggregation [15]. This process of altering the structure of the macromolecule with no change in its molecular weight is defined as denaturation [1]. The disruption occurs in the secondary interactions such as van der Waals, hydrogen bonds, ionic and hydrophobic interactions [16] whereas the primary peptide bonds [17] are largely unaffected.

Various levels of a protein's native structure are stabilised by a range of non-covalent interactions. Hydrogen bonds contribute to the secondary structure, while hydrophobic and van der Waals forces are responsible for the tertiary structure. Denaturants are known to disrupt these interactions resulting in a random coil

structure [3,18]. This original model of the absence of structure in the denatured state was proposed by Anfinsen [19]. The theory was based on *in vitro* denaturation experiments on ribonuclease [20], Staphylococcal nuclease [21] and RNase [22,23]. This led to the random-coil hypothesis, which suggested that during denaturation, proteins lose their structure and adopt a random coil statistical state. Proteins were also understood to regain their native conformation upon removal of the denaturing conditions. But, recent evidences have shown the presence of ordered structures in the denatured state and these results have changed the random-coil hypothesis.

Various studies point to transiently populated structures in the denatured state which contain information on folding initiation sites [24-30]. Residual structures in proteins are also known to affect protein thermodynamic stability, as evidenced by the changes in melting temperatures of RNase H [31]. The late events along the folding pathway are well understood, while early events are largely undetermined. Hydrophobic collapse is the earliest event that has warranted significant attention, although residual structures and nucleating structures are present even during initialisation. Studies on the denatured states of Staphylococcal protein G variants, with similar sequences and dissimilar structures, show the presence of structural determinants in the denatured state [32]. Such native-like tertiary contacts have also been identified in the ensemble of the four-helix Acyl-CoA binding protein (ACBP) [33,34] and Staphylococcal nuclease [35]. These regions are reported to determine the native topology of each variant and prevent one variant from adopting the structure of the other.

The mechanism of disruption of protein structure by denaturants was poorly understood until recently. NMR data shows that urea binds directly to the peptide backbone [36] and elongates the protein chain [37] with increasing concentrations [38].

Unlike the folded state, the denatured state cannot be a single point of reference or a set of structures but rather a complex result of the protein sequence and it's solvent. The denatured state is described as a distribution of various protein micro-states that depend on its denaturing conditions and sequence [39]. Taking these into consideration, the denatured state is represented in the form of an ensemble, the denatured state ensemble (DSE). The DSE of a given protein depends on many factors such as its sequence, denaturing condition, etc.

Different denaturants disrupt the structure of proteins differently, evident from their dimensions in different solvents. Denaturation by urea and Gd-HCl results in significant differences in the hydrodynamic radius of proteins [40]. Refolding from different solvents could mean different starting points with the same resultant folded state. Studying denatured protein structure under different conditions has implications due to their bearing on the folding pathway.

Residue types also play a major role in determining the properties of the DSE. Electrostatics has been known to make significant contributions, favourable or otherwise, to the energetics of the DSE [41-44]. Mutating seven hydrophobic residues to serine in kinase-inducible activation domain shows no significant change in the dimension (Rgyr) of the DSE, while mutation of eleven charged residues leads to sizable compaction [45]. Hence, contribution of charged residues is larger than non-polar residues to the stability and compaction of the DSE, although they can stabilise or disrupt the DSE [46,47].

It is imperative to fully characterise the denatured state of proteins since they can provide important clues on protein refolding. The denatured state can be thought of as a starting point in the protein refolding/folding pathway with its termination at the folded state. Understanding the make-up of the denatured state would provide us with information on the starting state from which protein refolding happens.

1.1.1.1 Structured?

The denatured state of proteins was assumed to be largely random-coil and follow polymer chain behaviour [1], until NMR studies on a globular protein in strongly denaturing conditions provided experimental observations of residual structures [12]. Similar studies also showed the presence of ordered structure in strongly denaturing conditions [48,49]. RNase A was originally thought to unfold completely and populate a random-coil conformation in strong denaturing conditions such as 6 M Gd-HCl and 8 M urea. SAXS and FTIR analysis of RNase later disagreed with a completely random-coil model with evidences of a more compact structure. The denatured structure, unlike the folded state, is a heterogeneous ensemble of conformations and makes it very hard to characterise. SAXS studies demonstrate that the DSE contains a range of compact to elongated structures [50]. The ensemble is structurally diverse and requires a range of methods to provide a good estimation of its properties.

1.1.1.2 Radius of gyration

The dimensions of proteins are a good indicator of whether the protein is structured and compact. Folded proteins have a lower Rgyr values due to hydrophobic collapse. This compact structure is held in place mainly by disulphide, electrostatic and van der Waals forces. The average radius of gyration of a molecule is an indicator of its compactness. In the case of proteins, it represents how folded or unfolded the structure is. The radius of gyration is the root-mean-squared distance of the centres of all the individual atoms from the centre of mass of the whole molecule.

$$R_{\mathrm{g}}^2 \stackrel{\mathrm{def}}{=} \frac{1}{N}\sum_{k=1}^{N}\left(\mathbf{r}_k - \mathbf{r}_{\mathbf{mean}}\right)^2 \qquad \text{Equation 1}$$

r_k is the position of each atom in the structure and r_{mean} is the mean position of all the atoms. In essence this tells us how compact the structure is.

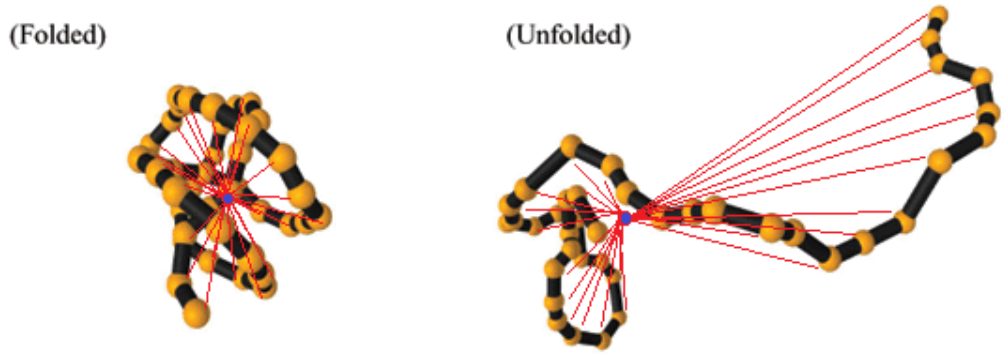

Figure 1.1 Radius of gyration as a measure of foldedness. Illustration of the difference in radius of gyrations between different structures. A folded structure has a smaller radius of gyration than an unfolded structure due to its compactness.

A random-coil model is devoid of all these interactions and is expected to have a large Rgyr value, since it completely extended. But, SAXS and FTIR studies show that chemically and thermally denatured RNase A has a Rgyr value of ~33 Å, which is lower than the expected Rgyr value for a random-coil model. The folded average radius of gyration is determined to be 15.9 Å, while the random-coil denatured state is expected to have an Rgyr value of ~48 Å. Radius of gyration measurements for the folded state are more precise because folded proteins typically exist as native structures, while the denatured state exists as a heterogeneous ensemble in solution. The fact that RNase in the denatured state has a lower Rgyr value than the predicted value in the random-coil model re-emphasises the presence of ordered structures in it. These residual ordered structures compacts the ensemble and results in a lower Rgyr value.

1.1.1.3 Experimental characterisation

NMR can provide valuable insights on the denatured state ensemble (DSE). NMR can provide NOE, chemical shifts, paramagnetic relaxation enhancement and residual dipolar coupling data to obtain useful information from the denatured ensemble [51]. NMR RDCs, PREs, hydrogen-exchange protection factors, SAXS and other measurements were used as constraints to filter and choose good structural ensembles from computationally generated structures of the drkSH3 domain [52]. Alpha-

synuclein has been studied similarly using PRE measurements as restraints in molecular dynamics simulations [53]. Secondary chemical shift analysis has been used to describe non-native transient structures in the SH3 domain of an all-beta protein c-Src [54]. The report also notes the differences in residue-wise structural propensities between various SH3 domains. Paramagnetic resonance is one of the most successful techniques used for the detection of long-range contacts in disordered protein ensembles [34,49,55-57]. RDCs are also very popular in understanding such disordered ensembles [35,58,59]. PRE appears to be better at picking up transient contacts that occur in such ensembles compared to RDCs [57,60].

Experimental data for the denatured state share the same caveats with the disordered state. The data obtained from most of the techniques are ensemble-averaged. Some experimental data such as NOE and PRE are biased towards close contacts. These are weighted averages dependent on the r^{-6} on the dipole interactions. This creates a major hindrance in using these data to understand and interpret structures in the ensemble.

1.1.2 Nascent state

The least studied of all the protein states is the nascent peptide state. It is not even considered a part of the landscape since it manifests only inside the ribosome tunnel and shortly after release. It is the starting point of the folding process from synthesis. Its absence from literature can be explained by its very short life-time outside the ribosome before it transitions through folding intermediates into a folded state [61] and they are protected by nascent chain-binding chaperones [62]. This limited time-frame poses a significant challenge to biophysical and biochemical methods that are used to study other states. Another complication is the process of co-translational folding, where nascent polypeptides fold during translation even when they are attached to the ribosomes [63]. A large number of proteins have been shown to attain their native structures during translation [64-66].

This does not explain folding of proteins with β-topology since the N-terminal residues may have to wait for the C-terminal residues to form long-range interactions to form β-strands. The nascent starting structure for protein folding is different from the denatured starting state for protein refolding, since they are shown to follow different pathways. The nascent chain folding is much faster than denatured protein refolding [61,63]. Chaperones do play an important role in protein folding but they are responsible for less than 20% of the *E. coli* cytosolic protein folding [62]. There is, in addition to chaperones, a mechanism that helps proteins fold faster and better than refolding. The nascent state of proteins contain favourable conformations that can pass through kinetic and thermodynamic traps [66,67] faster and might go through a lesser set of conformations before reaching the folded state. This advantage is partially coded onto the sequence, and sequences with "frustrated" energy landscapes [68] are selected out by evolution. But this does not explain how the same sequence folds faster co-translationally compared to its refolding. The nascent state is a better starting point than the denatured state and becomes a key parameter to understand the faster rates of co-translational folding.

The co-translation folding has also been shown to happen inside the ribosome with the tunnel containing folding zones [69] and the detection of folded tertiary structure inside the tunnel [70]. This implies that the peptides may already be well into the folding pathway even before they are exiting the ribosome. The tunnel is itself a matter of debate. Purely structural studies report that the tunnel cannot accommodate any folded structure [71] but the tunnel is also known to expand during translation [72] for which high-resolution data is unavailable. These conflicting results can be explained by the different states of the ribosome in each of them (normal state and translating state). The nascent polypeptide passes through a constriction and the constriction acts as a gating mechanism [73]. In turn, the peptide also seems to have an effect on the tunnel and its geometry. Study of the whole system including the

tunnel and the peptide will provide new insights on the effect of constriction of the tunnel on the peptide and that of the peptide on the tunnel. So, in order to understand the nascent state of proteins, it is important to decipher their structures inside the ribosome together with a cause-and-effect relationship with the ribosome tunnel.

1.1.2.1 Folded?

There are evidences of enzymes that can show their catalytic activity as soon as they exit from the ribosome tunnel [61]. Enzyme activity has been detected in peptides that are still bound to the ribosome [74-76]. Nascent chains possess the capacity to bind to their specific ligands during their synthesis. This is also true for cofactor binding studies. These show that the nascent peptide has an active conformation during translation and while attached to the ribosome. This has been thought to happen outside the tunnel even as they are tethered to the ribosome. PEG coupling of folded, engineered cysteines inside the ribosome has been captured as they cause significant shifts in gel electrophoresis. Cysteines have been engineered in beta and alpha secondary structures and both show considerable folding and acquisition of secondary and tertiary structure inside the tunnel, closer to its exit [70]. Recent evidence of structure acquisition inside the peptide tunnel and the presence of folding zones inside the tunnel have given rise to the possibility that these structures could be pre-formed inside the ribosome tunnel.

1.1.3 Disordered state

Since early structural studies, protein function has been attributed to its rigid three-dimensional (3D) native folded structure [18,19]. This suggests that a protein can fold into its designated low-energy native state and its function is a consequence of its rigid 3D structure [77-79]. Later, this theory was relaxed for induced-fit model, which is currently the most widely accepted structure-function relationship [80]. Recent evidences demonstrate the presence of many functional proteins that do not possess a defined stable 3D structure [81]. These complete or parts of proteins that exist

entirely or partially in disordered state are exceptions to the above model and are referred to as intrinsically disordered proteins (IDPs) [82-84]. The amount of disorder and its range vary between IDPs. Proteins like p53 contains stretches of ordered and disordered structures [85], while the Tau protein is entirely disordered over its 440 residues [86].

Their disorder can be attributed to their sequence composition. They contain higher than average ratios of surface or hydrophilic amino acids (A, R, G, Q, S, P, E and K) to core hydrophobic amino acids (W, C, F, I, Y, V, L and N) [40,87,88]. The low composition of hydrophobic residues reduced hydrophobic collapse, which is mainly responsible for folding the core protein [89]. IDPs have low sequence complexity and high sequence variability [84].

IDPs lack a unique, stable 3D structure but rather exist as an ensemble of structures and conformations. IDPs are attributed to many important functions such as transcription regulation and signal transduction [83,90]. IDPs have a structural advantage due to their disordered structures. The disorder in their protein structures allows IDPs to bind to multiple partners and is involved in multiple interactions. This binding promiscuity [91] is the main reason for IDPs to take up important hubs in protein interaction networks [92,93]. These are involved in important signalling pathways for recognition and in the regulation of various binding partners [94,95]. Due to their importance in multiple protein pathways and networks, IDPs are involved in various diseases [96]. An unfoldome has been implicated in a network analysis of human genetic diseases, suggesting that disorder is prevalent among proteins involved in them [97,98].

1.1.3.1 Challenges with disordered states

IDPs, due to their high propensity for disorder, are best represented by ensembles of structures. They have been referred to as ‘protein clouds’, where the conformation of the backbone varies with no equilibrium values [99]. These disordered states are

heterogeneous and rapidly interconvert between different conformations posing challenges to the field of structure determination [100]. It is also not possible to use global definitions for disordered proteins since they have been shown to exist in distinct groups with different structural compositions. Reports based on far-UV, circular dichroism (CD) spectra and hydrodynamic radius clearly indicate that some proteins possess larger amounts of secondary structures than others [40]. CD and hydrodynamic studies have allowed the classification of unfolded proteins into two structurally distinct categories: intrinsic coils and intrinsic pre-molten globules. This makes it harder to define general rules and develop methods for studying structure in unfolded proteins.

1.1.3.2 Structure in disordered state

The presence of structure in the disordered state is an ongoing debate in the literature. It is unclear if the disordered state should be described as a random-coil or to consider them to contain fluctuating structures [101]. But, it is currently well established that disordered proteins contain considerable amounts of local and long-range structures. Alpha-synuclein is one of the well-studied intrinsically disordered proteins. Using PRE distance constraints and MD simulations, synuclein has been found to contain structures with long-range interactions in the C-terminal tail [102]. This has been supported by the detection of large populations of conformers that exhibit long-range contacts between the N- and C-terminal domains using residual dipolar coupling [57,58]. The residual structures present in the disordered proteins are implicated in forming initial contact points with structured binding partners [103].

1.1.3.3 Experimental characterisation

Experimental characterisation of IDPs poses challenges to the currently available range of techniques. Obtaining the unique 3D structure of IDPs from crystallography is difficult and would provide very little information, since IDPs are characterised by a multitude of conformations. A few crystallography studies have reported being able

to obtain structures of IDPs in the presence of other proteins [104-106]. The disadvantage of such studies is that the structural mosaicity of IDPs is lost once they are crystallised. Small-angle X-ray scattering (SAXS) [107] is one of the most widely-used methods to study IDPs by providing insights on their dimensions and indirectly on their disorder or unfoldedness [108,109].

NMR is the most widely-used technique to study IDPs [110,111] since it is the only technique that can provide atomic-level information of IDPs in solution [112]. Chemical shift data [113-115] is used to estimate secondary structures in proteins by calculating deviations of NMR parameters from random coil values. The presence of such chemical shifts in IDPs are excellent indicators of secondary structural preferences, since they mostly take up random coil conformations [24,116].

Paramagnetic relaxation enhancement is also another powerful technique that is applied for the study of disordered proteins. PRE provides information on transient tertiary organisation of the unfolded and partially folded proteins. This information is used to understand r^{-6} relationship within the proteins and has been successfully applied to study many IDPs [117].

Residual dipolar coupling are used to refine and narrow down the large structural ensemble of disordered proteins [58,118,119]. Heteronuclear NMR is a more comprehensive method to understand conformational conversion but is challenging due to its large protein demands [112,120].

FRET reports energy transfer between the donor and acceptor chromophores and this transfer efficiency is proportional to the distance between them. This gives a measure of the average distance between two parts of a protein that have been tagged with chromophores. This can be used as a molecular ruler to measure if and how much they are folded. Single molecule fluorescence resonance energy transfer (SM-FRET) and FRET have also been very useful in understanding the conformational dynamics of such disordered proteins [121-124].

A combination of such techniques are applied together to provide a better understanding of IDPs [125]. IDPs have also been characterised using other methods such as differential scanning micro-calorimetry [126], Bayesian statistics [127] and single molecule fluorescence [128].

1.2 Folded state

The folded state is well studied and characterised by various classical structural studies. The important role of the folded state is because most protein functions, described today, are carried out by folded proteins. Proteins fold into specific structures that carry out different functions. Protein function depends on the cooperation between positions in the protein structure. The notions of protein being rigid structures were dispensed long ago. Protein function is a property of various contributions from various local and long-range interactions between different parts of a protein. Although the functional segment might form a relatively small part of the protein, the rest of the structure is also crucial and indispensable. Proteins communicate between different sites located close or distant to each other. Close communication is brought about by steric forces, hydrophobic interactions and hydrogen bonding, while mechanisms of distant contacts are harder to perceive from its structure. Allosteric regulation is the ability of a protein to modulate its activity based on conformational signals from distinct sites in the protein [129,130]. Allosteric property is the essential modulating factor in most of a protein's functional mechanisms. Modulation is an important part of protein response [131]. Modulation typically occurs when a protein interacts with another protein or a ligand resulting in change(s) to its function or kinetics. This is brought about by conformational changes that occur in response to binding and is called allostery. Allostery is well characterised in multi-domain proteins [132,133]. Single domain proteins exhibit allostery through conformational change but its mechanisms are not clear [134]. Several questions remain unanswered. Binding sites have been characterised in terms

of interactions, but energetic requirement for binding sites are not clear. Transfer of information within a protein by conformational changes has also been challenged by allostery without conformational changes [135]. A fundamental mechanism that can be applied to allosteric systems is missing, and the possibility of such a common mechanism is also in doubt.

Conformational change is no longer a requirement for allosteric communication between the different sites in the protein [135]. Thermal fluctuations from ligand binding are shown to generate several kJ mol^{-1} of energy. The absence of conformational change in allostery further tests the current methods to understand and dissect protein modulation. However, the common criterion in any sort of modulation is the perturbation in energy. This energy is used to transfer information across different functional and scaffold sites in proteins and help the protein communicate within itself [136,137].

The most popular methods that study such energetic interactions are based on conservation of amino acids at energetically important positions [138,139]. Perturbation of such positions provides information on their importance. The approach and results are challenged by double mutant cycle analysis [140]. These approaches also have been shown to contain phylogenetic noise that interferes with the identification of such interacting networks [139].

1.3 The sampling problem

The sampling problem in protein folding is the accurate prediction of the structure of a protein from its sequence. It is one of the most important unsolved problems in science [141-143]. Levinthal's paradox states the practical impossibility of brute force examination of all the possible states [144]. The biggest problem is the huge protein landscape that makes it very difficult for thorough sampling. This has necessitated the development of faster and more intuitive methods for conformational searching.

1.3.1 Approaches to structure sampling

Numerous tools to sample the conformational space have been developed. There are different approaches that have been considered for searching through sample space. Hierarchical approach begins with an extended chain and randomly changesing the conformation of residues, thus increasingly working with longer stretches [145]. The two popular approaches to sampling are Monte Carlo (MC) and molecular dynamics (MD). Monte Carlo is a sampling approach, where every change made is accepted or rejected based on a set of probabilities. The change is always accepted if it is beneficial. The change can be accepted with a given Boltzmann probability distribution, even if it is not beneficial. The change is evaluated based on the change in energy caused by it. Any change is generally considered beneficial if it reduces the energy. The Boltzmann distribution allows the sampling to evade local minima in the energy landscape. ROSETTA is one of the most successful MC sampling methods which incorporate this MC approach, but changes are made with respect to fragments rather than individual residues [146]. Fragments are short polypeptide samples taken from the known structures in PDB. While MC depends on probabilities, MD samples are based on molecular mechanics and force-fields by the numerical integration of Newton's laws of motion. These numerical integrations are carried out from a set of initial and boundary conditions. Verlet algorithm is a popular numerical approach to calculate trajectories by breaking down the simulation to a discrete set of time steps [147]. As the time steps are reduced, they give rise to more accurate models and lead to higher computational complexity. Simulating a considerably large protein is practical if the solvent is not considered. Solvents play an important part in determining the protein space [148,149]. Implicit solvent models with a bulk term perform faster at a considerable reduction in accuracy [150-152]. Explicit solvents provide excellent accuracy since every solvent molecule is included in the simulation [153].

1.3.2 Trajectory Directed Ensemble Sampling

Trajectory Directed Ensemble Sampling (TraDES) is an *ab initio* structure prediction tool developed initially for structure prediction [154,155]. TraDES builds all-atom samples of sterically plausible protein structures in excluded volume space by a self-avoiding random walk [156,157]. It generates protein structure from the N- to C-terminus by adding one residue at a time. There is a backtracking algorithm which will backtrack and resample residues in case of atomic collisions in the generated chain. The approach is based on sampling dihedral probabilities of amino acids from dictionaries obtained from the PDB. This gives rise to sterically probable structures with dihedral distributions similar to those observed in PDB. The backbone is first generated and the rotamers are assigned from the Dunbrack rotamer library [158]. The N-terminal C-α atom is placed at the origin in 3D coordinates (0,0,0). Every residue has an independent distribution of dihedrals that is unique to itself. The second C-α atom is placed by randomly selecting dihedrals from the distribution. The bond lengths are also sampled from high-resolution PDB structures. It progressively places C-α atom for every residue and adds the backbone atoms subsequently followed by the side-chains. Initially, a trajectory distribution is generated which contains all the probability distribution functions (PDF) for every residue of the protein chain. The PDFs are chosen based on different sampling options, such as standard, extended, random-coil, etc. It is possible to generate a trajectory distribution for sampling different residues in different conformations.

2. Chemically Denatured Proteins and Residual Structures

2.1 Introduction

Most of the structural focus is on the functional states of proteins, while the chemically denatured landscape is still largely unexplored. The unfolded state is considered a useful tool to study protein folding. The chemically denatured state together with the disordered state forms the unfolded realm of a protein. All such unfolded states are dynamic and cannot be represented using a single structure. So, the experimental data obtained are spatially averaged. The denatured state is an important experimental starting point for the protein folding process. SAXS studies indicate that the dimensions of a protein (e.g. radius of gyration) during early refolding are very similar to its denatured state [159]. This denatured state is an ensemble of compact to expanded structures [160] referred to as the denatured state ensemble (DSE) [161]. The denatured ensemble contains partially structured states which have implications in protein folding [162]. These states are recalcitrant to conventional techniques that study folding [163] since these only look at the overall dimensions of the protein. DSE structures are sequence-dependent and single mutations can cause significant differences in them [164,165].

The denatured state maintains some structural integrity, unique to its sequence and denaturing conditions. These residual structures are not uniformly distributed along the sequence and can influence folding thermodynamics [166]. Residual structures in proteins have been studied by small-angle X-ray scattering [167], NMR [168,169], single molecular FRET [170], circular dichroism spectroscopy [171] and hydrogen/deuterium exchange [172]. The alpha-helical content decreases while beta and polyproline conformations dominate the denatured state [171,173]. NMR studies have shown that urea- and Gd-HCl-induced structures correlate well with simulations

of partial beta and polyproline II sampling [174]. But, it is unlikely that beta and polyproline II conformations alone can represent the denatured state of the proteins [175]. The residual structures and their effects on protein refolding have not been understood completely. The denatured state of a protein behaves like a random polymer and its Rgyr has a power-law relationship with its residue length [2,176] [177]. Ensemble sampling is one of the best methods to study such conformers that do not follow any ordered structural pattern. Ensemble methods [177] reproduce the random-polymer behaviour of denatured proteins originally established by SAXS [176].

Different types of residual structures (native-like structure, intermediate structures and off-pathway intermediates) populate the DSE of a protein. Native-like structures are observed in the unfolded state [26,178] including transient tertiary structures [33]. These can act as nucleating structures during protein folding. Residual dipolar coupling shows long-range interactions and native-like spatial positioning in harsh denaturing environments, such as 8 M urea [35]. The denatured state has a local conformational bias towards native-like structures [179], and the presence of precursors for transition states have been detected using RDC [180]. The off-pathway intermediates usually occur with incomplete denaturing conditions (up to 3 M urea), while native-like structures are often found in completely denaturing conditions (6 M urea) [181]. The DSE can be made completely devoid of any structure by adding excessive amounts of denaturants (8 M urea or 6 M Gd-HCl). Secondary structures form ahead of tertiary structures and are present even in the absence of tertiary interactions [182]. A completely random model of protein structure formation from unstructured state has the disadvantage of forming secondary and tertiary interactions simultaneously. Formation of tertiary structure from incomplete local secondary structure can lead to wrong contacts and the protein getting stuck in local energy minima. The denatured state follows global random-coil behaviour [176] and still

contains native-like structures [183]. The denatured state is difficult to study using experiments due to their structural diversity and inherent randomness. Some disagreements also exist between results from popular methods like SAXS and single molecule FRET [184]. Denatured state has been simulated using molecular dynamics [185] but the ability of MD simulations to capture diverse conformations of the proteins in different solvents is not convincing. There is still a huge void in atomistic experimental data for the denatured state due to the heterogeneity of the denatured ensemble.

Trajectory Directed Ensemble Sampling (TraDES) [155] is part of the ENSEMBLE package used to fit structures to NMR data for intrinsically disordered proteins with specific spatial constraints [186]. TraDES is a tool for fast probabilistic sampling of protein conformations generated in continuous three-dimensional space with realistic bond lengths, angles and dihedrals as observed in native structures [154]. Large numbers of conformers are generated in real conformational space by a Monte Carlo all-atom, off-lattice build-up. The sampling constraints used here are based on previous studies which show increased beta and polyproline II conformations in denatured proteins [174,177,187]. Ensembles of 23 proteins, previously studied using SAXS [176,177], were simulated using TraDES by varying dihedral angle populations to best match their denatured SAXS Rgyr. Using the optimised parameters, a consistent ensemble model was generated to accurately predict the Rgyr of denatured proteins. This model is taken as a representation of the denatured state and further analysed for residual structures. Fifteen proteins whose native structures are known were examined for residual native-like conformations in their denatured state. Three additional proteins with experimental residual structural data were used to test the efficiency of the model. The residual native-like elements contain diverse secondary structures and a few short motifs. Structures like the β-turn also appear to be important.

2.2 Methods

2.2.1 Denatured state ensemble

The 'TraDES-2' package was used to generate the DSE of 23 proteins. TraDES contains conformational definitions in Ramachandran space (Φ and Ψ) for all amino acids in various conformations, generated from a non-redundant dataset of 7030 structures in the PDB database, as of June 2012 [188]. These contain conformational probabilities for every amino acid in different secondary structures and can be optimised to sample the denatured state. TraDES can use Garnier-Osguthorpe-Robson (GOR) [189,190], one-state, three-state secondary structure, uniform and standard structural constraints to generate structures. These are different approaches to assign secondary structures for amino-acids in the sequence. GOR is an information theory method to predict secondary structure of amino-acids in proteins. This uses probability parameters calculated from solved PDB structures. This also considers neighbour effects by using conditional probabilities to determine the secondary structure of residues by including information about its adjacent residues. One-state assigns the residue to take up either coil or extended or helical secondary structure based on user input. Three-state offers more flexibility where a ratio of these three states (coil : extended : helical) can be assigned instead of one. The ratio of extended (E) and coil (C) conformations was systematically modified (0%E+100%C to 100%E+0%C) using the three-state constraint and tested to match SAXS Rgyr data. The sampling ratios were modified in 5% increments. These different ratios were used in the program `seq2trj` to create trajectory distributions for each of the 23 proteins. The trajectory distributions contain the respective dihedral probabilities for a specific sampling ratio (e.g. 75%E+25%C). The program `trades` was used to generate an all-atom structural ensemble for each protein and sampling ratio. The TraDES-2 package executables and source code can be obtained from http://trades.blueprint.org. TraDES generates structures by self-avoiding walk of

proteins in excluded volume space. This self-avoiding walk represents the unfolded protein model and follows scaling law statistics [191,192].

5000 all-atom structures were generated for the 23 proteins and analysed using the TraDES-R-Analysis Package (TRAP) (http://trades.blueprint.org). The `trades` program outputs a log file with structural, geometric parameters and calculated statistical energy for all the structures. The log files can be parsed and ensemble properties such as Rgyr, N-C distance and energy can be calculated using TRAP. The Rgyr values for the 23 proteins from the TraDES ensembles are compared with SAXS Rgyr to test the accuracy of the sampling ratio. The Rgyr values are not normally distributed and skewed due to the constrained dihedral sampling. The Rgyr distributions of different proteins differ in the degree and direction of skew-ness due to varied amino acid compositions. So, the mean value is not a proper representation of the Rgyr values and the peak value was taken as the Rgyr value of the ensemble. The SAXS Rgyr values were used to compare the performance of the sampling ratios using two calculated measures of similarity and the best fit ensemble is chosen. The Average Absolute Difference (AAD) is the absolute deviation of the TraDES Rgyr from the SAXS Rgyr averaged over 23 proteins. The Average Ratio (AR) is the absolute deviation of the ratios of TraDES Rgyr and SAXS Rgyr from the ideal value of 1. The AR is also averaged over the set of 23 proteins for a given sampling ratio (%E+%C). The denatured state conformational space is understood by analysing the Ramachandran map of the resulting ensemble.

Protein	**Length**	**SAXS Rgyr (Å)**	**TraDES Rgyr(Å)**
GroEL	549	82	75.22
Phosphogycerate kinase (PGK)	416	71	65.27
Tryptophan synthase α subunit (TSA)	268	48.8	49.58
Carbonic anhydrase	260	59	51.07
Outer surface protein A	257	49.3	48.35
Apomyoglobin	154	40	38.97
Staphylococcal nuclease	149	37.2	37.63
Lysozyme	129	35.8	35.44
CheY	129	38	35.44
Ribonuclease A	124	33.2	37.49
mACP	98	30.4	29.33
ctACP	98	30.5	30.78
Protein L	79	26	28.21
Ubiquitin	76	25.2	26.93
drK	59	21.9	23.84
Protein G	52	23	23.14
Cytochrome C	39	18.4	18.45
AK-37	37	16.9	17.09
AK-32	32	14.5	16.96
AK-27	27	12.8	15.40
AK-16	16	9.8	11.49
Angiotensin	8	9.1	7.28

Table 2.1 Comparing predicted and experimental Rgyrs of chemically denatured proteins (CDPs). The set of 23 proteins for which Rgyr and power-law relationship is calculated. Predicted Rgyr is calculated using TraDES specific sampling (50%E+50%C sampling with ensemble size of 5000 structures) and experimental Rgyr values were determined by SAXS.

2.2.2 Flory's relationship

The Rgyr from TraDES for the 23 proteins were fit to a power-law equation (Equation 2) using MATLAB and its Flory parameters [2], υ and R_0, were calculated. Unlike the previous study [176], the parameters were calculated without considering confidence intervals. So, as a standard comparison, the υ and R_0 for earlier studies [176,177] were similarly recalculated.

$$R_g = R_o N^{\nu} \quad \text{Equation 2}$$

2.2.3 Conformational analysis

TraDES `RamangL` program was used to calculate the dihedral angles of the ensemble that best fit the SAXS Rgyr data. The torsion angles were plotted on four different Ramachandran maps: all amino acids, all amino acids except proline and glycine, only proline and only glycine. The maps were plotted using `TRADES.RamaPlot()` function in TRAP. Similarly, all-coil and all-extended Ramachandran maps were also generated. For native-like residual structure map, the phi-psi values of positions in the ensemble that match the respective native conformations were also plotted.

2.2.4 Native-like structure analysis

Using the optimised sampling parameters, larger DSEs of 30,000 structures were generated. Fifteen proteins with medium residue lengths (39 to 260) were chosen from Table 1 for the analysis. For experimental validation, DSEs were generated for three additional proteins, "apoflavodoxin", "hUBF HMB Box 1" protein and "outer membrane protein X", whose residual structures have been experimentally determined. Every structure in the DSE was compared to the crystal structure and the dihedral similarities of individual residues were noted. Dihedral angle comparison was used to calculate native-like similarity. A match was recorded if a residue's Φ and Ψ values are both within 2° of its crystal structure Φ and Ψ values and every match was assigned a score of 1. Stretches with above-threshold consecutive positional matches are considered to contain native-like structures in the DSE.

Different proteins have varying amounts of residual structure and setting a global match threshold is not possible. For most proteins, 4 matches per position were set as a threshold to be considered a part of the consecutive stretch. The cumulative scores of consecutive positions were used to identify native-like stretches. If the number of matches for a position is below the threshold value, the cumulative score was reset to 0. The native-like residual positions were compared to the crystal structure and their secondary structures in the folded state are reported using DSSP [193,194] were noted. The program `get_residual.pl` in the TraDES package was used to carry out the dihedral comparisons.

2.3 Results

2.3.1 TraDES predicts denatured Rgyr

Rgyr of ensembles with specific 50%E+50%C sampling closely resemble the SAXS Rgyr (Table 2.1). The TraDES Rgyr values correlate well with SAXS Rgyr with a linear fit R^2 value of 0.9828 (Figure 2.1). However, the R^2 values do not provide a good comparison of the different sampling ratios. The Average Absolute Difference (AAD) (Equation 3) and Average Ratio (AR) (Equation 4) were used to evaluate the performance of different sampling ratios against SAXS Rgyr data (Figure 2.2) and 50%E+50%C ensembles have the minimal AAD and AR values of 2.38 Å and 1.00577 respectively. AAD is the average difference between the SAXS Rgyr and TraDES Rgyr per protein. An AAD value of 2.38 Å is close to the mean SAXS error of 2.11 Å, based on the standard deviation of the Rgyr values obtained by SAXS, and suggests a good agreement. This 50%E+50%C sampling models the SAXS data of denatured proteins accurately and was used for further residual structure analysis.

$$Average\ Absolute\ Difference\ (AAD)_{x\%E+y\%RC}$$

$$= \frac{1}{N}\sum_{i=1,N}\left(\left|Rg_{i,(xy)}^{TraDES} - Rg_{i,(xy)}^{SAXS}\right|\right) \qquad \text{Equation 3}$$

$$Average\ Ratio\ (AR)_{x\%E+y\%RC} \qquad \text{Equation 4}$$

$$= \left(\left|1 - \frac{1}{N}\sum_{i=1,N}\frac{Rg_{i,(xy)}^{TraDES}}{Rg_{i,(xy)}^{SAXS}}\right|\right)$$

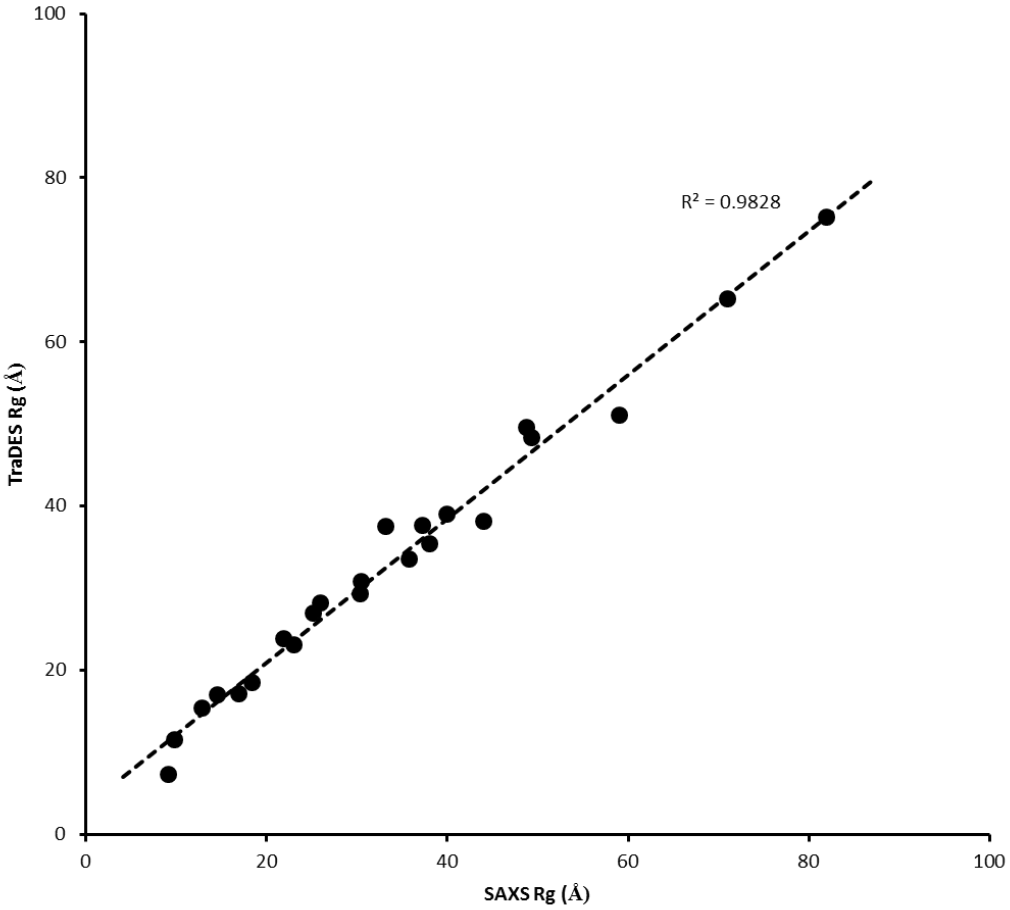

Figure 2.1 TraDES Rgyr (50%E+50%C) vs SAXS Rgyr. The dotted line is the line of linear-best fit of the peak Rgyr values (black solid circles) with an R^2 fit of 0.9828. Rgyr values are the mean values obtained from an ensemble of 5000 structures.

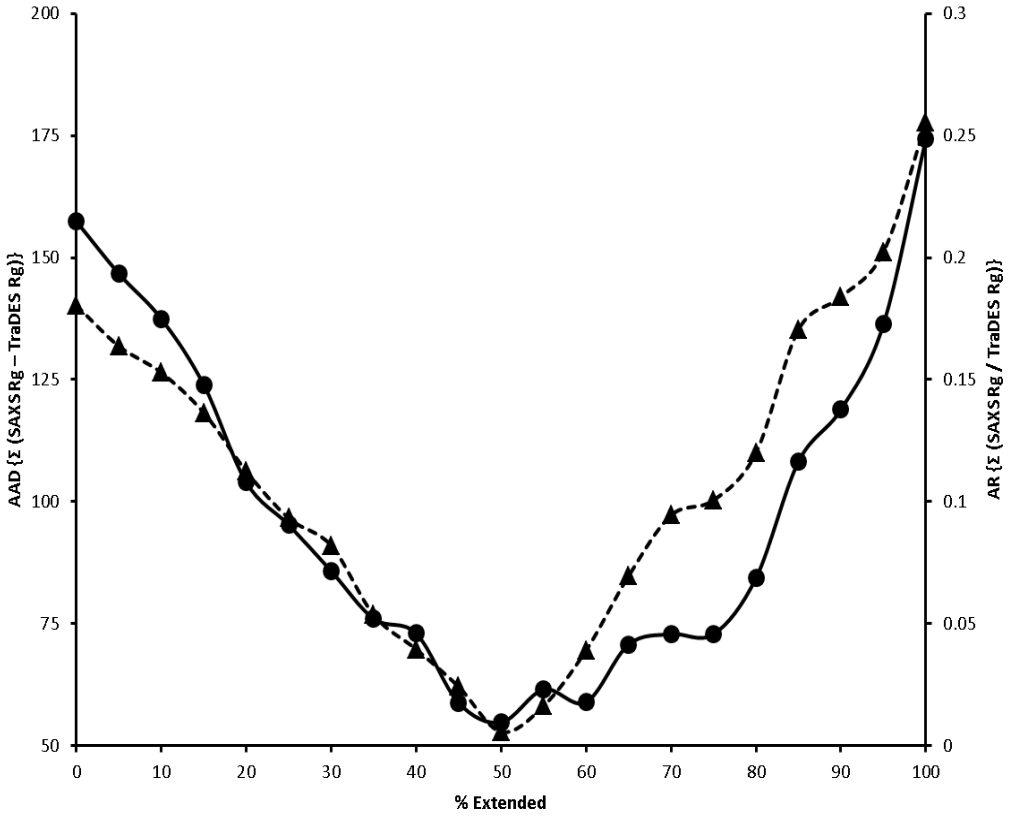

Figure 2.2 Comparison of different sampling ratios. The Average Absolute Difference (AAD) (▲) and Average Ratio (AR) (●) values are plotted for different compositions. The 50%E+50%C ratio gives the lowest values for both the parameters and agrees closely with the SAXS Rgyr. The AAD and AR values are calculated from an ensemble of 5000 structures.

2.3.2 Flory's relationship

Denatured Rgyr values have a power-law relationship with the length (N) of the protein [176,177]. The υ values for TraDES and SAXS Rgyrs are 0.5365 and 0.5677 respectively. The power-law relationship (Equation 2) shows good agreement between the length and the Rgyr of the denatured proteins. The Flory equation for TraDES Rgyr has the parametric values of 1.0523 for R_0 and 0.5328 for υ.

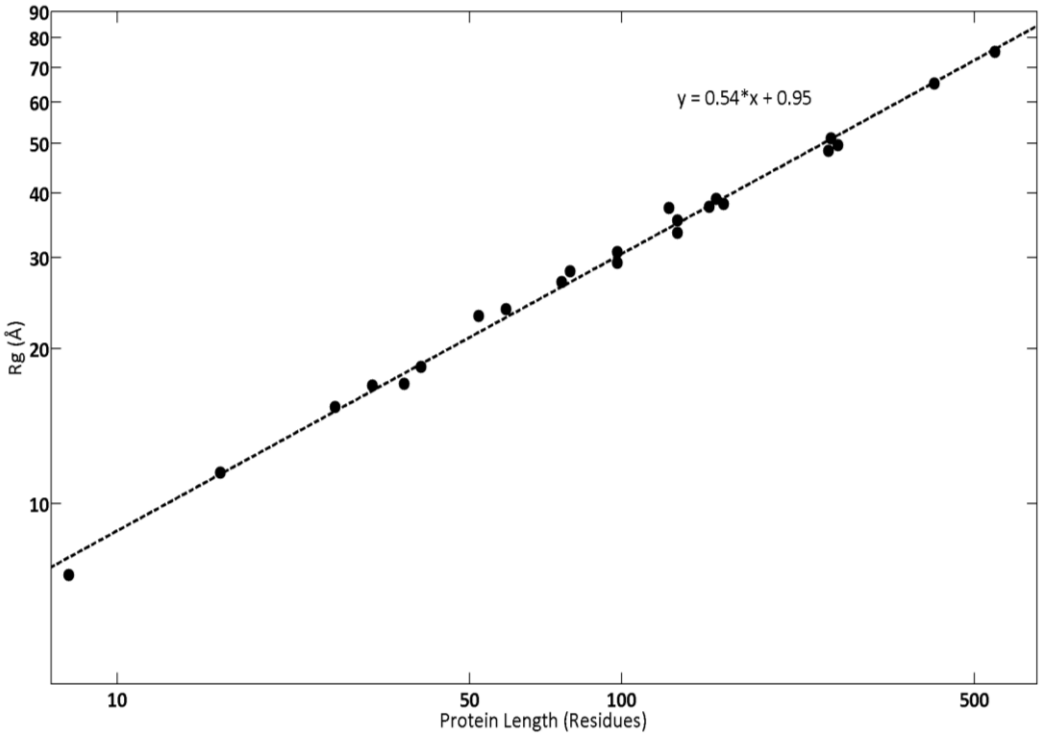

Figure 2.3 Flory's power-law Relationship between Rgyr and residue length. The dotted line is the linear best-fit line with slope 0.5328 (υ) and intercept 1.0523 (R_0) for TraDES Rgyr against residue lengths for CDPs.

2.3.3 Effect of sampling on conformational space

Different sampling ratios have an effect on the conformational space. In the protein carbonic anhydrase, the 50%E+50%C has more regions in the beta, extended and polyproline type II helical regions than a random coil model (Figure 2.4). The 100%C sampling gives a more distributed phi and psi space with peaks in extended and helical regions while 100%E sampling gives a Ramachandran space completely biased towards the beta and polyproline type II helix conformations. Extended sampling (100%E) overestimates the Rgyr while the coil sampling underestimates it. The 50%E+50%C model provides a good Rgyr model in agreement with experimental data (Figure 2.5). The linear fits of 100%E, 100%C and 50%E+50%C have slopes of 1.21, 0.7 and 0.88 respectively. The 50%E+50%C model is closest to the ideal slope of 1 and comparable to a slope of 0.89 from an earlier study [177].

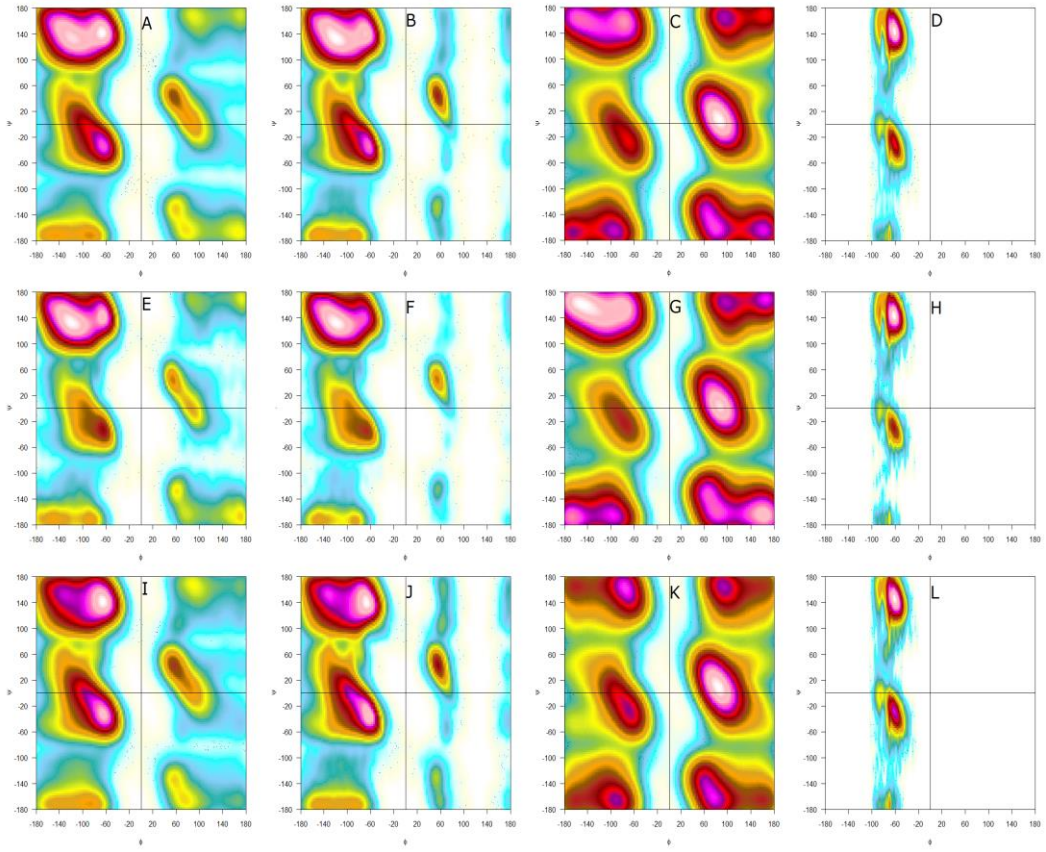

Figure 2.4 Effect of sampling on conformational space for carbonic anhydrase. The Ramachandran maps in four columns represent: all amino acids, all amino acids except proline and glycine, only glycine and only proline, respectively. The first (A, B, C, D), second (E, F, G, H) and third (I, J, K, L) rows correspond to 50%E+50%C, 100%E and 100%C, respectively. Ensembles of 5000 structures were used for each sampling ratio.

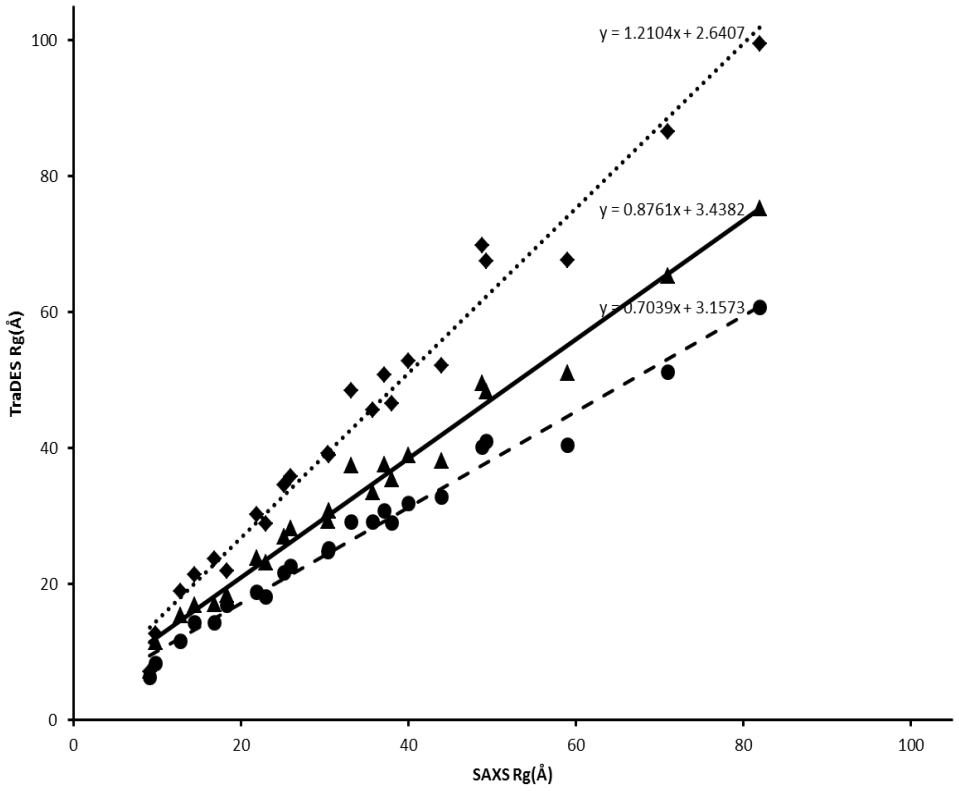

Figure 2.5 Effect of sampling ratios on Rgyr. 100%E (◆), 50%E+50%C (▲) and 100% RC (●) samplings are compared with SAXS Rgyr. The 100%E and 100%C sampling over- and underestimate the Rgyr while 50%E+50%C is able to match the denatured Rgyr values closely with ensembles of 5000 structures. The slopes of the linear fit are 1.21, 0.88 and 0.7 for 100%E, 50%E+50%C and 100%C respectively.

2.3.4 Native-like structures in DSE

Native-like structures are consecutive individual positions in the denatured ensemble which resemble the folded conformational state. All such stretches in a set of eighteen proteins have been identified (Table 2.2). Residual stretches differ in the amount of phi-psi matches, length and secondary structure. The amount of residual structure also varies between different proteins. This is calculated as "*residual structure %*" (Table 2.2), and is the percentage of total matches to the total number of structures (30,000). The *"% residues in residual structure*" is the fraction of residues that maintain native-like structure in the DSE (Table 2.2). These are relative measures to compare different proteins and do not represent any absolute experimental quantity.

Protein	Crystal Structure PDB Code	Residues with Residual Structure	Secondary Structures of the Residual Stretches	Residual Structure % (x100)	% Residues in Residual Structure
Carbonic anhydrase (CA)	1V9E[195]	37-48 58-63 84-101 112-124 & 140-149 151-156 & 178-187 210-225 227-239 244-250	Beta + Turn + Beta Beta + Bend Beta + Bend + Turn H-Bonded Beta Ladder 3/10 Helix Bend + Beta + Helix Bend + Bridge + Turn No Structure	16.17	42.69
Chemotactic protein Y (CheY)	1DJM[196]	21-27 53-57 & 84-89 93-102 & 111-121	Helix H-Bonded Beta Ladder Alpha Helix	10.17	30.23
Common-type acyl-phosphatase (ctACP)	2ACY[197]	5-14 22-33 & 54-67 73-97	Beta Whole Helix Beta + Bend + beta	12.91	62.24
Cytochrome C (cytC)	1V54[198]	14-22 28-38	Whole Helix Bend + Helix	8.62	51.28
Dihydrofolate reductase (DHFR)	1DDS[199] 3DRC[200]	2-10 32-43 & 44-62 105-115 140-159	Whole Beta Helix Turn Beta Whole Beta Turn + Beta	19.58	44.38
Drk-SH3 domain (drk)	2A36[201]	24-26 36-41 & 44-55	Beta H-Bonded Beta Ladder	10.95	33.9
Lysozyme (Lys)	1LSG[202] 3B6L[203]	48-56 114-118	Turn + Beta + Turn Turn	7.39	10.85
Muscle acyl-phosphatase (mACP)	1APS[204]	6-15 28-33 38-41	Whole Beta Helix Beta	9.16	20.41
Myoglobin (myo)	3RGK[205]	3-20 & 22-43 48-79 100-119 & 125-149	Whole Helix + 3/10 Helix Two whole Helix Whole Helix	14.39	75.97
Outer surface protein A (ospA)	1OSP[206]	44-65 117-131 133-151 166-188 237-255	Beta + Turn + Beta Helix + Turn + Beta Beta + Bend + Beta Turn + Beta + Helix + Turn Beta + Turn + Beta	18.16	38.13
Protein G (ptnG)	3GB1[207]	3-8 & 10-20 21-41 43-56	H-Bonded Beta Ladder Whole Helix Beta + Turn + Beta	22.00	88.14
Protein L (ptnL)	2PTL[208]	31-35 41-53 62-66 71-76	Beta Whole Helix Beta + Bridge Whole Beta	5.97	36.71

Ribonuclease (Rnase)	1RBB[209]	2-13 41-47 72-87 95-113	Whole Helix Whole Beta Beta + Turn + Whole Beta	12.44	20.16
Staphylococcal nuclease (Snase)	1EY0[210]	7-14 16-37 39-44 67-116 120-137	Beta Beta + Turn + Beta Bend Beta + Turn +Beta + Turn + Helix Helix	16.32	69.8
Ubiquitin (Ubi)	1YX5[211]	11-32 35-45 65-74	Beta + Turn + Helix 3/10 Helix + Beta Beta	17.49	56.58
Apoflavodoxin	1YOB[212]	5-9 11-21 29-35 37-53 61-67 83-101 108-125 127-154 157-171	Beta Bend + Helix No Structure Helix + Turn + Beta Bend + 3/10 Helix Turn + Beta + Turn Helix + Turn + Beta Turn + Bridge + Bend + Beta 3/10 Helix + Helix	14.42	70.95
Outer membrane protein X (ompX)	1QJ9[213]	24-50 & 65-83 98-130 135-147	Beta + Turn + Beta Beta + Turn + Beta Whole Beta	30.93	62.16
hUBF HMG Box 1 (HUBF)	1K99[214]	15-20 & 23-32 41-50 & 59-67	Helix + Turn Helix	46.24	35.35

Table 2.2 Identified residual structures from the DSEs. This table shows the proteins and the residual structures identified in the TraDES DSEs. This shows the PDB crystal structure used for phi-psi comparison along with the positions and the secondary structures that they form in the native structure. The secondary structures are notations from DSSP.

As an example, we look at the RNase residual plot which shows four regions of different secondary structures (Figure 2.6). The two main stretches are the β-strands that give high matches to the crystal structure. The denatured RNase protein contains these conformational stretches populated in the DSE and these short secondary structures may have an implication on the refolding of the protein (Figure 2.7).

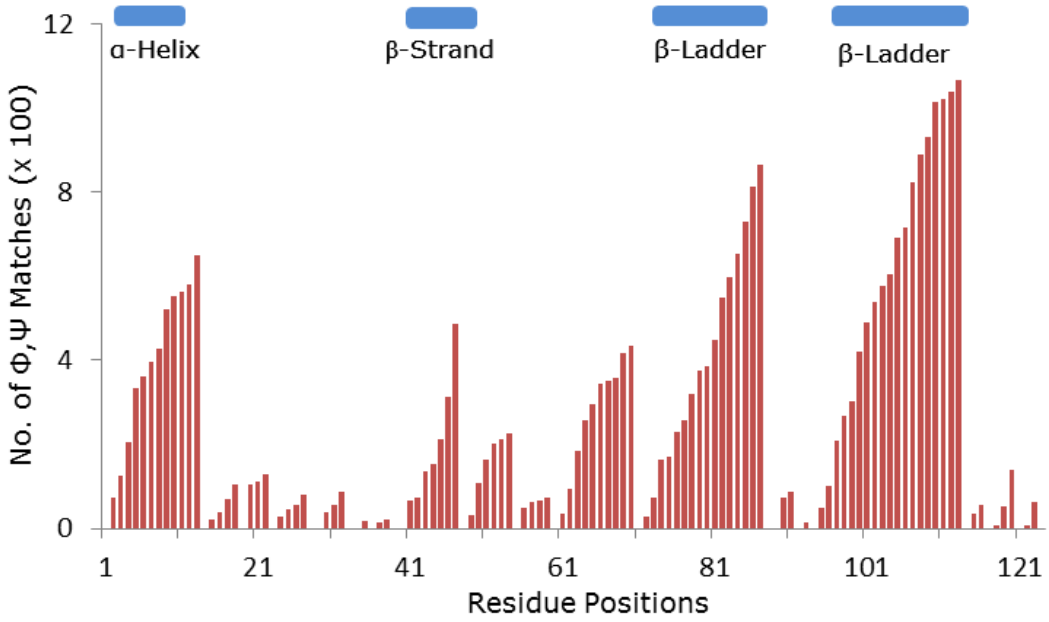

Figure 2.6 Plot of native-like matches for different positions in RNase. Each bar is the cumulative score of discrete phi-psi matches in contiguous positions in an ensemble of 30,000 structures (refer Methods). A helix and four β-strands in the ensemble resemble and retain the native-like geometry in the denatured state. Blue stretches have low propensities (cumulative match scores of 100 to 500), while red stretches have high propensities (scores > 750).

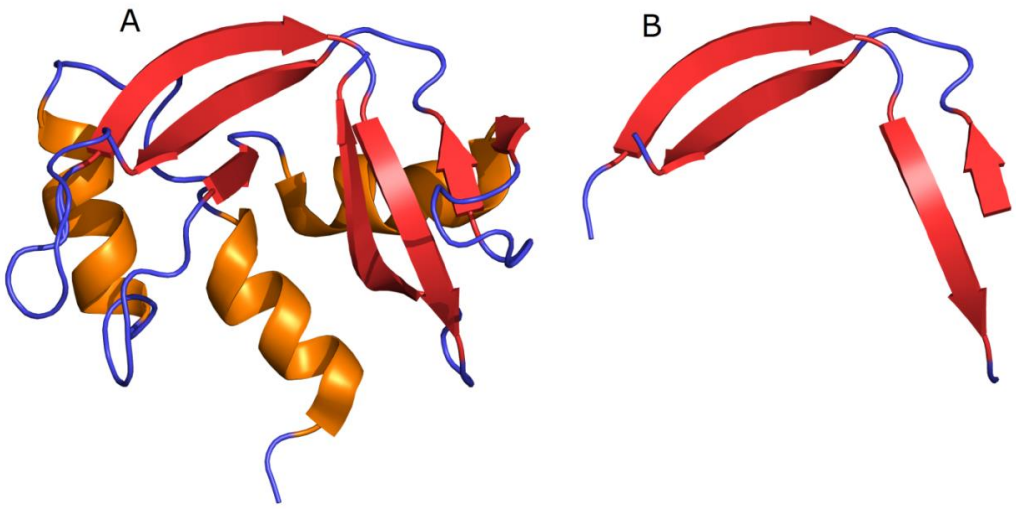

Figure 2.7 Residual structure of Rnase mapped to its structure. The high propensity residual structures of RNase (B) inferred from Fig. 2.6, compared to the complete structure of the RNase protein (A). There is a perpendicular symmetry to the whole protein with two set of strands and helices running perpendicular to each other. The residual structure seems to contain four strands that can act as starting structures for the rest to fold into a native tertiary structure.

These structures form two anti-parallel β-ladders that are perpendicular to each other. The RNase protein has an L-shaped symmetric structure with a groove in between. It is quite plausible that the formation of the β-strands and the perpendicular topology can help act as the starting point or nucleus for the other secondary structures to fold around it.

2.4 Discussion

TraDES computationally predicts the denatured Rgyr values of 23 proteins accurately (Figure 2.1). Rgyr calculation is highly time-efficient and takes approximately 6 hours for 23 proteins, each with an ensemble size of 5000 structures. Accurately computing denatured Rgyr values of proteins with time efficiency is advantageous. The representative all-atom structural ensemble is also available for detailed analysis on the denatured state.

Denatured state models are easily verified without any experimental data by their self-consistency. The chemically denatured proteins follow random polymer behaviour with a power-law relationship between its size and its residue length. The slope (υ) of the power-law fit for an ideal random polymer in a "good" solvent has been calculated to be 0.5. For proteins in non-ideal solvent, υ is expected to be 0.587 to 0.589 [215]. The υ values for the 23 proteins from earlier studies are 0.5677 [176] and 0.5322 [177]. The TraDES υ value is 0.5365 and is closer to ideal Flory than SAXS υ value. The ideal value of 0.588 is calculated for a real polymer with non-zero thickness and non-trivial interactions between monomers. The SAXS Rgyrs are calculated in different denaturing conditions and result in higher υ value since different denaturing conditions destabilise proteins differently [216,217]. TraDES uses a constant condition for all proteins and this uniformity provides a υ value closer to ideal υ value than SAXS and is also true for other such ensemble studies [177]. The υ value also depends on uniform randomness, which is absent due to the different amino acid compositions in the group of proteins. Every amino acid has its own

propensity for various secondary structures and these structures are stabilised or destabilised differently in denaturing environments. Proteins also contain varying amount of residual structures (Table 2.2) and this also causes deviation from ideal polymer behaviour. The presence of varied residual structures and compositions prevents TraDES to attain ideal υ of 0.5. SAXS in addition has different denaturing conditions as well. So, TraDES takes up a υ value to 0.5365, between the SAXS value of 0.5677 and ideal value of 0.5.

Over- and underestimation of Rgyr by 100%E and 100%C sampling is an indirect evidence for residual structure in the denatured state. In the absence of any residual structure, the proteins would be fully extended with higher Rgyr values similar to those predicted by 100%E (Figure 2.5). The denaturants bind to the protein backbone and extend the polypeptide [173]. The coil model has considerable structure, similar to intrinsically disordered proteins, and gives lower Rgyr values. The 50%E+50%C sampling accounts for the extended protein backbone by 50%E sampling and also considers residual structures by providing 50% coil sampling. This sampling ratio which best matches the SAXS Rgyr values is different from a previous study [177]. The difference could be due to the different tools (TraDES and Flexible-Meccano (FM) [125]), since individual conformational libraries are different. The ratios that TraDES and FM describe are different and not directly comparable. TraDES and FM have different libraries for coil and extended conformations and it is not straightforward to derive actual relative ratios for comparison. TraDES uses all-atom structures compared to the spherical side-chain volume exclusion model used in FM. However, in both cases, the ratios are chosen with respect to their abilities to match experimental data. So, the sampling ratios are irrelevant as long as both are able to match experimental data accurately. The conformational space also agrees with the NMR determined space of ubiquitin denatured by urea [173]. This 50%E+50%C model is able to follow the power-law relationship, accurately calculate the Rgyr of

23 denatured proteins and provides 3D ensembles for native-like structure analysis.

The denatured conformational space (Figure 2.8B) is very similar to the original starting trajectory distribution (Figure 2.8A) and is a measure of the sampling accuracy of TraDES. The native-like conformations (Figure 2.8C) are observed around the crystal structure dihedrals (black dots in Fig 2.8C) as small distributions and this spread is dictated by the 2° phi-psi threshold. There are some points in the crystal structure that do not have any residual structures in the ensembles which is due to their ensemble size (30,000). If the sample size is increased, the ensemble should sample those conformations as well. An increase in sample size would not affect the residual structures that are identified since the threshold would also be increased to match sample size. Some residues in the crystal structures are sampled more than others and this quantity is the number of dihedral matches. This consecutive cumulative match score is used to infer native-like stretches. This is similar to how alpha-helices are predicted. Alpha-helices are stretches of consecutive residues in alpha-helical conformations and a residue having a helical conformation does not make it part of a helix. Similarly, the residual stretches are consecutive residues, each of which has high native-like similarity. The stretches are compared by their cumulative match scores.

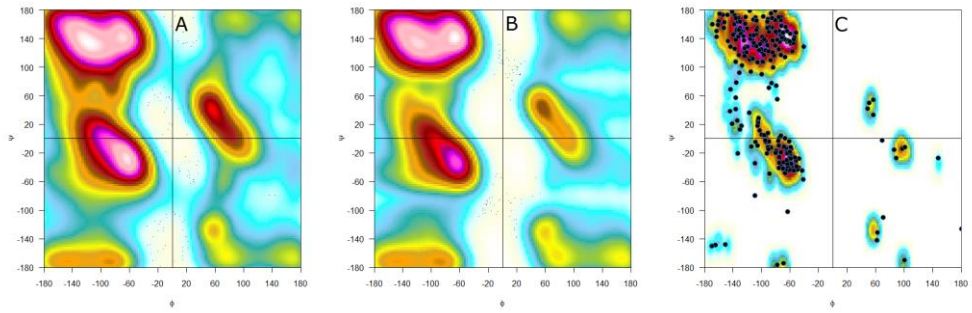

Figure 2.8 Sampling bias does not affect residual structure (carbonic anhydrase). The ensemble 50%E+50%C space (B) is different from the matched ensemble residual structure space (C) and demonstrates that the sampling bias does not directly affect the residual space. The crystal structure of carbonic anhydrase is plotted as black circles in (C). (A) is the starting trajectory distribution (see Methods) used by TraDES to create (B).

Although the DSEs follow global random-coil behaviour, local interactions are still present in individual conformers. A C-α distance matrix for carbonic anhydrase was plotted to demonstrate the structural heterogeneity in the denatured ensemble (Fig 2.9). The diagonals are expected to have high values since the residues are close in sequence. High off-diagonal values represent long-distance interactions and are an indirect measure of the structure present in the conformer. The example structures (Figure 2.9B, C) show close contacts between residues distant in sequence, while the ensemble (Figure 2.9A) shows no such contacts. This is best explained by the averaging method used to improve signal-to-noise ratio in imaging. The signal (diagonal interactions) is more consistent irrespective of the structure while the noise (off-diagonals) is different between structures and averages out when considered as a whole (ensemble). This explains how the DSEs can contain local structures and also maintain global random-coil parameters. The interactions are not completely random and some structures are more common than others. For example, carbonic anhydrase shows two prominent interactions, between residues 40 to 100 and residues 120 to 180. These structures are part of the native-like structures calculated earlier (Table 2).

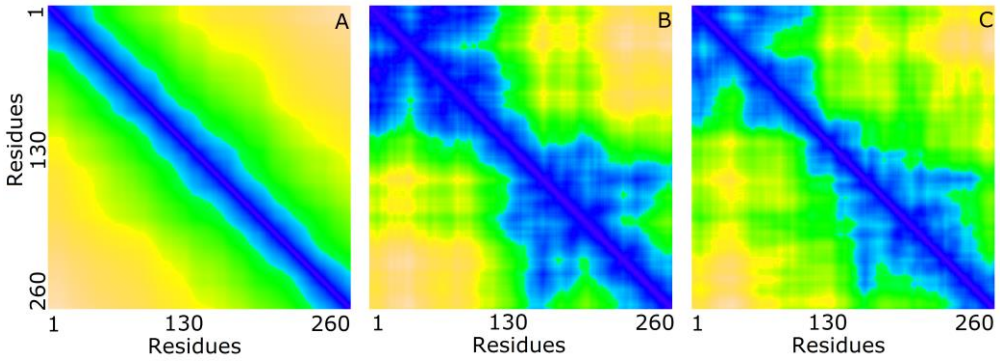

Figure 2.9 Global properties and local structures in carbonic anhydrase distance matrix. C-α distance matrix in DSE. (B) and (C) are the distance matrices for two example structures from the carbonic anhydrase ensemble and (A) is the distance matrix of the average of all the ensemble structures. This shows that the ensemble can show global random-coil properties (A) and still maintain local individual structures (B, C). The distance matrix (A) matrix was generated by averaging 100 structures.

The native-like structures identified in 18 proteins have different compositions. Most of the native-like stretches form important secondary structures in their respective

folded states. Results show that β-strands are more biased than alpha-helices. The denatured state is known to destabilise helices and favour beta and PPII extended states [173]. These matching stretches are not in the same conformer, but are consecutive individual positions that are biased towards the native-like conformations. Sequences that form β-structures in the native fold are energetically more prone to sample diverse states than those sequences that form alpha-helices due to the difficulty of forming β-structures [218]. The presence of native-like beta nucleating structures in the denatured state reduces the probability of those residues taking up other wrong conformations. This could skew the folding energetics in favour of non-switching structures during refolding. Interestingly, there are a lot of β-turns and bends in the residual structure, which can help in the refolding of the protein. A β-turn, once formed, can easily bring together two β-strands to zipper and form hydrogen bonds with each other [219], which otherwise are hard to fold by getting kinetically trapped [67]. Beta turns have been identified earlier as nucleating residual structures in the folding of fatty acid binding protein [168]. Adding β-turns into proteins has been shown to increase folding rates [220]. The importance of β-turns in nucleating β-sheet folding has been well characterised [221-223]. This explains the presence and importance of the turns in the DSEs. Similarly, there are a few short motifs like beta-turn-helix, etc. which might also act as nucleating fragments for protein refolding.

These residual structures have been experimentally identified in a few proteins. Three native-like structures have been identified in the denatured state of Staphylococcal nuclease by NMR [178,224]. The alpha-helix 2 (residues 98-106) and two turns (residues 83 to 86 and 94 to 97) are captured by the TraDES DSE as well. In addition, there are a few other β-strands and turns that are present in the TraDES DSE. In protein G, a turn (residues 9 to 12), α-helix (residues 22-38) and a hairpin (residues 42 to 56) have been shown to contain native-like residual structures [26]. These

match the residues identified by the TraDES ensemble, where the hairpin residues 43 to 50 and 52 to 56 have been identified precisely, along with the helix from 31 to 40 and 13 to 29. Ubiquitin is reported to contain residual structures in residues 1-17 and 23-34 [225]. TraDES predicts residues 3 to 34 (except residue 9) to contain native-like residual structures, agreeing with NMR results. A lysozyme deletion mutant in residues 47 and 48 has been shown to affect its refolding rate [226]. In the TraDES analysis of lysozyme, there are residual β-strands on the either side of those residues (residues 43-46 and 49-57) and it is plausible that the deletion of those two residues 47 and 48 adversely affected the residual β-strands and indicates their importance in refolding. Another point mutation at residue 58 in protein G has been shown to influence the structure and stability of the denatured state [165]. This mutation is right in the middle of a residual whole β-strand, identified by TraDES, which explains the contribution of that residual structure in the DSE stability.

Apoflavodoxin contains residual structures in three helices, residues 41 to 45, residues 108 to 118 and residues 160 to 169 [227], all of which are identified by TraDES. According to DSSP, residues 41 to 52 contain a helix-bend-beta conformation while residues 108 to 118 and residues 160 to 169 contain whole helices. In the protein OmpX, residual structures are reported in residues 73 to 82 and 137 to 145 [228] while TraDES reports residues 65 to 83 and 135 to 147. Residues 65 to 83 form a beta-turn-beta structure and residues 135 to 147 form a whole β-strand. Residues 2 to 14, which hydrogen bond with the residues 135 to 147, have high cumulative scores and shows propensity for forming a residual β-ladder. The hUBF protein has residual structures reported in helix III from 63 to 79, helix I from 18 to 28 and part of helix II from 38 to 43 [27]. TraDES identifies a part of the helix I, III and whole of helix II from 39 to 50. It is clear that TraDES can accurately report experimentally validated native-like structures in chemically denatured proteins. Given the limited experimental residue-level data currently available for the

denatured state, this method of accurately predicting residual structures will quicken and further our understanding of the denatured state.

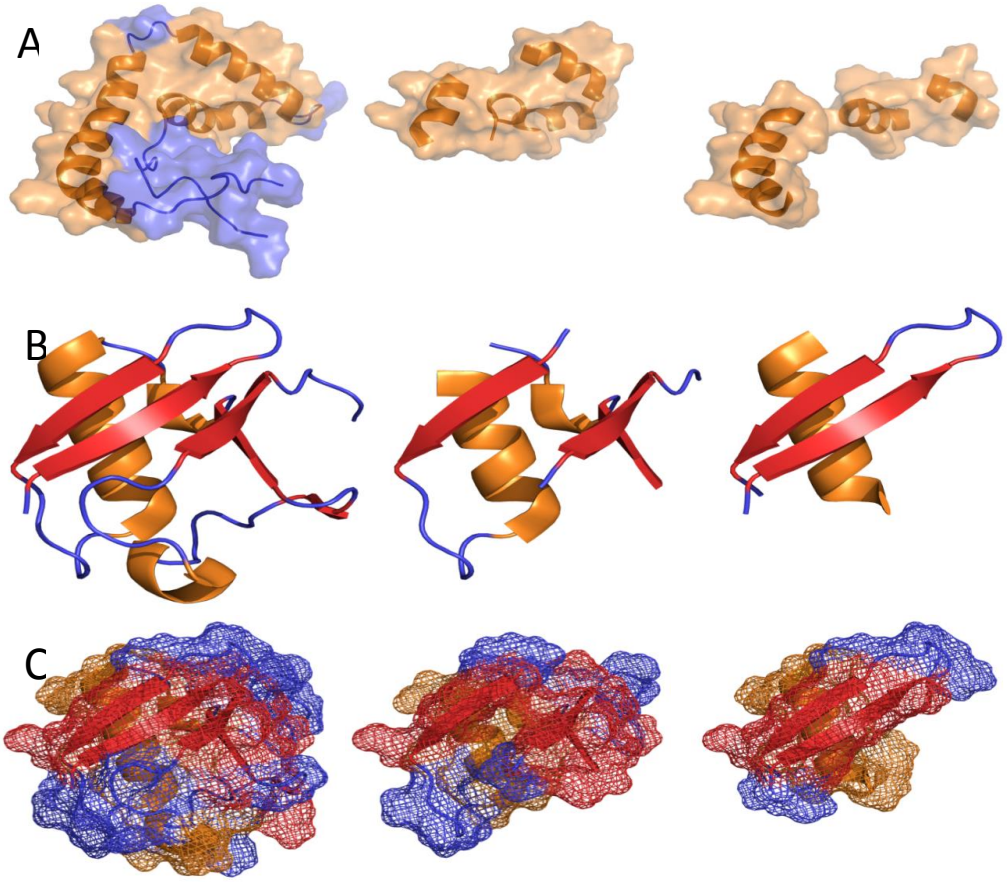

Figure 2.10 Residual structures from TraDES vs experiments. Residual structures identified by TraDES are compared to experimentally determined residual structures for proteins hUBF (A), protein G (B) and ubiquitin (C). The columns (from left to right) correspond to their complete structure, TraDES residual structure and experimental residual structure, respectively.

It is surprising that a simple *de novo* sampling is able to predict such structures under complex denaturing conditions and is further proof that the majority of information underlying the folding mechanism is present on an amino acid level. TraDES reports additional residual structures compared to experimental data. This could be due to the effect of different denaturants, their concentrations [216,217,229] and techniques, while TraDES assumes constant solvent and experimental conditions. It is quite possible that under such different denaturing conditions, proteins may provide us different subsets of TraDES identified residual structures. The different subsets could

be (de)stabilised differently in different denaturing conditions. This may result in only a few residual structures being identified in a given denaturing environment. It is possible that the residual structure in urea is different from that in Gd-HCl and the nucleating sites in the two conditions may differ. More detailed analysis is required to establish such denaturant-based refolding pathways.

The residual structures are not completely formed in every single structure in the ensemble or even completely in a few structures. These are parts of different structures, which on an ensemble average, are skewed towards native-like structures when compared to a random volume-excluded extended model.

The presence of such residual structures including turns and short motifs in the chemically denatured state is surprising. It is now reasonable to assume that these could be directly or indirectly involved in initiating protein refolding. The parameters, dihedral threshold (2°), match threshold (4) (refer Methods) have been generalised in this study. They should ideally be standardised for every protein for better and specific results. This is the first such report of an all-atom ensemble model to calculate the denatured Rgyr of proteins and use the model to describe the inherent residual structure. This could be an important step in understanding protein refolding. These ensembles, if used with other methods such as NMR and FRET, can provide further detailed insights. TraDES also provides N-C distance that can be correlated and validated by FRET. Other studies from our lab indicate that TraDES can predict folded FRET N-C distance for protein p53. This can be extended to denatured proteins and FRET end-to-end distance values can also be used to filter ensemble structures. Similarly, experimental data obtained from SAXS, NMR, FRET, etc. can be used together with DSE or as starting constraints to provide more accurate and specific ensembles. By no means is this a complete analysis of the denatured residual structure. In addition to native-like structures, transient non-native structures and off-pathway intermediates are also present in the denatured state [230]. The mechanism

of folding from these transient states is not directly forthcoming, while native-like residual structures can act as nucleating sites in protein refolding [30]. Such diverse native-like residual structures have also given rise to the consideration that proteins may follow a condition-based unfolding and refolding pathway. The folding process is more robust than it was thought to be, although further detailed studies are required for it to be conclusive.

2.5 Conclusions

In summary, chemically denatured states of proteins contains significant residual structure, which can act as nucleating sites during protein refolding. Describing unfolded denatured states of proteins pose methodological challenges due to their structural diversity. Determining the conformation space of the denatured proteins has been carried out by using SAXS-derived dimension data for a set of proteins with a range of residue lengths. This demonstrates an excellent model for predicting the Rgyr of proteins from their sequence alone.

Further, the sampling ratio was used to generate ensembles to identify native-like residual structures in the denatured state for a set of fifteen proteins. The results demonstrate the presence of short motifs and secondary structures like the β-turn, which are proven to act as nucleating sites in proteins. The approach was verified by comparing experimental results obtained for three proteins from literature with TraDES. Sampling model shows good agreement with experimentally validated residual structures. These residual structures are starting points in protein refolding and can provide a deeper understanding of its mechanism. Protein refolding is the closest model available to *in vivo* folding and insights from refolding can be applicable to the protein folding model. By providing all-atom structures that correspond to these results, this method can be used directly for any further structural or biophysical studies to filter and arrive at specific protein dependent models of protein folding.

In summary, this chapter provides a computational approach to understand denatured protein structures from their sequence by sampling specific regions of the protein conformational space. The residual structures suggest short residual structural motifs that act as nucleating sites for protein refolding.

3. Nascent Polypeptide Structure

"Everything that living things do can be understood in terms of the jiggling and wiggling of atoms"

-Richard Feynman

3.1 Introduction

Ribosomes are the machinery required for polypeptide synthesis by mRNA translation and have been intensively studied for this reason. The ribosome is an assembly of two subunits (large and small), each made up of RNA and proteins. Once the structure of the ribosome was solved, the mechanism of peptide bond formation became clear. The peptide bond is formed in the peptidyl transferase centre (PTC) in the large subunit and the peptide is elongated from its N- to C-terminus. The PTC, where the protein chains are synthesised, lies deep inside the ribosome. The terms "nascent chain" or "nascent polypeptide" are generally used to denote polypeptides that are still attached to the PTC in the form of peptidyl-tRNA.

The nascent chains in the ribosome pass through a tunnel-like cavity inside the large subunit, referred to as the peptide exit tunnel. Early evidence indicated that parts of the nascent chains (30-40 residues) were protected by the large subunit from degradation by proteases [231,232]. This property of the large subunit was later explained by the presence of a tunnel using three-dimensional image reconstruction of the ribosomal large subunit [233]. This tunnel was suggested as the path taken by the nascent chain through the ribosome [233]. High-resolution crystallographic and cryo-EM reports confirm that proteins traverse a narrow tunnel from the peptidyl transferase centre to the surface [234,235]. This results in the nascent chains travelling almost 100 Å along the peptide exit tunnel before they are released [236,237]. The ribosomal peptide exit site identified earlier by the use of antibodies [238] also matches the structural data obtained later. Like the rest of the ribosome, the

tunnel is composed of both proteins and RNA [234]. The tunnel is non-uniform and its diameter varies from 10 Å in its narrowest part to around 20 Å near its end [234].

The biophysical properties of the tunnel are not completely understood. The tunnel, originally thought of as a passive conduit for the nascent polypeptide, has been recently associated with diverse functions. The tunnel is reported to play a functional role in protein development. The tunnel can arrest protein elongation by sending specific signals, through the proteins and RNA, to the peptidyl transferase centre [239]. There have been reports of the tunnel acting as a gate discriminating between different protein sequences [240,241]. Macrolide antibiotics also bind to specific sites on the tunnel, blocking protein elongation [242]. Although it is clear that the exit tunnel helps mediate certain structural changes that affect protein movement, it is unclear how its non-uniform branched structure and its geometry contribute to its other functions such as folding, gating, etc.

Protein folding and the effects of the ribosome on it are topics of intense study. Co-translational folding, effects of RNA on folding, recruitment of chaperones by the ribosome and folding inside the ribosome are different aspects of the same problem. Although investigating the folding inside the ribosome is experimentally challenging, the remaining functions take place outside the ribosome and can be studied with relative ease. Proteins start to acquire secondary and tertiary structures as soon as they emerge from the exit tunnel [243,244] and make distinct, highly conserved contacts with the tunnel exteriors. Theories on the ability of the tunnel to accommodate secondary and tertiary interactions are inconsistent. Some report that the geometry of the tunnel cannot accommodate any secondary structures [71]. However, it has been proposed that the shape and size of the tunnel could accommodate some alpha-helical proteins [245]. Fluorescence resonance energy transfer (FRET) and cysteine scanning “molecular tape measure” studies are consistent with a model in which transmembrane alpha-helix folding is initiated in the

tunnel [72,246,247]. Furthermore, there is experimental evidence from PEGylation studies of tertiary interactions within the tunnel [70]. The discrepancy lies in the fundamental difference in approaches. The studies based on the rigid crystal structure point towards an inability of the tunnel to accommodate folded subunits while *in vivo* dynamics studies point to a more accommodating tunnel. This dynamics view is supported by cryo-EM studies, which show that the tunnel of a translating ribosome expands during protein synthesis [72]. Hence, it is plausible that the dynamic translating ribosome could accommodate secondary and tertiary interaction while a rigid non-translating ribosome cannot.

A cryo-EM structure of a ribosome with a stalled nascent polypeptide in the exit tunnel [239] is available, which provides details on the interacting residues. The finer details are unavailable since its resolution is only 5.4 Å, but this expanded stalled ribosome structure provides a great starting point for understanding the dynamics of the tunnel and polypeptide during translation.

Structural dynamics of biomolecules are hard to capture by conventional techniques. Crystallography can only describe static ordered structures and is restricted in its ability to determine protein dynamics since conformational heterogeneity cannot be captured by X-ray crystallography. NMR can provide structural and dynamics data [248] but is limited in the size of the molecules that can be studied due to their shorter NMR signal relaxation times and slower tumbling rates [249]. This limitation on the size of molecules makes it practically hard to study ribosome dynamics by NMR. Molecular dynamics (MD) is a well-established alternative to study the dynamics of macromolecules. All-atom models of the ribosome have been simulated but results mostly focus on protein synthesis, folding outside the ribosome or dynamics between different parts of the ribosome. Large-scale conformational changes in the ribosome have been recently analysed using fluorescence and mostly focus on movements of individual ribosome components. Co-translational folding of a nascent chymotrypsin

inhibitor has been simulated using a coarse Go-like model, which reinforces earlier experimental results [244]. All-atom models of the ribosome have also been studied using MD simulations [250-252] to help understand polypeptide movement and conformational changes in different segments of the ribosome. None of these studies focus on the geometry and dynamics of the tunnel nor do they shed any light on the effect of the polypeptide on the tunnel geometry and vice versa. A better understanding of the tunnel expansion due to the nascent peptide would assist in developing more suitable and effective macrolide antibiotics, which bind to the tunnel and prevent protein translation.

The large voids in the ribosome pose a significant challenge in elucidating the geometry of the tunnel. Numerical methods are available for computing the volume and surface area of a union of balls [253,254]. One such approach, the rolling ball algorithm, has been used earlier to identify the tunnel [71]. However, the surface area and/or volume computed by numerical integration over a set of points, even if closely spaced, are not accurate and cannot be readily differentiated [255]. There is a smoothing effect produced by the rolling ball algorithm and its resolution is only as good as the radius of the ball. A smaller radius of the ball will include more voids, which are not part of the tunnel, while a bigger radius will reduce the resolution and details of branching.

An alternative approach is to use coordinate geometry and Delaunay triangulation to identify the tunnel. Delaunay triangulation represents the protein as a union of balls with standard van der Waals radii. The centres of the spheres are used to segregate overlapping regions between adjacent atoms. The lines connecting the centres of spheres result in triangles and tetrahedrons in two and three dimensions, respectively. A cavity is detected if the point of intersection of the spheres lies outside the individual spheres. This approach has been used to identify cavities in protein structures [256-258]. Here, we apply Delaunay triangulation to the ribosome structure

and try to capture the tunnel and its geometry. The large subunit stalled ribosome structure with the peptide is available and is simulated using molecular dynamics. The ribosome is simulated with and without the peptide in the tunnel and the effects of the presence of the polypeptide are analysed. Delaunay triangulation is applied to the trajectory of the simulations and the time-resolved tunnel geometry is quantified. Volume and surface area of the tunnel are calculated for the entire trajectory to study its dynamics.

The confined space in the tunnel, in turn, has an effect on the structure of the nascent polypeptide [259]. The tunnel geometry limits the space available to the polypeptide and restricts it to a smaller subset of conformations. The heterogeneity of the tunnel also means that conformations accessible to the nascent polypeptide change along the tunnel. The tunnel is known to expand during translation, but there are no structures for nascent polypeptides completely traversing the tunnel. A cryo-EM structure is available for a 20-residue polyalanine peptide that covers around 60 Å of the 100 Å tunnel. The tunnel has narrow regions close to the PTC, including a constriction [240,242] while regions close to the exit are reported to accommodate tertiary interactions [69]. Here, we analyse available nascent peptide conformations using snapshots of the nascent peptide in the tunnel from MD.

MD trajectories provide snapshots of the peptide over time, but do not guarantee a complete sampling of the conformational space available to it inside the tunnel. Similar to earlier simulations, we can at best capture the local structural fluctuations around the starting structure [251]. Hence, to get a comprehensive sample set of nascent polypeptide structures, Monte Carlo conformational sampling is carried out using the Trajectory Directed Ensemble Sampling (TraDES) package. Dock by superposition is used along with Ensemble Sampling to generate a nascent polypeptide ensemble that can fit inside the exit tunnel. A simple dock by superposition also reports the quality of a structure or an ensemble to fit in the tunnel

by the number of clashes with the atoms that make up the tunnel. For comprehensive sampling, the cryo-EM structure is unfolded to provide a diverse ensemble, which is then filtered using dock by superposition.

3.1.1 A threshold of tunnel constraints

The varying local geometry along the tunnel can have different spatial effects. Since the tunnel is constricted close to the PTC and wider near the exit, there are bound to be spatial thresholds in the tunnel where the tunnel stops acting as an active constrictive apparatus and becomes more of a guiding apparatus. We define a threshold as an end of a constriction or gating after which nascent polypeptide structure is less constricted or not actively under tunnel constriction. Once past the threshold, the peptide may no longer be under strict structural regulation and is more likely to take up favourable residue-based conformations. The ensemble structures are aligned inside the tunnel based on the cryo-EM structure. This results in mapping the consecutive residues in the chain to different positions in the tunnel. By progressively constraining residues, from the ones closest to the PTC to the ones closest to the tunnel exit, different ensembles are generated that represent different thresholds. Looking at the different ensembles and the quality of their member structures, it is possible to infer which parts of the tunnel are more constricting or accommodating. For instance, if the ensemble generated by constraining residues one to seven has a substantially better subset of acceptable conformations (from dock by superposition) than residues one to six, then the position of the 7^{th} residue has a more accommodating geometry. By successively generating such ensembles and analysing them, we can indirectly find if there are any spatial thresholds in the tunnel geometry and analyse their positions and geometry, if present.

The ensembles of the polypeptide generated (by MD and MC) to understand its conformation are based on the cryo-EM structure of the peptide taken as the initial structure. Unfolding provides a wide distribution of conformations, but it is entirely

wrong to assume that these are the only sets of structures that the tunnel can accommodate. It is possible that this ensemble might be one of the subsets possible inside the tunnel. On the other hand, this could be the only possible closely clustered structure. In other words, can the restrictive tunnel environment give rise to only a small ensemble of closely related structures similar to the cryo-EM structure, or is that only a subset of a much larger population? This can be answered by generating *de novo* ensembles of structures that can fit into the tunnel by using standard sampling in TraDES. The standard sampling uses libraries that are derived from the overall set of conformations that are observed in the PDB. The tunnel geometry is the only restraint that can be applied to filter the structures that are generated through the standard sampling. This is a very cumbersome procedure since every residue has a large set of conformations it can choose from. Over a set of twenty residues in the peptide chain, the overall sample set exponentially increases and is computationally intensive. So, a steepest descent algorithm is used to filter and direct the structure sampling towards the optimal set. The steepest descent is a simple first-order optimisation procedure which iteratively chooses the best result in every step and moves along minimising/maximising a given parameter over multiple runs. Here, the quality of the structure is minimised using the steepest descent algorithm. The quality of the structure is defined by the number of clashes in dock by superposition, and understandably, lower clashes reflect a better quality structure. The tunnel geometry alone is used to derive a *de novo* ensemble of possible nascent peptide structures using a steepest descent method based on maximising the quality of the ensemble (i.e. minimising the number of clashes in dock by superposition). These results will show us if the original cryo-EM sample set is the only possible set of structures in the tunnel or if they are only a subset.

In all the analyses, the residue conformations are a direct consequence of their respective tunnel positions. This can be used to apply all the results obtained from the

peptides to infer geometry in the local tunnel regions. This first study of the conformation regulation and dynamic interplay between the nascent polypeptide and the ribosome allows us to understand constrictions and important spatial and geometric thresholds along the tunnel.

3.2 Methods

3.2.1 50S Ribosome MD

The *E. coli* ribosome large-subunit (PDB ID: 2WWQ) [239] was simulated using AMBER [260] using FF-03 force-field parameters [261]. The large subunit structure included the 5S ribosomal-RNA, 23S ribosomal-RNA, P-site tRNA, mRNA and 31 large subunit proteins. This is the structure of the large subunit of the ribosome stalled during translation, with the TNAC leader peptide still in the peptide exit tunnel. This structure was already refined by Molecular Dynamics Flexible Fitting [262] using an earlier *E. coli* ribosome structure [263]. The nascent polypeptide was a 20-mer polyalanine structure in the ribosome tunnel. The ribosome structure that had the nascent polypeptide inside the ribosome tunnel was retained for simulation of the tunnel in the presence of the peptide. To understand tunnel dynamics without the peptide, the peptide structure was removed and the rest of the ribosomal structure was equilibrated and simulated. In both simulations, other parameters were maintained as follows. The maximum distance between atom pairs considered for pairwise summation in calculating the effective Born radii was set at 12 Å. So, atoms whose associated spheres are farther away than 12 Å from any given atom will not contribute to that atom's effective Born radius. The non-bonded cut-off for the Generalised Born (GB) model [264] was set at 12 Å. The GB model used was a pairwise GB model with default radii set up by LEaP. The simulation was carried out until the energies and the volume and surface area are equilibrated. Triplicates of the simulation were run and analysed. The length of the simulation was decided based on the time taken for the energy, volume and surface area of tunnel to be stabilised. The

potential and kinetic energy was constantly checked to decide on the time and were analysed to check the accuracy of the simulation.

3.2.2 Delaunay triangulation

Delaunay triangulation (DT) is a method in coordinate geometry that separates the space containing a set of points into regions closest to those respective points. The protein or RNA is represented as a union of balls, where each atom is a sphere with their respective atomic radii. For any given set of finite spheres S_i with centres Z_i and radii r_i, the Voronoi region of S_i consists of all points x, which are closest to S_i than to any other sphere.

Given a finite set of discs, the Voronoi diagram decomposes the plane into regions in which one circle minimises the square distance measure as $\|x-z_i\|^2 - r_i^2$. In Figure 3.1, the Voronoi diagram is restricted within the portion of the plane covered by the discs to get a decomposition of the union into convex regions. The dual DT is obtained by drawing edges between circle centres of neighbouring Voronoi regions. To draw the dual complex of the discs, we limit ourselves to edges and triangles between centres, whose corresponding restricted Voronoi regions have a non-empty common intersection.

This decomposes all the spheres into convex regions such that the boundary of each such region consists of spherical patches and planar patches on the boundary of the Voronoi diagram. Delaunay triangulation is the dual of the Voronoi diagram, obtained by drawing an edge between the centres of two spheres if they share a common face (Figure 3.1).

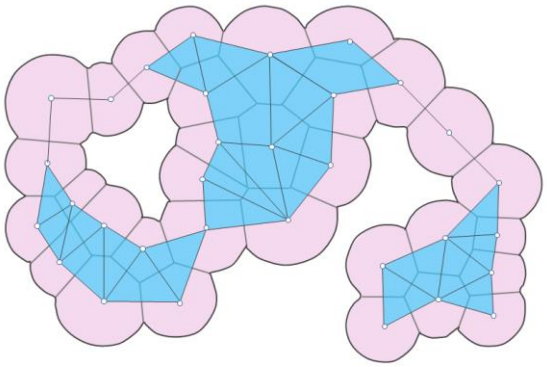

Figure 3.1 Voronoi decomposition and dual complex of a set of points. The regions in pink are the Voronoi regions of the points and their Delaunay triangulation is in cyan [255].

Furthermore, a triangle is drawn connecting the centres of three spheres if they intersect in a common line segment and a tetrahedron is drawn if four centres meet at a common point. These tetrahedrons are called simplices and can be directly extrapolated to the union of balls representation of protein or RNA atoms. The algorithm for triangulation is based on the alpha shape theory [255], specific to molecular simulation applications implementing the weighted surface area, the weighted volume and the derivatives of both. The Delaunay triangulation uses the improved incremental algorithm by Anglada [265] which uses the method of point insertion. It is based on a building approach that follows the observation that a new point only modifies the triangles whose circum-circles contain the point. Therefore when a new point is added, only the triangulation around the point needs to be updated. A triangle containing all the points is used as the starting point. The points are incrementally added and the local triangulation is updated. It has been proved that this iterative process converges, after a finite number of steps, towards the complete, accurate Delaunay triangulation [265]. The algorithm uses a modified version of the incremental approach described by [266]. This method uses a randomised incremental model and is implemented in FORTRAN and C. The incremental algorithm is more time-efficient and applicable to large biomolecules like the ribosome. The scripts for

preparing the structures, running the algorithm and extracting the results were written in-house in PERL.

Using the simplices as representations of the union of balls, the inclusion-exclusion formula of the dual complex accurately calculates the volume [267]. Surface area is also calculated by using an extension of the same principle formula as shown in [268]. These approaches to calculate volume and surface areas have been detailed in earlier reports [269,270]

Pockets or cavities in biomolecules are defined as when the intersection of the spheres occurs outside the spheres themselves. Thus, pockets or voids can be essentially obtained by subtracting the limited dual triangulation, constructed as mentioned above, from the total dual triangulation of all the centres. To make them follow solvation patterns seen in biomolecules, often the radii of the spheres are increased by a default value of 1.4 Å, which is roughly the radius of a solvated atom compared to one without.

Delaunay triangulation was tested using the crystal structure to identify the atoms of the peptide exit tunnel. Due to the presence of large empty spaces inside the ribosome, it was not straightforward to obtain the tunnel by using the algorithm and so two different strategies were used to capture the tunnel. Atoms within 20 Å of the stalled peptide were considered to be part of the tunnel and DT was applied to those atoms. In another method, the radii of atoms, not part of the tunnel, were increased to collapse the space and triangulation applied. In both cases, the solvent radius of the atoms needed to be optimised. Different radii were checked to find an optimal value, which included the water default of 1.4 Å. The tunnel was well established when the solvent radius was set at 2.5 Å to overcome the large space in the ribosome.

In the second method, the non-tunnel atom radii also needed optimisation in addition to the solvent radius. Various radii were tried and the resulting tunnel geometry was compared. As mentioned earlier, the large gaps inside the ribosome combined with

the concave 30S binding pocket made it very challenging to capture the tunnel alone. Based on an earlier report of the geometry of the tunnel [71], we were able to compare the two methods with different parameters. It showed that the first method of considering the tunnel atoms alone with a solvent radius of 2.5 Å seemed to capture the tunnel geometry well. This was applied to the trajectory of the 50S ribosome simulation and the various changes to the peptide tunnel were captured. For the simulation with the peptide, the tunnel was captured disregarding the presence of the peptide so that the two simulations with and without the peptide could be compared. The same atoms that make up the tunnel were considered for tunnel triangulation in both simulations so that the exact changes due to the peptide could be measured.

The trajectories of both the simulations were used and the snapshots of the trajectory were extracted as individual PDB file using VMD [271]. An in-house program was used to iterate through each of the structures and apply Delaunay algorithm and extract the volume and surface area of the tunnel cavity. The other cavities, like the 30S ribosome binding site and other concave surfaces, which show up in the analysis, were disregarded to focus on the necessary data for just the ribosomal peptide tunnel.

3.2.3 Trajectory Directed Ensemble Sampling (TraDES)

Trajectory Directed Ensemble Sampling (TraDES) is a Monte Carlo structure sampling tool to sample protein conformational space [155]. This uses a volume-exclusion model to build the protein structure from the N-terminal to the C-terminal by residue-wise addition. There is a backtracking algorithm which is used when a residue cannot be place due to the steric clashes with the previous residues [154]. This results in protein structures being generated in a random walk mechanism without running into the previously generated residues. TraDES has a residue-wise probability distribution derived from high resolution structures in the PDB. Figure 3.2 shows an example distribution of alanine using standard sampling.

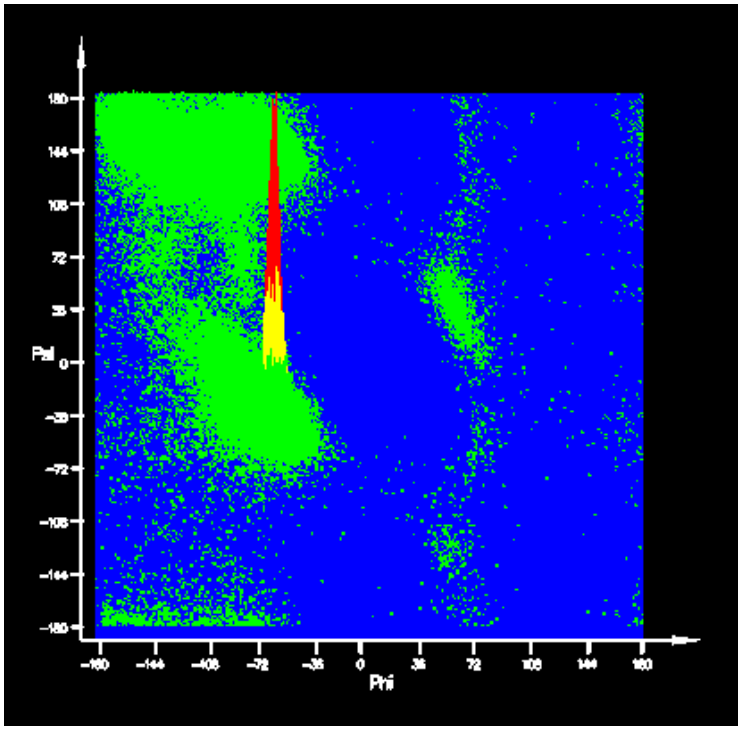

Figure 3.2 TraDES standard probability distribution for alanine. These probabilities are taken from high resolution structures in the PDB. Green, yellow and red represent low, moderate and high probabilities respectively.

TraDES has different dictionaries for standard, extended, coil and alpha-helical sampling. The sampling can also be modified using different set of ratios. These dihedral distributions are sampled in 3D space based on the sequence of the protein. TraDES can also sample conformational space with reference to a structure. The `str2trj` program can use a native MMDB ASN.1 structure file [272] as input and can create a trajectory distribution with the same dihedral angle observed in the crystal structure. This can be extended to sample the conformational space around the structure by unfolding. The extent of unfolding can be dictated using the temperature and time parameters. This will unfold the structure and create a distribution based on the time and temperature. Higher time and higher temperatures give rise to a larger distribution. Figure 3.3 shows the effect of unfolding using a temperature of 350 K and a time step of 150 fs based on the crystal structure for one residue. The centre of the unfolded distribution is the native dihedral of the residue, based on which it is unfolded.

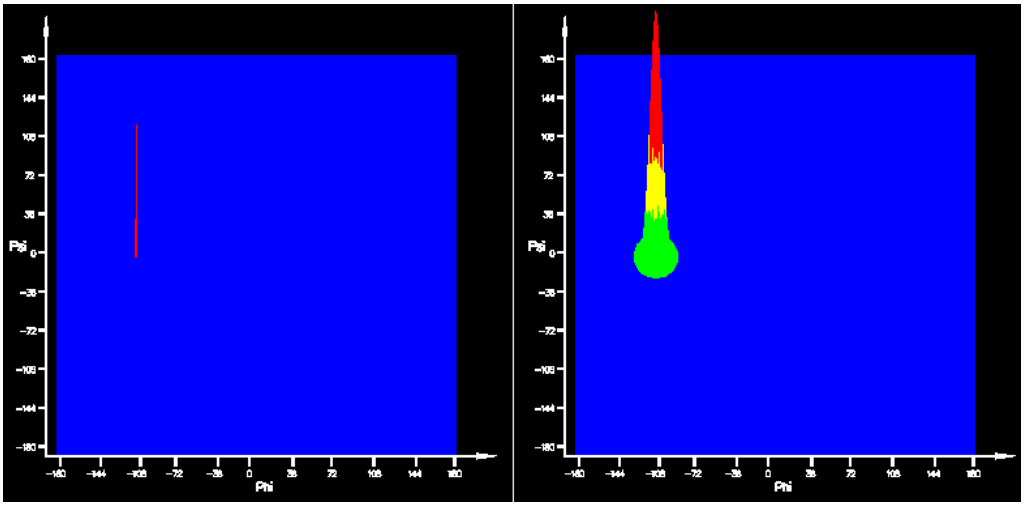

Figure 3.3 Effect of unfolding on the dihedral distribution. Unfolding of the conformation of residue 12 in the nascent chain (left) at a temperature of 350 K and time step of 150 fs results in the unfolded dihedral distribution (right). The colours represent probabilities with green being least likely, yellow as moderately likely, and red, the most likely.

The nascent polypeptide structure from cryo-EM was taken as the starting reference structure and a trajectory distribution was generated in Ramachandran space using the TraDES program `str2trj`. `str2trj` takes a protein structure in MMDB ASN.1 format [272] and outputs a trajectory distribution that can be sampled by the program `trades`. The trajectory distribution is a set of residue-wise dihedral probabilities along the protein chain. The individual dihedral distributions are the dihedral probabilities for every amino acid taken from the PDB, as mentioned above. The standard dictionary is the complete set of dihedral probabilities from all secondary structures. To avoid secondary structural bias, this dictionary was used for the polyalanine nascent polypeptide trajectory distribution for sampling using only its sequence.

3.2.3.1 Sampling based on the cryo-EM structure

The peptide was unfolded at a temperature of 300 K and time-step of 150 fs and was sampled using the program `trades`. The first five residues closest to the PTC were not unfolded and their dihedrals were exactly set to the crystal structure. This unfolding was done to eliminate grossly inaccurate conformations while still

providing a good sampling. This trajectory distribution was then used to generate 500,000 structures of the peptide. These structures were aligned to the nascent peptide crystal structure with reference to residues one to four (all-atom alignment). These were then docked by superposition against the crystal structure of the ribosome without the peptide and the Van der Waals clashes were evaluated by the program `crashcheck.pl` (http://trades.blueprint.org) with a leniency of 0.25 Å for elastic collisions. Dock by superposition between the crystal structure of the peptide and the crystal structure of the ribosome produced five clashes. Considering this value and manual checking of a sample of structures, the cut-off was set at five clashes for the generated peptide structures to fit into the tunnel and be a measure of the conformational space inside it. The peptide structures thus filtered were analysed using Ramachandran plots and taken together to provide the tunnel space.

3.2.3.2 Tunnel geometry constraint using steepest descent

De novo sampling using steepest descent was carried out by sampling the polyalanine chain with standard sampling. A 20-mer alanine sequence was the input in the TraDES `seq2trj` program to generate a trajectory distribution from the sequence. The first four residues closest to the PTC were constrained based on the cryo-EM structure for reasons related to alignment. The first four residues were used as references to align the ensemble structures inside the tunnel. The rest of the residues from five to twenty were unconstrained and given standard conformational probabilities in the `str2trj` program, which represent the standard dihedrals observed in the PDB [188]. The standard sampling contains the whole set of dihedrals observed in protein structures in PDB and the probability of obtaining the required structures reduces dramatically by the use of the standard sampling. To perform dock by superposition for a huge ensemble is computationally very intensive compared to generating those using TraDES. So, to improve the ensemble quality and eliminate grossly inaccurate structures, two filters were used to select only possibly accurate

sequences. End-to-end distance and C-C distance filters were applied to every generated structure and only structures that pass these filters were considered for dock by superposition. The end-to-end or the NC filter filters structures based on their end-to-end distances (Figure 3.4). This is based on the fact that polypeptide structures do not fold back on themselves inside the tunnel. The structures that had an end-to-end distance of less than 40Å were discarded, since this meant that they folded back on themselves which is not possible in the tunnel. The C-C distance is the distance between the C-α atom of the last residue in the generated structure and the C-α atom of the cryo-EM structure inside the tunnel. This is a measure of directionality of the chain. Structures that have C-C distances less than 10 Å are directed towards the exit of the tunnel. The 10 Å CC filter gives a sphere of radius 10 Å near the exit of the tunnel with its centre as the C-α atom of the final residue in the cryo-EM structure (Figure 3.5). The CC filter is only applied to structures that pass the NC filter. So, the structures that pass these filters are both elongated and directed towards the exit of the tunnel. By giving it a 40 Å NC distance filter, it eliminates inaccurate structures while still not forcing any rigorous elongated state.

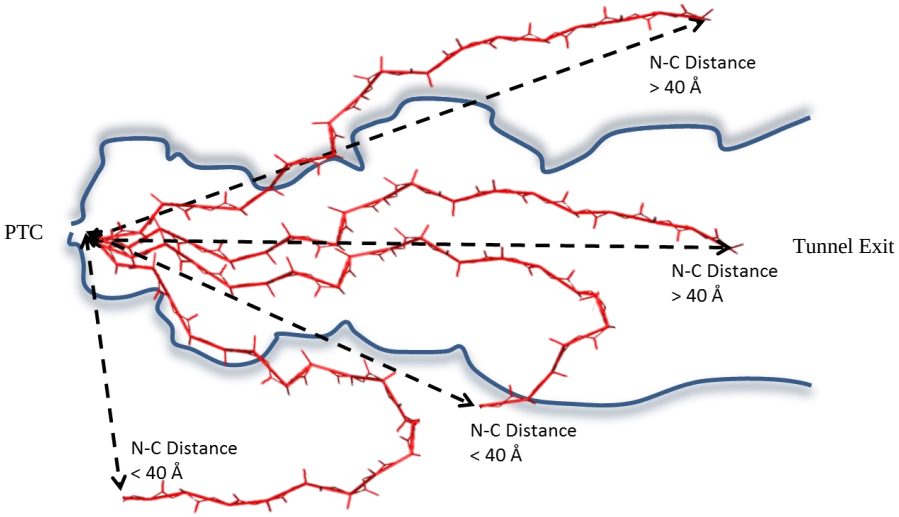

Figure 3.4 Illustration of the N-C distance constraint. Nascent peptide structures (red) that have their end-to-end distances shorter than 40 Å are expected to have folded back on the tunnel (blue), which is not possible. The fully-extended cryo-EM structure has an N-C distance of 50 Å. So, 40 Å filter will serve to eliminate grossly inaccurate structures while providing some freedom for the peptides. From left to right, the tunnel is represented from the PTC to its exit.

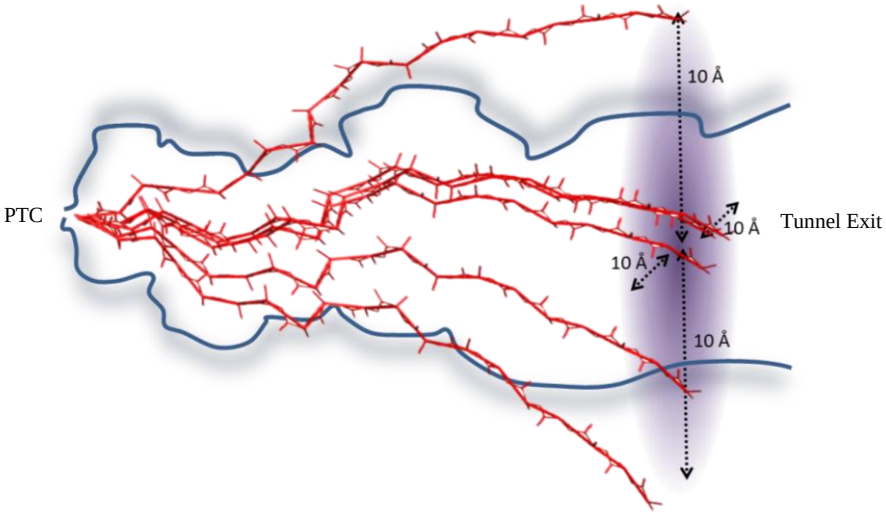

Figure 3.5 Illustration of CC distance filter. Nascent peptide structures (red) that have their last C-α carbon farther than 10 Å from the cryo-EM C-α are filtered out. Structures that are farther than 10 Å from the cryo-EM structure would not fit in the tunnel (blue) since they would have hard sphere overlaps with the tunnel residues. From left to right, the tunnel is represented from the PTC to its exit. The CC filter is essentially a sphere of radius 10 Å with the 20th C-α of the cryo-EM structure in its centre. Ensemble structures that do not have their 20th C-α atom inside the sphere will be filtered out.

The steepest descent method is applied by using multiple runs of ensemble generation based on the best structure from the previous run. Initially, a 20-mer polyalanine with standard sampling probabilities is taken as the starting structure. The first four residues are fixed to crystal structure dihedrals to aid in alignment. 4000 structures are generated per ensemble that pass both the above NC and CC filters and are analysed by dock by superposition. After every run, the best structure with the lowest clash is taken, unfolded and used as the starting structure for the next run. The average clashes of the 4000 structures and the lowest clash are continually monitored every run. The minimisation continues until these two parameters have reached a plateau minimum. The final resulting ensemble and the structures that pass the dock by superposition are compared to the crystal structure ensemble in Ramachandran space.

3.2.3.3 Spatial thresholds in the tunnel

Tunnel spatial thresholds are analysed by using the same polyalanine sequence but by progressively constraining residues. The first three residues are always constrained to their cryo-EM dihedrals as a reference for aligning the peptide into the tunnel. Different ensembles with 2,000,000 structures are generated for every position from residue four to residue fourteen. For example, for the 'residue 10' ensemble, all residues from one to ten are constrained similar to the cryo-EM structure and residues eleven to twenty are sampled using standard sampling. The same NC and CC filters are used to restrict and eliminate inaccurate structures. They also provide a coarse estimate of the threshold by reporting the number of structures that pass the filters for every threshold. Eleven different trajectory distributions (residue four to fourteen) were generated with the respective residues constrained (Figure 3.6). The spatial threshold for each of the distribution is the position of the last residue in the tunnel. This is a measure in terms of distance from the first C-α atom to the C-α atom of the last constrained residue. The numbers of structures that pass the NC and CC

constraints and the average crashes are plotted against different thresholds in the tunnel. The filtered structures are analysed using Ramachandran plots to shed light on any distinct characteristics imparted by the different thresholds.

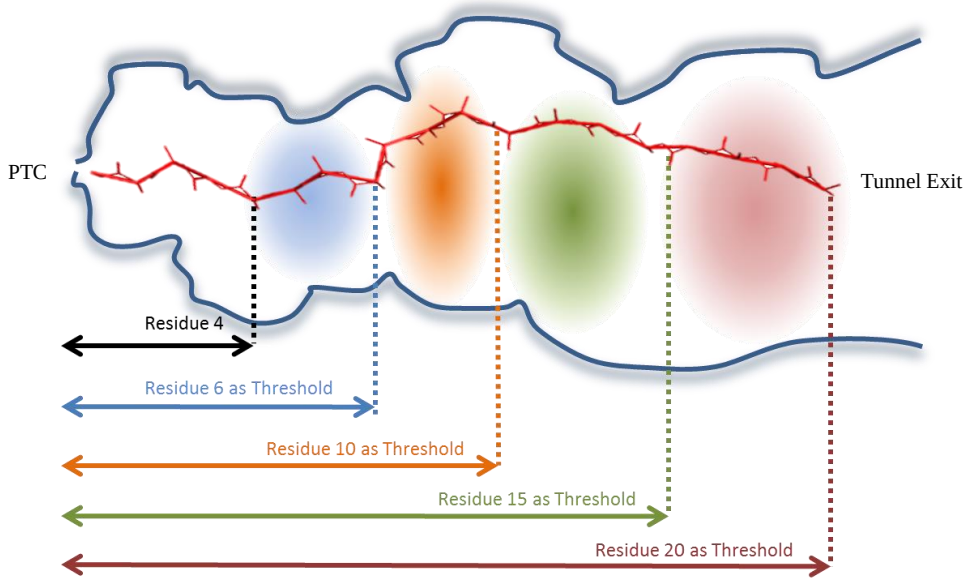

Figure 3.6 Description of the spatial thresholds.in the tunnel. Different residues correspond to different segments of the tunnel. By constraining the peptide progressively and analysing the quality of the resultant ensemble, it is possible to understand spatial thresholds in the tunnel. For example, if constraining residues 1 to 6 and allowing the rest of the peptide to sample standard Ramachandran space gives a worse ensemble than by constraining residues 1 to 5, we can predict that the tunnel segment around residue 6 is not accommodating. Similarly, if there is large improvement in ensemble quality by constraining residues 1 to 9, compared to 1 to 10, then we can predict that the tunnel segment corresponding to residue 9 is a threshold after which the tunnel is more accommodating. From left to right, the tunnel is represented from the PTC to its exit.

3.2.4 Nascent polypeptide analysis

From the simulation of the ribosome with the nascent polypeptide in the tunnel, the trajectory of the peptide structures were extracted and made into an ensemble using an in-house PERL script. This represents the entire set of structures that fit into the tunnel during the entire simulation. The torsion angles of the various amino acids were studied using Ramachandran plot [273]. The torsion angles were calculated using PyMOL [274] with a Python code to run through all the peptide structures and the residue based Ramachandran plot was generated using OriginPro [275]. The residue-wise spatial interactions of the nascent polypeptide were calculated manually

using residue-residue distance of 5 Å in PyMOL and were correlated with the Ramachandran plot to arrive at conclusions. This provides details of the freedom available to the various residues in the chain. Given that the chain is polyalanine, any conformational bias should be directly related to the local structural space available to the amino acid inside the tunnel rather than amino acids composition. To avoid end effects, we report the torsion angles of only residues 2 to 19. Thus, by correlating the torsion angle data to the actual spatial presence of the residue in the tunnel, we can inspect the various zones of the tunnel and report where in the tunnel are constrictions that might alter the conformational space available to the amino acid that passes through it.

3.3 Results

3.3.1 Capturing the tunnel using DT

Delaunay triangulation was applied to the entire ribosome structure to identify the tunnel and understand its dynamics. Delaunay triangulation of the ribosome alone without the nascent peptide was done and the tunnel was identified (Figure 3.7). It also shows the superposition of the nascent polypeptide from the ribosome and the tunnel derived from the triangulation.

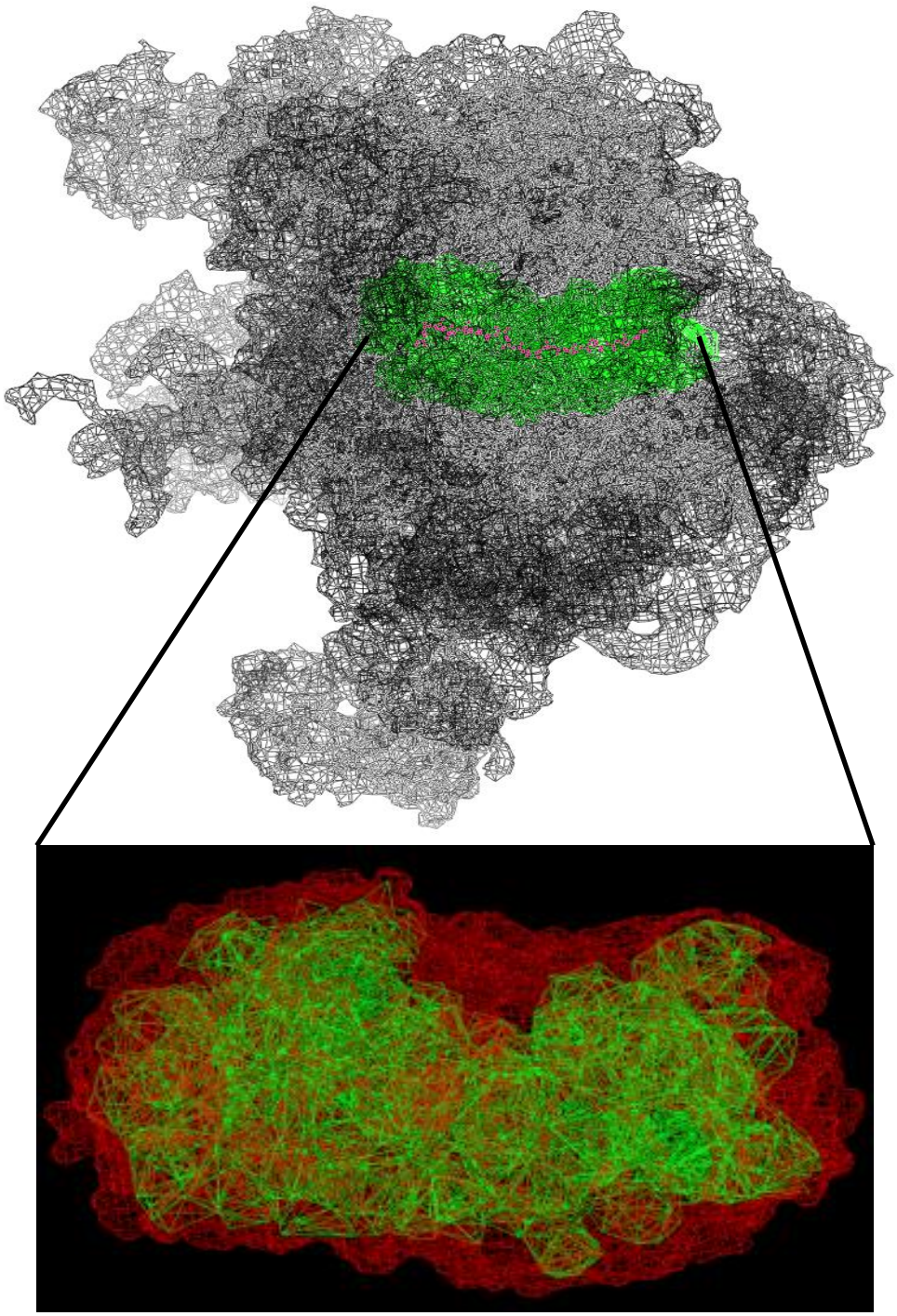

Figure 3.7 Ribosome exit tunnel captured by Delaunay triangulation. The large subunit of the ribosome (black mesh) is shown with the captured tunnel (green) and the nascent polypeptide (pink). A closer look at the tunnel geometry (green) along with the residues that make up the tunnel (red mesh).

It is clear that the tunnel is captured effectively and it encompasses the polypeptide. Atoms in the ribosome which are at a distance of 20 Å from the nascent polypeptide are considered to constitute the tunnel, and tunnel was captured using these atoms.

This is done to overcome the problem of the huge empty space within the ribosome. The other method of increasing the radius of the atoms depending on their distance from the tunnel also provided similar results, with the former method proving to be better. The tunnel identified is superimposed with the polypeptide showing an almost perfect fit, and also provides details on the features of the tunnel. In Figure 3.7, the green lines are triangles that form the voids. The tunnel is not a uniform structure and contains many small branches and constrictions along the way as previously observed [71]. So, the nascent polypeptide need not necessarily behave the same along the entire length of the tunnel and could tend to achieve local secondary structures, even if they are restricted to only a few residues. Although this gives a rough idea on the tunnel characteristics, it sheds no light on how dynamic it is or how it might change due to the presence of the polypeptide. So, the same Delaunay method was applied to the trajectory and volume and surface area captured.

3.3.2 Ribosome MD

The ribosome simulations were carried out with and without the nascent polypeptide in the exit tunnel and the trajectories were analysed by calculating the various energies like potential and kinetic energies. The simulation seems to be very stable and reaches equilibrium of constant potential energy and kinetic energy. The simulation was run until it stabilises to make sure it has reached its equilibrium state. The various energies were calculated and showed that the simulation proceeds to a lower energy and the simulation is stable in both cases. The trajectory is very smooth, gradual and reaches a plateau in both potential and kinetic energies. Both the simulations were very similar in both energies. This shows that deleting the nascent polypeptide does not have any adverse effects on the simulation and its trajectory. It also does not lead to collapse or any large conformational changes, which would clearly show up as anomalies in the potential energy of the system.

3.3.3 Delaunay triangulation

Trajectory of both the above simulations was used and Delaunay triangulation was applied to identify the tunnel. Application of the Delaunay triangulation was not straightforward due to large voids in the ribosome. So, by using a solvent radius of 2.5 Å around the atoms, as mentioned in Methods, the tunnel was identified. The tunnel cavity along with the residues around it was captured well. The tunnel captured was superimposed with the nascent polypeptide structure and it completely encompasses it with a lot of space available (Figure 3.8). The space around the peptide seems to be present to accommodate various amino acids that contain bulky side chains and seems enough for residues with bulky side-chains unlike the alanine present in this structure.

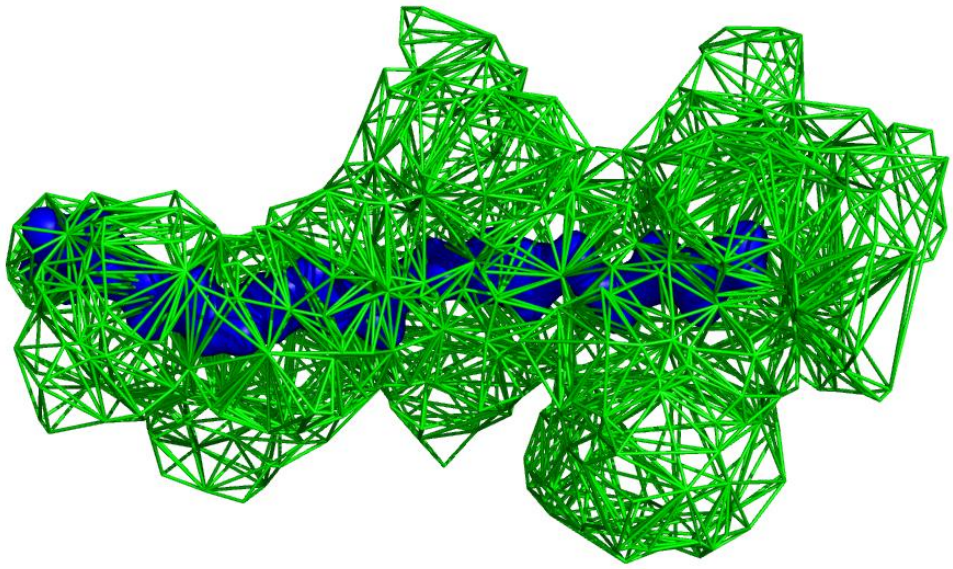

Figure 3.8 Tunnel captured by Delaunay triangulation. The tunnel captured (green) is superimposed against the nascent polypeptide (blue). The tunnel geometry is branched on all sides and provides room for the peptides to traverse it. Branch sizes vary and different branches can be expected to differ in their accommodating effectiveness.

3.3.4 Tunnel dynamics

Delaunay triangulation was applied to trajectories of the two above simulations and tunnel dynamics in both cases were compared. The volume and surface area of the tunnel was calculated with respect to time with and without the peptide in the tunnel

(Figure 3.9). The graphs show that they fluctuate and stabilise after about 200 steps of simulation. The initial jumps in the volume and surface area are due to the wriggling of the side chains to avoid high energy interactions. In such complex structures like the ribosome, the initial fluctuations are due to the various side-chain and inter-chain clashes. Since the tunnel is comprised of various proteins and RNA held together, movement of any one entity could cause major conformational changes.

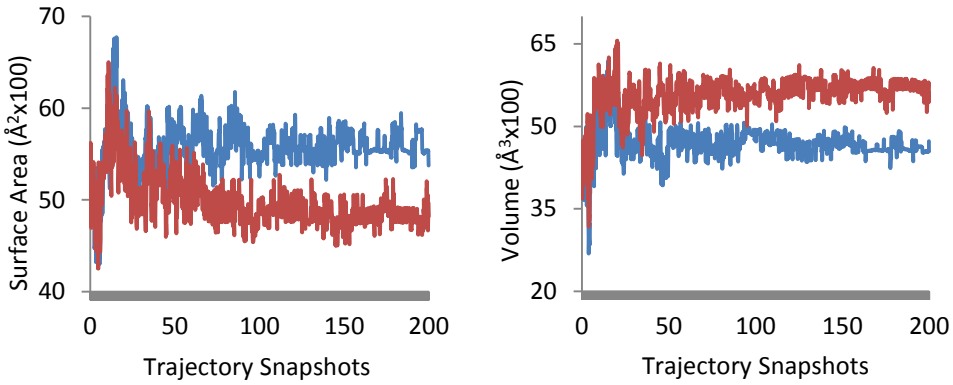

Figure 3.9 Volume and surface area of the tunnel captured using Delaunay triangulation . Blue represents the surface area and volume calculated with the polypeptide and red represents values without the peptide in the tunnel.

3.3.5 Nascent polypeptide ensemble from MD

The nascent polypeptide structures from the ribosome were taken and the various amino acid conformations were detailed. The entire ensemble of the nascent polypeptide is shown in Figure 3.10. This contains 2000 peptide structures, each a snapshot of the trajectory taken at each time step. This shows the various peptide structures that the tunnel can accommodate.

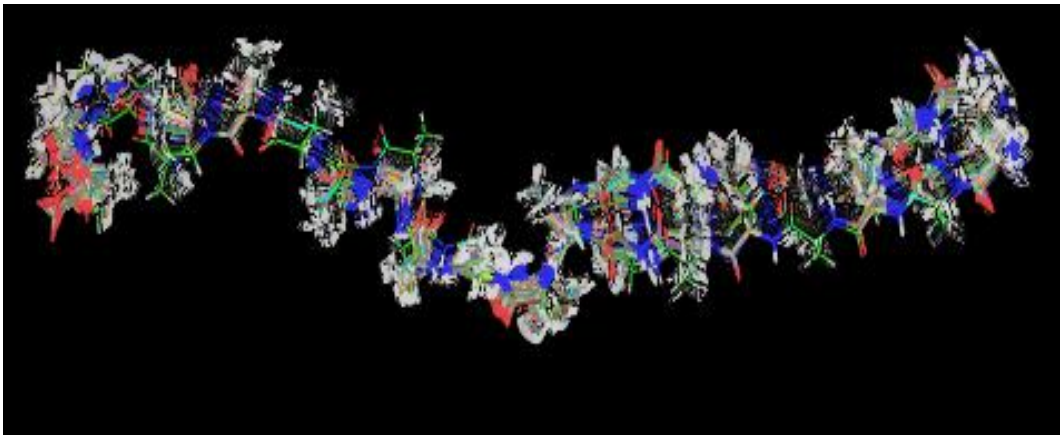

Figure 3.10 Superposed nascent polypeptide tunnel ensemble. These are the superimposed structural snapshots of the nascent peptide in the tunnel from MD.

The Ramachandran map, in Figure 3.11, for the various amino acids shows that not all amino acids enjoy the same conformational freedom and the space available to them dictates their structure locally. The diverse nature of the conformational ensemble shows that the tunnel is very diverse along its length. The nascent polypeptide takes up conformations according to the space available to it. Residues 15, 18 and 19 are completely in the extended β conformation while residues 3, 5, 6, 8, 10, 11, 13 and 14 are in the polyproline type II (PPII) conformation.

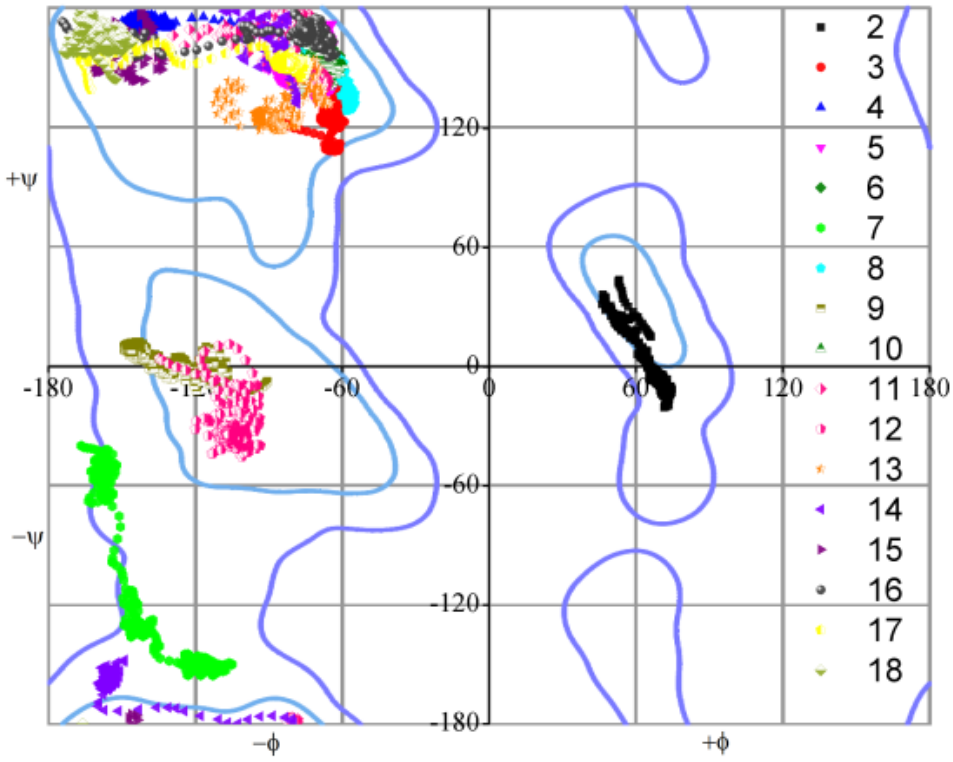

Figure 3.11 Ramachandran map of the nascent polypeptide conformational ensemble. The dihedral angles of the amino acid positions 2-19 in the 20 residue polyalanine peptide are plotted from the MD trajectory snapshots.

Residues 4, 16 and 17 are scattered between β and PPII while residues 9 and 12 are in an α-3_{10} helix structure. Residue 2 is the only residue in the left-handed helix. The most interesting residue in them is residue 7 which does not fit into the preferred region in the Ramachandran map, but is in the allowed range for all amino acids except Pro. We refer to this region as ς. A closer look shows that the residue 7 is close to the tunnel constriction but slightly closer to the PTC. Residues 10 to 13 are present in the constriction zone and are close to various tunnel residues. Residue 8 has close contacts with the ribonucleotide A2058 in the 23S r-RNA. The A2058 is located on the inner wall of the narrowest part of the exit tunnel [240]. The A2058 to G mutation is reported to confer erythromycin resistance. Similar mutations in amino acids close to A2058 in ribosomal proteins L22 and L4 are also shown to relieve elongation arrests. This is explained in the report as relieving a 'jamming-like effect' near the tunnel constriction. Residue 7 is present in the part of the tunnel which does

not provide enough conformation space for the alanine to take up any of its preferred conformations.

If this description of nascent chain conformation in the upper and central tunnel is correct, it implies that each amino acid transitions through the major backbone torsion angle regions of L-amino acid conformational space as it is extruded. For example, an amino acid would have its torsion angle rotated through the following order at each stage of tunnel: α_R, PPII, PPII-β, PPII, PPII, ς, PPII, α_L, PPII, PPII, α_L, PPII, PPII, β, β-PPII, β-PPII, β and β while it goes through various parts of the tunnel.

3.3.6 MC ensemble sampling

3.3.6.1 Sampling around the cryo-EM structure

The nascent polypeptide crystal structure was unfolded and the conformational space around it was sampled using TraDES. The Ramachandran map shows most of the residues taking up very similar conformations even when they are unfolded (Figure 3.12). The degree of freedom allowed by unfolding gives rise to wider distributions. The biggest change is noticed in residue 7, which was earlier in the less-favoured ς region. Residue 7's conformation is almost at the boundary of the right-handed alpha-helical region. This region is not a very favoured region in the Ramachandran plot. Interestingly, residue 12 has also taken up a similar conformation from the previous alpha-helical state in MD.

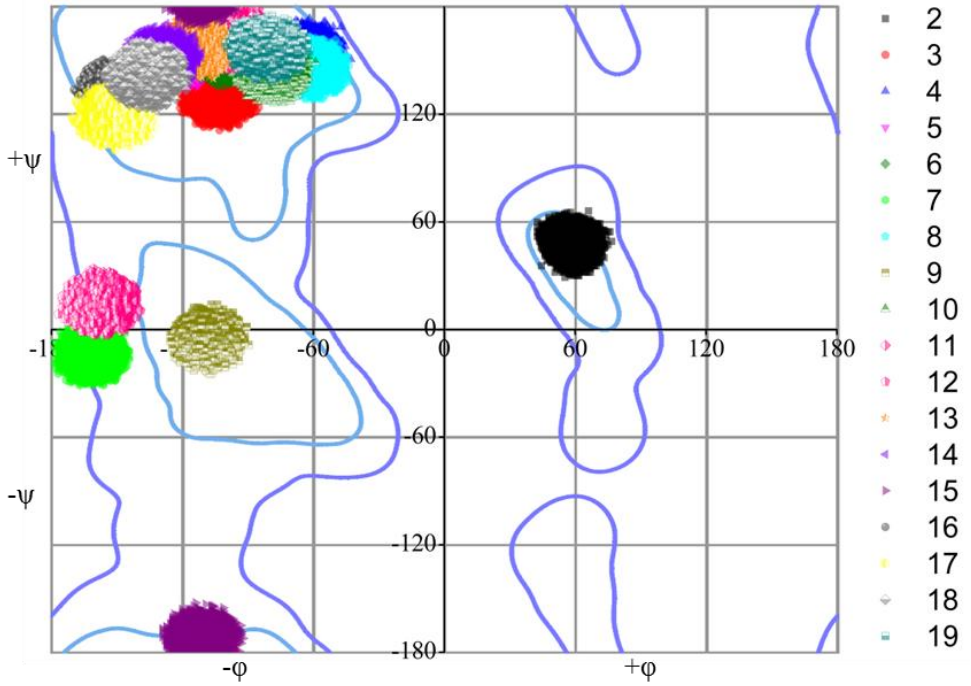

Figure 3.12 Ramachandran map of the NP ensemble from TraDES (MC). The NP was unfolded and sampled by TraDES. This is compared to the conformational space obtained from MD.

Interestingly, residue 12 has also taken up a similar conformation from the previous alpha-helical state in MD. It is important to understand why residues 7 and 12 take up

such non-favourable conformations. 15 out of the 19 residues take up extended conformations based on results from MD and MC simulations. This is expected since extended state is more favourable in such narrow environment. Residues 2 and 9 take up left-handed and right-handed alpha-helical states, respectively. 17 of the 19 residues are consistent and have similar distribution in MD and MC. Residue 7 and residue 12 vary between TraDES and AMBER and need to be analysed.

3.3.6.2 *De novo* sampling

De novo sampling is expected to provide progressively better structures as the number of cycles of steepest descent go on. The final best structures from the *de novo* sampling had a very good quality compared to the native structure. The native polypeptide structure had five clashes when docked with the large subunit of the ribosome. The final run of steepest descent produced structures that had just one clash when it was docked. These produce sets of structures that fit well in the tunnel, even better than the native structure. Comparison of the top three structures with the native polypeptide in the tunnel, show that these are very similar to each other (Figure 3.13).

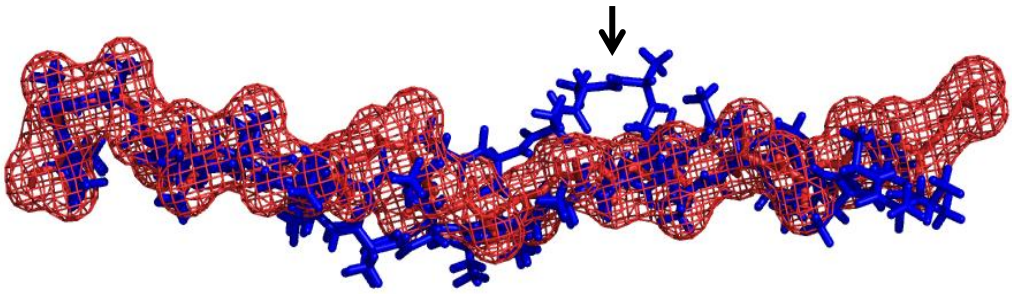

Figure 3.13 Comparing the top results from *de novo* sampling with the native structure. A few structures do not follow the same trajectory or path followed by the native structures. One such example of a local structure deviation is shown by the black arrow.

Evident from Figure 3.13, one of the structures branches away from the native structure and it is not clear how this structure fits into the tunnel. To understand this further, we need to take a closer look at these structures in presence of the tunnel.

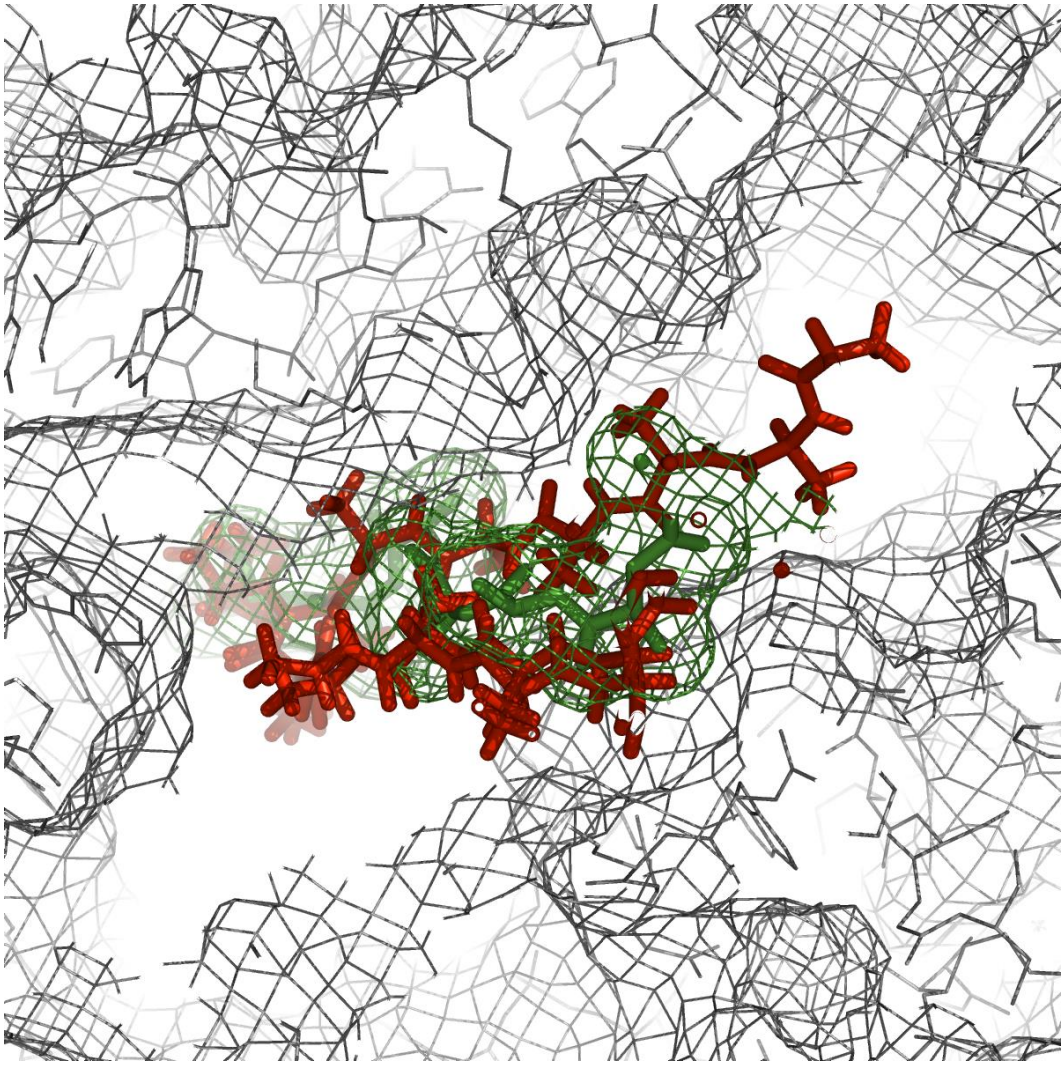

Figure 3.14 Understanding deviations of conformers inside the tunnel. The various conformers (red) take up structures that differ from the general path taken by the native structure (green). This shows tunnel voids are branched and can accommodate multiple paths towards the exit.

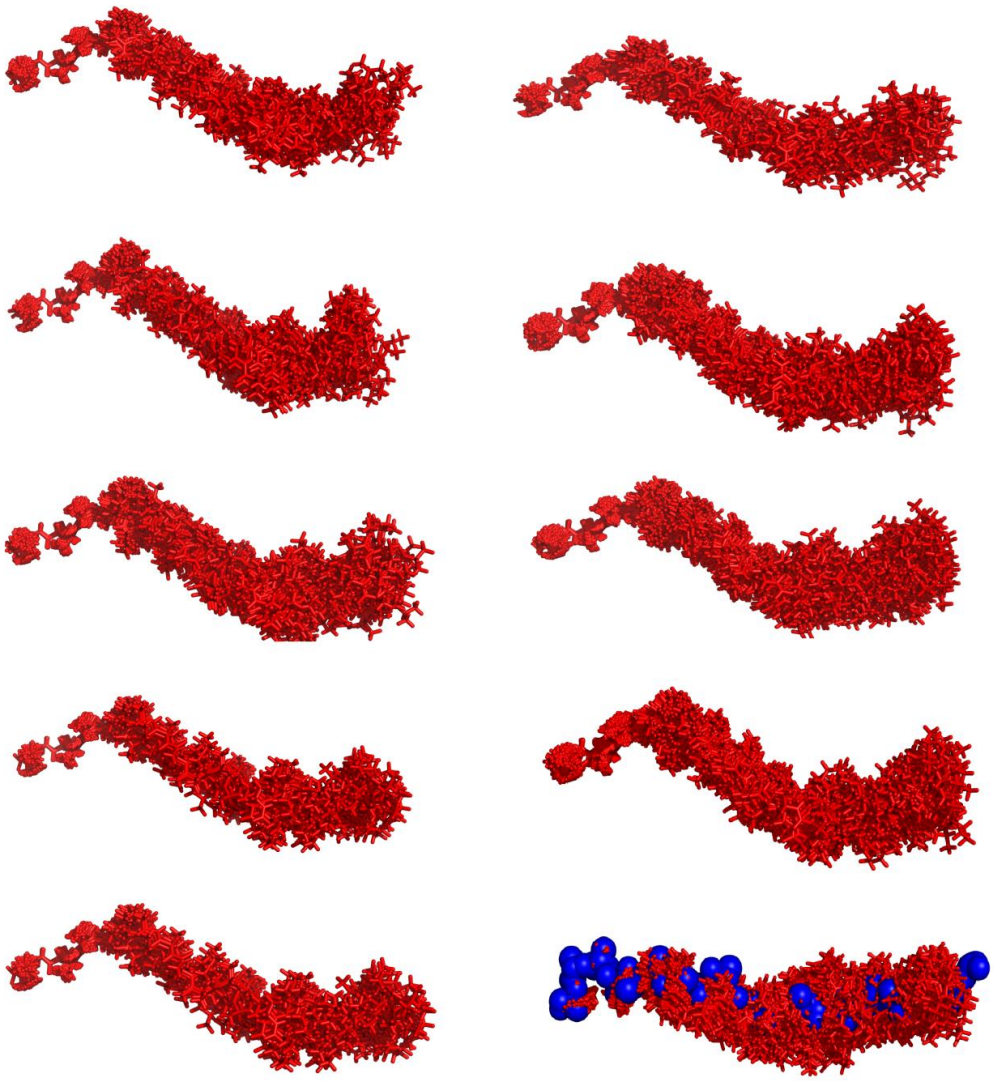

Figure 3.15 Ensemble of structures that fit into the tunnel at every run (side-view). These structures are a direct result of the space available to them in the tunnel at every steepest descent run. The peptides follow a "Z"-like path in the tunnel. The starting segment is constrained, the middle segment is constant and the last segment is wider and funnel shaped. The last figure shows the native structure (represented as blue spheres) superimposed with the final ensemble (red sticks) from Run 10.

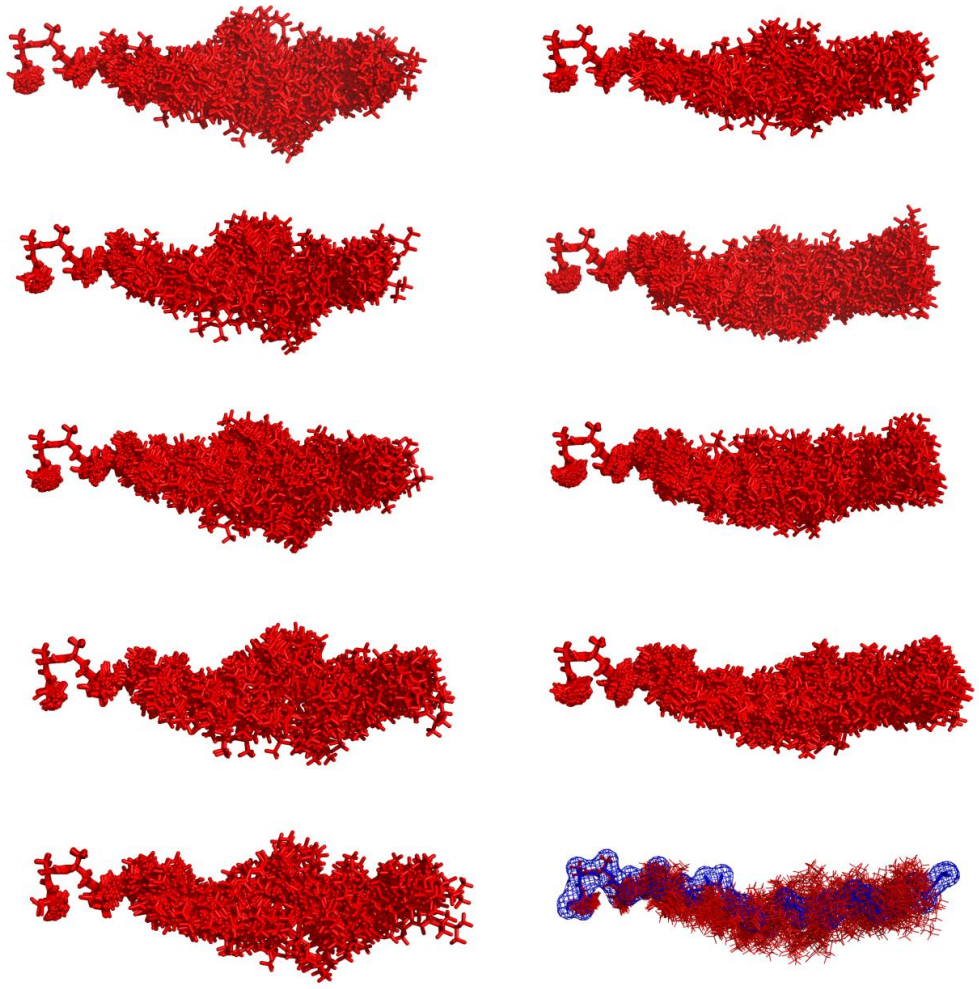

Figure 3.16. Ensemble structures that can fit into the tunnel at every run (top view). This is almost a perpendicular to the view in Figure 3.15. This can be considered as looking from the top, compared to Figure 3.15 which looks from the side. The tunnel is wide in the middle and more structures utilize branches for traversing the tunnel. These wide geometries disappear and the ensembles become more uniform. The final illustration shows structures from the final ensemble (red lines) superimposed over the native structure (in blue mesh and sticks)

It is clear from Figure 3.14 that the structure utilises the branches in the tunnel to explore alternate conformations or paths towards the exit of the tunnel. Looking from the exit of the tunnel down to the PTC, the tunnel seems to wide enough to accommodate different routes for polypeptides to traverse it. The tunnel is irregular and this irregularity is reflected in the ensembles of these structures. These are just three of the best structures taken from the final run of steepest descent. It is expected

that utilisation of the tunnel completely is more of a norm than exception. Due to the large sample size of 10,000 structures for every run, the distribution of the conformations is large enough to represent structures that traverse all the projections comprehensively. The whole set of ensemble structures that can fit in the tunnel are shown in Figures 3.15 and 3.16. These show the two different views of the same ensemble.

Figure 3.15 reports on the space available for the ensembles and in turn the tunnel geometry. The tunnel is 'Z'-shaped, with the first four residues until the first turn. These residues are constrained and hence form a very tight set of structures that are almost perfectly superimposable. The second part of the tunnel is wider but is constant in width from this view. The last six residues form a funnel-like shape due to the increasing space at the end of the tunnel.

An alternate view, perpendicular to the previous view (Figure 3.15), of the same set of structures is shown in Figure 3.16. This can be considered as viewing the tunnel from the top, compared to earlier view of looking from the side. The distribution of structure is different in this view as we see a wider middle portion of the tunnel and is narrower in the exit. Interestingly, the large difference in width between the middle and the end of the tunnel decreases as the minimisation carries on. The final ensemble has a reduced difference, but it is clear that the tunnel geometry on one of its axes is wider in the middle than in the end.

This difference between the geometry of the tunnel across its axes is interesting. Although the immediate effect of this is not clear, it might be acting to induce a spiral motion to push the peptide out of the ribosome. These ensemble illustrations do not provide quantitative information on the changes that take place in the ensembles over minimisation.

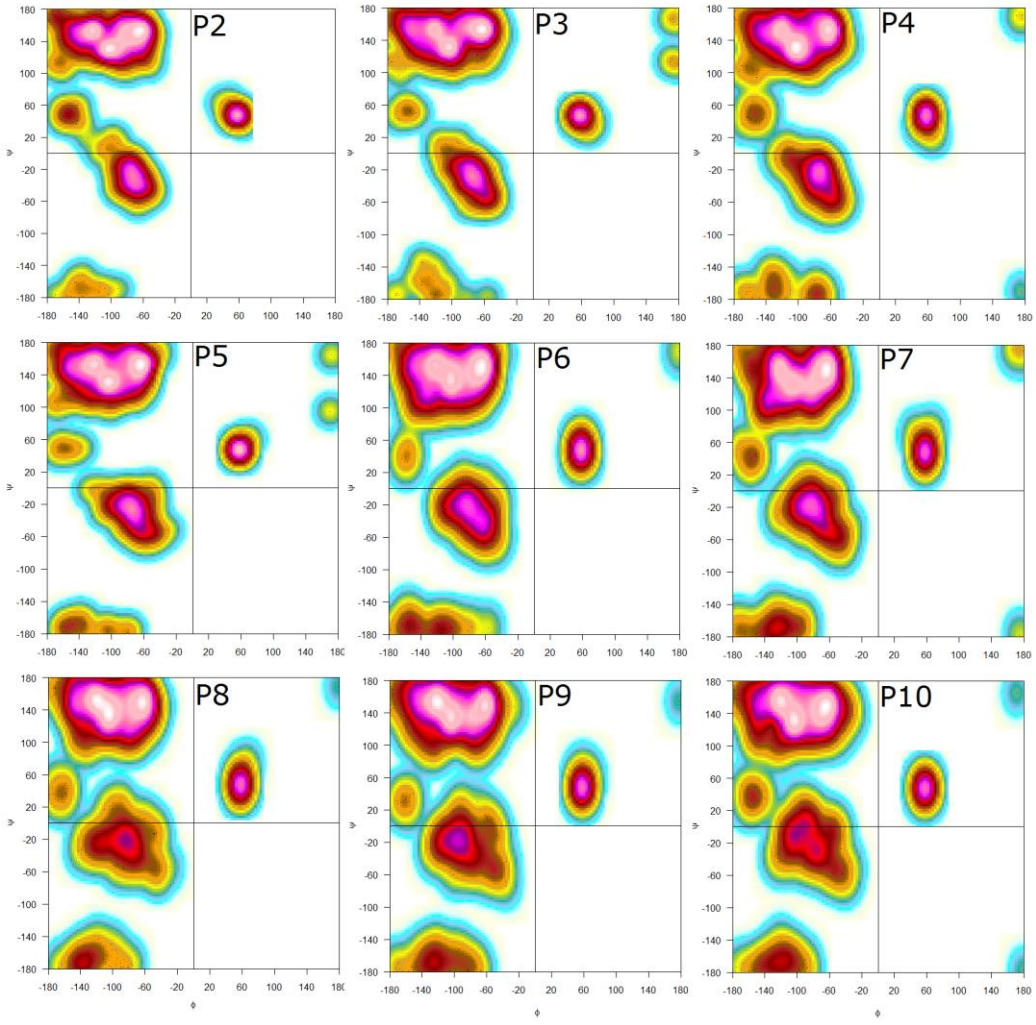

Figure 3.17 Change in conformational space over minimisation runs. At every run of steepest descent of minimisation, the ensemble moves towards a better set of conformations. The Ramachandran plot shows the gradual progression towards increased extended conformations as the minimisation continues. White-cyan-yellow-orange-red-pink-white is the colour gradient used for increasing intensity.

The change in conformation space takes place towards increasingly extended regions in the Ramachandran plot. The extended regions refer to the beta and polyproline type II helical conformations. The P2 conformation is close to the starting standard sampling distribution used in TraDES which has a mixture of alpha and extended states. The psi values of the extended space are largely positive in the starting distributions. The increase of extended space is coupled with the progress of minimisation (Figure 3.17). Unlike the starting distributions, the extended states take up negative psi values in addition to positive values. The alpha-helical regions get

increasingly wider moving away from the core alpha-helical regions. This suggests that minimisation prefers non-helical states in the first twenty residues of the nascent polypeptide.

3.3.6.3 Spatial thresholds

Progressively constraining successive residues result in an increase in tunnel space as it gets closer to tunnel exit. At every threshold, 200,000 structures are generated and docked into the tunnel by superposition. The number of structures that passes the docking at each residue shows the extent of involvement of the tunnel. If we take into account only the first nine residues, the tunnel is absolutely essential since, in the absence of its surface, almost none of the 200,000 structures pass the dock by superposition. This shows that the tunnel is required to actively constrict the residues so that polypeptides can traverse and exit the tunnel.

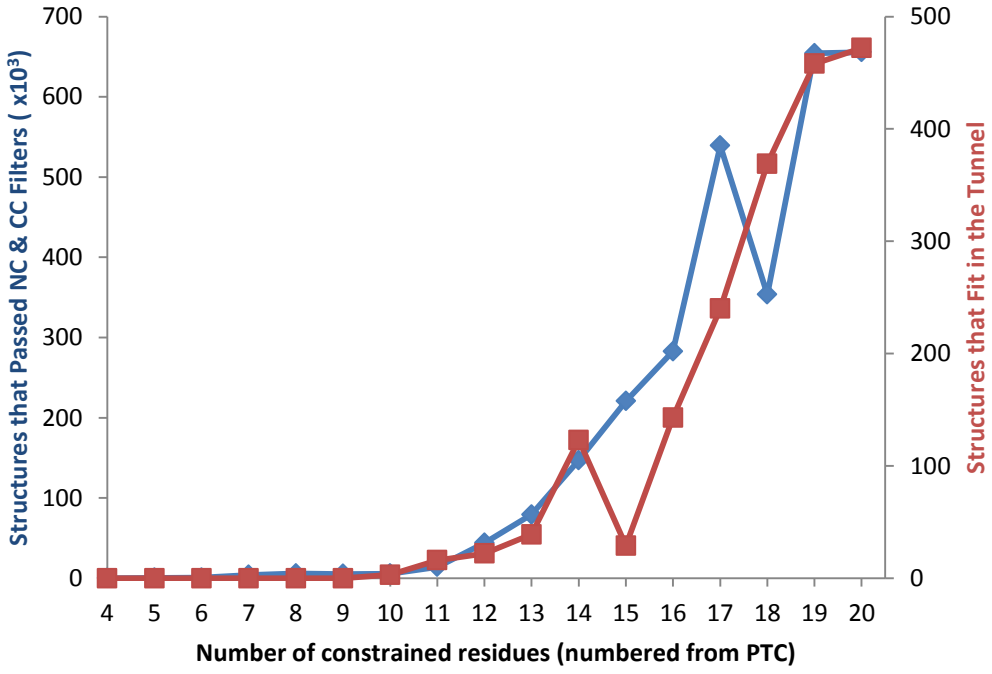

Figure 3.18 Identifying spatial thresholds in the tunnel. The structures that pass the NC-CC filters at every point in the threshold are plotted to show the increasing quality of the ensemble as more of the tunnel is actively acting as a constraint. The structures that fit in the tunnel (red dots) show that at residue 15, the tunnel has a geometry that is averse to accommodating the peptide, in comparison to residues around it.

After the threshold at residue 9, there is a constant increase in the number of structures that pass the docking step except at residue 15. This increase relates to the decreasing need for the tunnel to perform the constraining role. This is a result of the increasing space and freedom available for the peptide as it nears the tunnel exit. At residue 15, there is a clear decrease in the number of structures that fit in the tunnel. It can be expected that the tunnel geometry at residue 15 has limited freedom and could also point away from the exit. There are two kinks in the 'Z'-shaped tunnel seen from the earlier results from *de novo* sampling (Figure 3.15). It is possible that the kink is the reason that the tunnel geometry differs in that region alone. Once past that region, the number of structures continuously increases until residue 20. Although residue 15 acts as a gate which lets very little conformational freedom for the polypeptide, there seems to be no obvious spatial threshold in the tunnel after it. The structures that pass the N-C and C-C constraints are not biologically relevant but show us the capacity of the different tunnel segments to give rise to structures that do no fold back and can traverse the tunnel to the exit. The C-C constraint makes sure that the peptides are pointed towards the exit and the structures that pass the constraints are inclined to traverse the tunnel rather than crash into its walls. Tunnel segment at residue 18 has lesser structures that have passed the C-C distance. This shows that the direction of the segment is pointed away from tunnel exit.

3.4 Discussion

3.4.1 Tunnel and peptide dynamics from MD

A new computational procedure using MD and Delaunay triangulation has been shown to provide detailed insights into the ribosome tunnel and its dynamics. Until now very few simulations have shown us detailed information related to the general dynamics of the ribosome as it has been pretty hard to understand what goes on inside the ribosome. The Delaunay triangulation approach provides a fast and accurate method to evaluate empty spaces inside macromolecules. Since the method is based

on geometry alone, this can be applied to heterogeneous systems. This method is robust and flexible and can be extended to any macromolecular system comprising a cavity within or containing a concave surface. MD simulations have become a powerful tool by itself in providing various details and the Delaunay triangulation gives a refined new way to analyse and draw conclusions from it. The availability of a structure with the nascent peptide has given more depth to the analysis since earlier simulations were done either without the peptide or by building it inside the tunnel [252]. The polypeptide conformations show the amino acid preferences at specific points in the tunnel and indirectly at the freedom at those tunnel segments. The tunnel shows that in the presence of the polypeptide, it has a greater surface area compared to that in its absence. In contrast, the volume of the tunnel reduces in the presence of the polypeptide which could be due to mere presence of the polypeptide and its interactions with the residues, acting like a tether on the either side of the tunnel. This leads to the tunnel becoming more convoluted due to the RNA-protein interactions which explains the increase in surface area and decrease in volume.

From the volume and surface area plots, the volume and surface area of the tunnel seems to increase in the early part of the simulation and is due to the structure relieving itself of any close contacts and steric clashes. The volume and surface energy graphs show the dynamic nature of the tunnel despite the energy graphs showing a very gradual steady trajectory for the entire ribosome. So the tunnel in itself is very dynamic and undergoes a lot of changes.

The ensemble of nascent peptides from the simulation also provides us with new evidence on the conformation of the protein chains inside the tunnel. The tunnel is not even and has shown the presence of a folding zone and gating along its path. A poly-Ala peptide clearly shows the conformational differences along the tunnel as the only difference in residue conformation is due to the steric space available to it at that given point in the tunnel.

The residues 10 to 13 are close to the various amino acids and nucleotides that form the constriction and gating region of the tunnel. Residue 13 in the peptide is close to residue 91, 92 and 93 in the ribosomal protein L22 and residue 67 from the ribosomal protein L4. Residue 12 is close to G91 in L22 and residues 59, 65 and 67 in L4 proteins respectively. Residue 11 in addition is in close contact with the RNA chain B at base 751. It has been shown earlier that the ribosomal protein L22 is a part of the β-turn that protrudes into the tunnel forming the constriction. On the other side of the constriction is the alpha-helix from the protein L4 which also interacts with the peptide (Figure 3.19). The evidence from the simulation is in accordance with earlier reports of the tunnel structure and its constriction. The last few residues are able to take up an extended β conformation due to the wider nature of the tunnel at that position. The first residue is in α_R which may be due to the stalling or is the conformation taken as soon as the peptide bond is formed. The residues in the middle part of the tunnel are mostly in the polyproline conformation while some are distributed between polyproline and β state.

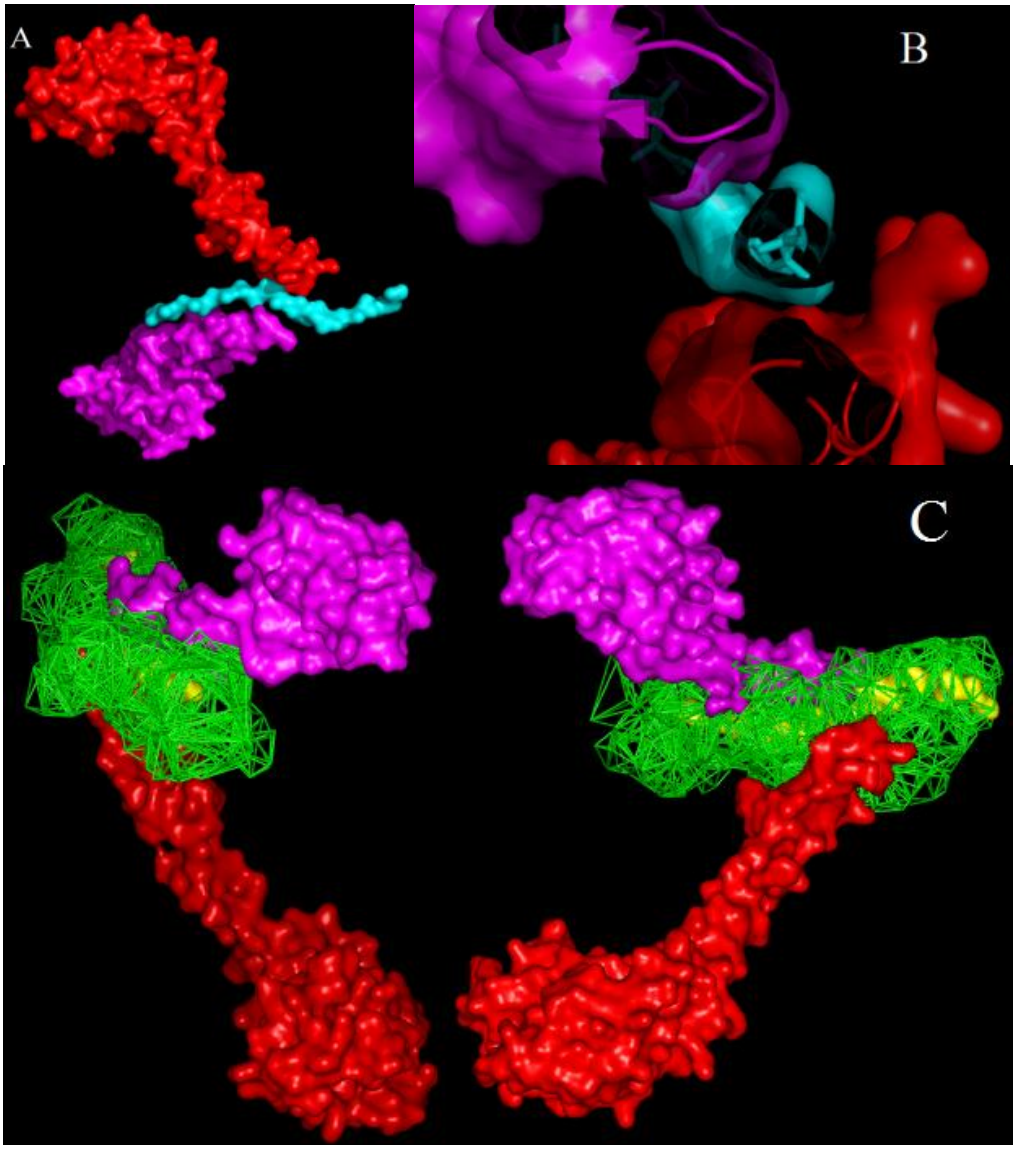

Figure 3.19 Ribosomal exit-tunnel constriction. Top left (A) shows the ribosomal protein L22 (magenta) and L4 (red) forming the constriction in the presence on the nascent peptide (cyan). (B) shows the constriction in detail, where the L22 β-turn and the end of the L4 helix are seen (viewing from inside to outside of the tunnel). (C) shows the constriction of the triangulated tunnel (green triangles) with proteins L22 (magenta) and L4 (red) and the nascent polypeptide (yellow) barely visible inside the tunnel.

The residues also need to pass through the constriction formed by extended loops of L22 and L4. Figure 3.19B shows the narrow space available for the peptide. The extended loops of proteins L22 and L4 seem to interact with the tunnel by creating a DNA-like groove on either side of the tunnel (Figure 3.19C). These loops form the constriction of the tunnel where the residues 10 to 13 are present. Residues 10 to 13

are in the polyproline conformation except for residue 12, which is in α_L structure when they pass the constriction in the tunnel. The residue 7's ς conformation is only marginally stable and is a constraint imposed by the tunnel. When the nascent chain traverses the tunnel constriction it has to go through a sterically-hindered environment which may be a method to gate peptides or this could be an annealing step before the residues are allowed to take up amino acid specific conformations. This could also be related to the discrimination against D-amino acids and/or cis-peptides, but will need further studies along the same lines to provide conclusive insights on this mechanism. Residues 15, 18 and 19, which have passed the constriction, are in an extended β conformation and are present in the start of the lower tunnel, which is much wider than the upper and central tunnel (Figure 3.15, 3.16). This is in correlation with earlier reports of the tunnel's ability to accommodate amino acids in foldable conformations at around 53 Å from the PTC [276], which is close to the distance of residue 18 from the PTC. The distance between the first and the last residue is 56 Å when measured in a line and around 59 Å in a molecular tape measurement. The tunnel considered is only around 68 Å in length which is its upper and central portion while the major folding zones are in the lower tunnel which is wider and contains a much bigger volume. Taken together, the tunnel is a very dynamic part of the ribosome and puts the nascent peptide through a range of conformations induced by local structural bias due to the limited space available in the tunnel.

This MD study does not reflect the translation motion of the ribosome since there was no force applied to the polypeptide for extrusion. Rather, this elucidates the available geometry in different parts of the upper tunnel for the polypeptide. Since the chain is polyalanine, the local geometry is dictated by the tunnel and its dynamics alone. The lack of explicit water and ions may also contribute, in part, to the lower surface area of the tunnel the absence of the peptide. This simulation may only be a representation

of the local conformational changes in the ribosome, as described by [251]. Large motions of the ribosome could also contributed to the volume and surface area changes observed in the ribosome, but these motions are mostly absent in the current study. If the entire dynamics of the tunnel and polypeptide is purely random vibration, we should observe an even spread in the Ramachandran map for all 20 residues in the chain. The differing individual conformations of the various positions along the chain show that these are in fact due to the local structural bias provided by the tunnel. It would be realistic to say that during translation, the individual residues might need to adopt different conformations dictated by the local tunnel geometry.

3.4.2 Unfolded polypeptide sampling

Ensemble sampling and dock by superposition provide a great tool to understand geometries that cannot be studied in detail using crystallography or NMR. This is a very novel pipeline technique that can solve the simple problem of investigating such enclosed geometries. Crystallography is a snapshot technique, where an ordered immobilised protein structure is solved. But, it is quite well known that protein and biomolecules are rarely static. Even if the high resolution structure of the nascent polypeptide inside the tunnel is solved tomorrow, it will only provide one possible conformation adopted by the protein in the ribosome. Ensemble sampling and dock by superposition are very simple techniques individually, but put together, they provide a great pipeline for understanding confined spaces. The time efficiency of this method is also largely greater compared to a molecular dynamics simulation. The coverage of a MC-based method is always higher than a MD approach but a time-resolved model cannot be obtained. MD simulation has provided good insights on the dynamics but a very general idea of the conformations. Figure 3.10 shows the narrow range of similar structures that have been identified by MD. In contrast, structural snapshots of the ensemble structures that pass docking alone have a very wide distribution (Figure 3.15 and 3.16). Despite all these advantages, this is not a very

popular technique and hence we need to apply caution before we can safely interpret these results. This was checked and re-checked with a large number of structures and manually curating clashes by looking at every docking result. These showed a very good reliability and no errors popped up. To visualize the effectiveness of this method, a few structures that passed these constraints are illustrated in Figure 3.20 below. The nascent polypeptide conformers (in red) that pass the docking tool fit in the tunnel well and do not have any bad contacts with the tunnel walls (green mesh).

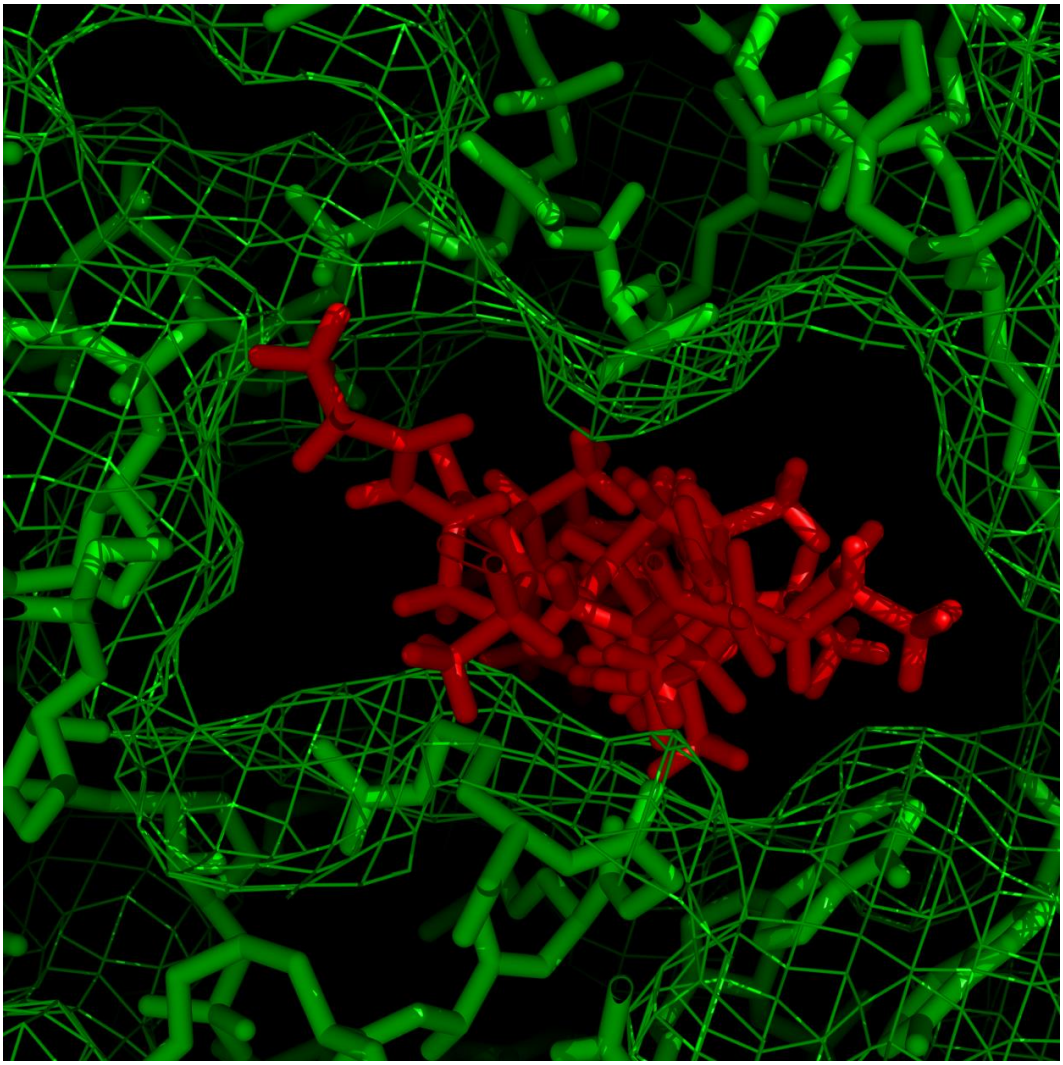

Figure 3.20 Illustrating the effectiveness of the filtering algorithm. A few conformers that pass the NC-CC filters and selected by the dock by superposition are shown in the tunnel in a cross-section view (looking from the exit towards the PTC) to show the effectiveness of the method. None of the conformers (red sticks) have any bad contacts with the tunnel (green mesh and sticks).

Some chains extend and traverse various branches in the tunnel. The tunnel is known to have a lot of branches and it is still unclear how these branches contribute to the protein moving along. These provide a good idea of how these branches indirectly affect the constrictions in the tunnel. A tunnel in its most constricted state will have a smaller volume and no branching, and a more accommodating part of the tunnel will have branches that can help the protein move along with a less restrictive environment.

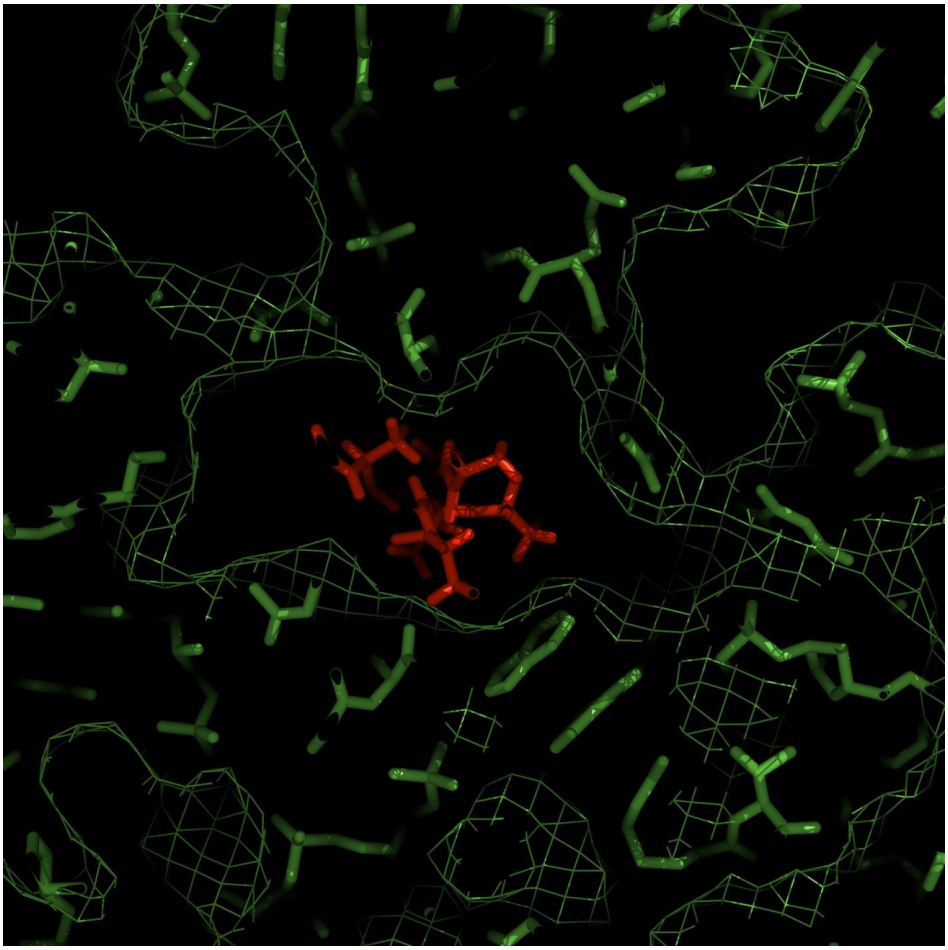

Figure 3.21 Irregularity of the tunnel surface acting as a sampling constraint. The irregular geometry of the tunnel (green mesh and sticks) is shown along with the effect it has on the conformers (red sticks) that sample it. Conformers are shown to sample branches and grooves formed by the irregular geometry of the ribosome tunnel. This also indicates the effectiveness and advantage of using the tunnel geometry as a constraint for peptide sampling.

A cross-section of the tunnel shows us a clearer picture in which the tunnel cavity is highly irregular and the peptides stay well within the cavity (Figure 3.21). This is

again an indirect proof of principle for the MC-docking pipeline which results in a set of structures that are constrained by the geometry of the tunnel, as it happens *in vivo*. Using the actual tunnel surface and atoms as constraint instead of a rolling ball or a cylindrical constraint gives us an accurate sampling of the polypeptide conformations.

3.4.3 *De novo* sampling

The MD and MC samplings based on the cryo-EM structure give great insights into the capacity of the method and the large ensemble of structures that can fit in the tunnel. Upon close observation, the structures are all closely related and can be clustered into a small group of dihedrals. It was observed in an earlier report that dynamics in such simulation are largely due to thermal fluctuations and local conformational changes that happen around the crystal structure [251]. This results in local sampling around the energy minima and hence it is incorrect to assume that the entire space inside the ribosome can be described by this alone. The comprehensive solution to this would be to start from a clean slate and scan the complete set of conformations available to the peptide and filter ones that can geometrically fit inside the tunnel. Although comprehensive, it is practically impossible to completely scan the structural and energetic landscape of the protein, even if it is only 20 residues long. Even if the whole landscape is represented, it is exponentially harder to filter and validate it. A very close approximation is to look only at the geometry and not energy as the minimisation property. Generally, energy is a great property to minimise since lower energy directly implies a better structure. But energy computation is very intensive and impossible for a large set of structures. Since we are more focused on the geometry and complete characterisation of the tunnel and its peptide, an alternative possibility is to use the quality of structures that fit into the tunnel. As interactions inside the tunnel are not well understood, it might be better to use geometry than energy, since energy needs to be well modelled to provide an accurate description. Minimising geometry also has its advantages as the peptide can

be modelled independent of the ribosome and can be fitted and filtered to increase efficiency. A brute-force approach might require a large ensemble to encompass all possible conformations. A steepest descent approach is used to overcome this problem by refining the ensemble at every step. Steepest descent is a very popular minimisation technique and has been applied largely in the energy minimisation step of MD. It was also been applied in analysing such large scale data in different biological contexts such as mining signal transduction networks [277]. This reduces the size of the ensemble at every step and improves time efficiency. This increases the probability of sampling accurate structures at every step. In addition to the steepest descent, NC and CC filters increase sampling accuracy by eliminating grossly inaccurate structures. Together, these small simple improvements make it possible to sample a large part of the nascent polypeptide conformational space.

It is possible to infer the tunnel space by looking at the ensemble alone. A visual inspection of the tunnel will provide details on the amount of branching and features of the tunnel. Two different views showcase the complex structure of tunnel (Figure 3.15 and 3.16). Along one axis, the tunnel is wider at the exit and narrower in the middle, and vice versa along another perpendicular axis. This is explained by the branching of the tunnel differently. The tunnel is heavily branched but not all branches can be expected to accommodate residues.

In the triangulated space of the tunnel we can see the two main branches that affect the ensemble. Although it is not clear from the figure, the two branches are along different axes and each one is responsible for the wider geometry of structures in different axes. This is a purely geometric view and the branches may or may not be accommodating based on their chemical nature to form or disrupt different types of interactions. The composition of the tunnel is expected to play a significant part in deciding if the NP will be harboured by the branch. This requires a very complex

model with various interactions that happen in the tunnel, which are not very well understood at this point of time.

The Ramachandran map over the course of the minimisation shows the structures moving towards a more extended state with every cycle of the steepest descent minimization. The absence of preference over negative or positive psi values show that they move along a general direction rather than a specific polyproline type II or beta conformation.

An interesting result was the identification of different paths taken by the various conformers in the ensemble within the tunnel. This shows the peptide has a lot more freedom than reported earlier. Figure 3.22 shows two such peptides that take up different conformations along the tunnel. The tunnel branches provide variations of broad and narrow spaces along the way. As the peptide moves along those specific parts of the tunnel, it takes up more dynamic or stricter conformations locally. It can be inferred from Figure 3.22 that the first part of the tunnel is very narrow and hence most of the ensemble takes up a small range of conformations while, at around 40 Å from the PTC, it widens and the peptides can take up a bigger range of dihedrals. Interestingly, again the tunnel narrows and we see that the range of conformations has decreased. This can be quantitatively viewed from the Ramachandran plots of the individual residues.

It would be interesting to check how the lower tunnel behaves given that there are few experimental reports of tertiary interactions [70] observed in it. It is tempting to say that the conformational ensemble in the lower tunnel would be much more diverse. The sequence might actually dictate the conformations in the lower tunnel unlike what we have observed in the upper and central tunnel in the current study and could possibly give rise to a few basic secondary structures.

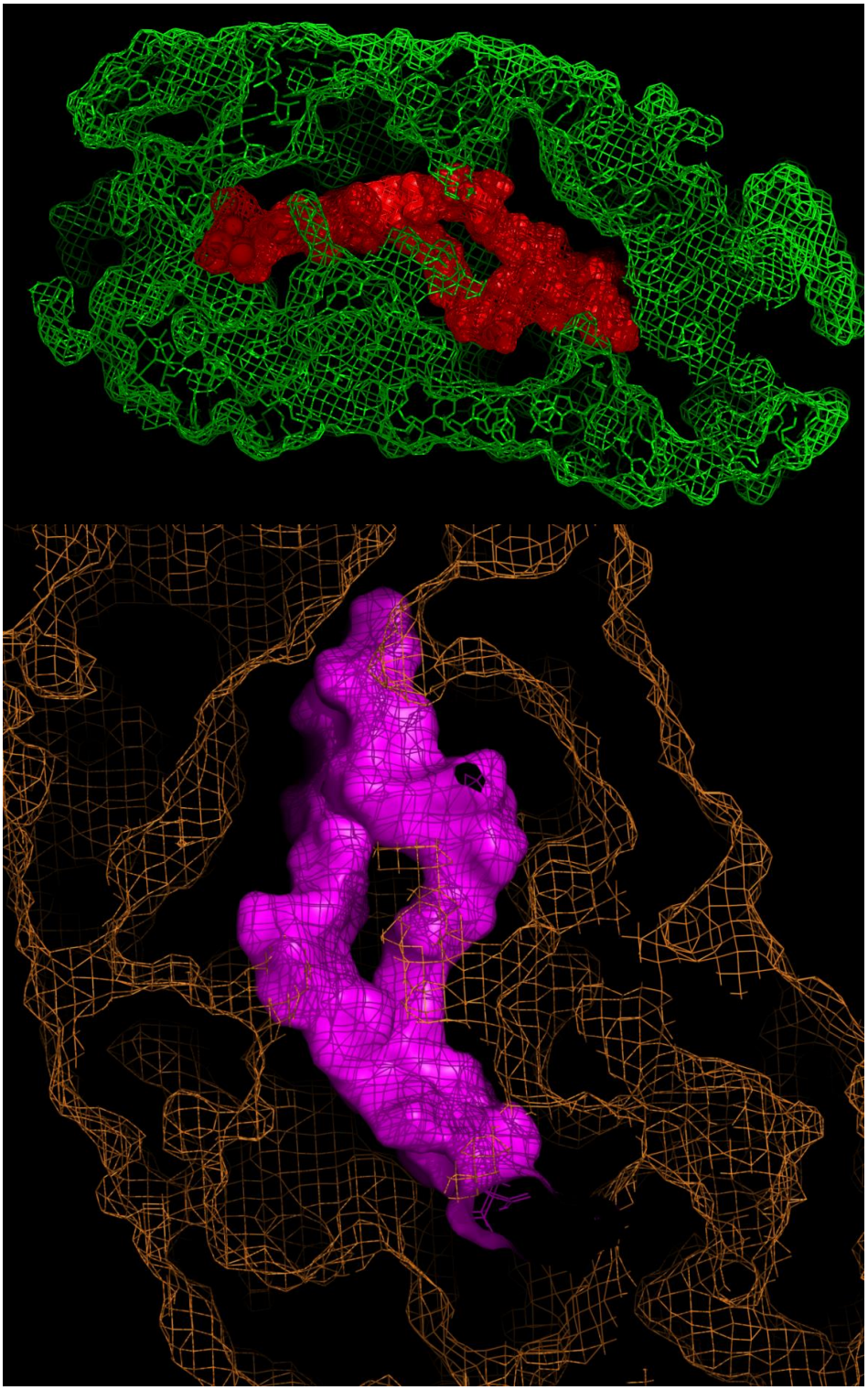

Figure 3.22 Different paths are available for the nascent peptides in the tunnel. The tunnel is able to accommodate peptides that take up different tracks along the tunnel. This may vary on the amino acid compositions; smaller amino acids might have larger freedom than bulkier amino acids.

The main objective of the *de novo* analysis was to ascertain if the nascent polypeptide ensemble generated by unfolding of the nascent polypeptide was a complete subset of the space in the tunnel. Comparing the dihedral distributions of the ensemble from *de novo* and unfolded sampling clearly show that the cryo-EM structure is only a very small set of the structures that can fit in the tunnel. The unfolded structures mostly are subsets of the *de novo* dihedrals. Three residues have completely different distributions. It is clear that each of the structures passes the geometric constraints of the tunnel and hence these dihedrals are a direct consequence of the space available at those positions in the tunnel. The first four residues are constrained (refer Methods) and hence their dihedrals are constant among the ensemble. There is a subtle yet discernible increase in the width and spread of the distributions as we move away from the PTC and closer to the exit. All the residue positions, except residue 7, prefer to populate the extended regions of the Ramachandran map. This reinforces earlier results of the nascent peptide largely preferring to traverse the tunnel in the extended state. It is also evident from Figures 3.27 and 3.28 that the later part of the tunnel also accommodates helical states. Residues 12 and 14 to 19 show propensities for the right-handed alpha-helical conformation in addition to the extended state.

Residue 15 is unique since it has a lot of freedom to take up most of the standard conformations available for amino acids including the left-handed helix, which is not common in native proteins. Residue 15 was earlier shown to act as a spatial threshold since providing the cryo-EM dihedral for MC sampling resulted in a worse ensemble than leaving it unconstrained. The cryo-EM dihedral only contained a small set in the extended regions and this elimination of the helical state from the sample space is possibly the reason why the quality of the ensemble turned out to be worse.

Residue 7 and 12 took up unfavourable positions in the MD and unfolded sampling. Residue 7, even in *de novo* sampling, is found in a similarly unfavourable region and confirms the early finding that the space in the tunnel is not permitting it to take up

favourable conformational states. Residue 12 was also implicated in the non-favoured region by MD, but *de novo* sampling shows it can take up the favourable extended and alpha-helical conformations in structures that fit the tunnel. A possible reason for discrepancy could be the presence of residue 12 in the constriction point formed by L22 and L4. It is also interesting that residue 12, which is the closest to the PTC, to have helical propensity. The conformation of residue 12 and its tunnel geometry is not clear immediately but requires more insight.

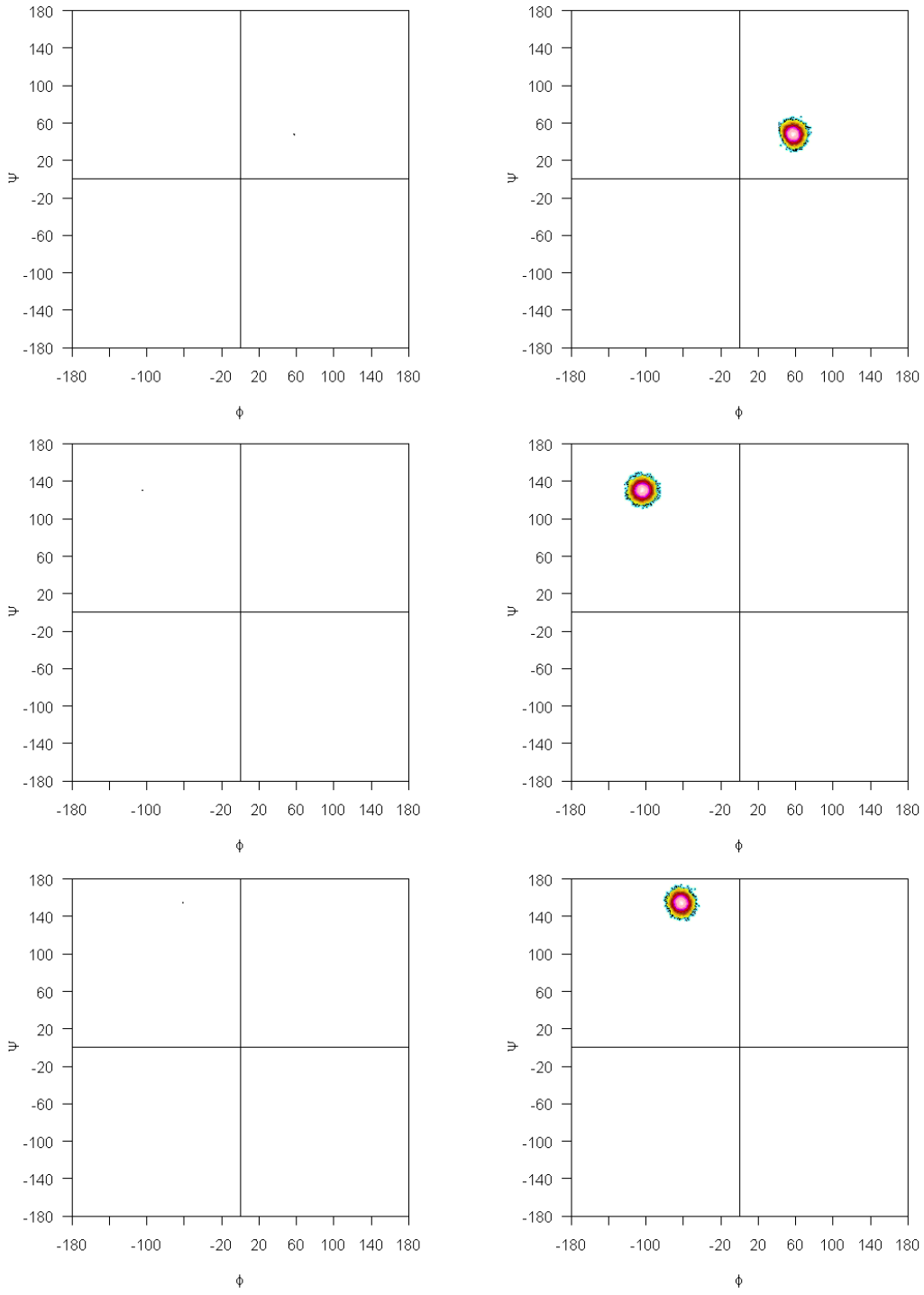

Figure 3.23 Unfolded sampling is only a subset of the complete NP space (residues 2, 3 and 4). The resultant space from *de novo* sampling (left) starting from the full conformational space gives a more complete sampling compared to the cryo-EM unfolded space (right), which is only a small set of all possible structures inside the tunnel, based on its geometry. These represent Ramachandran maps for residues 2, 3 and 4 for the first, second and third rows, respectively.

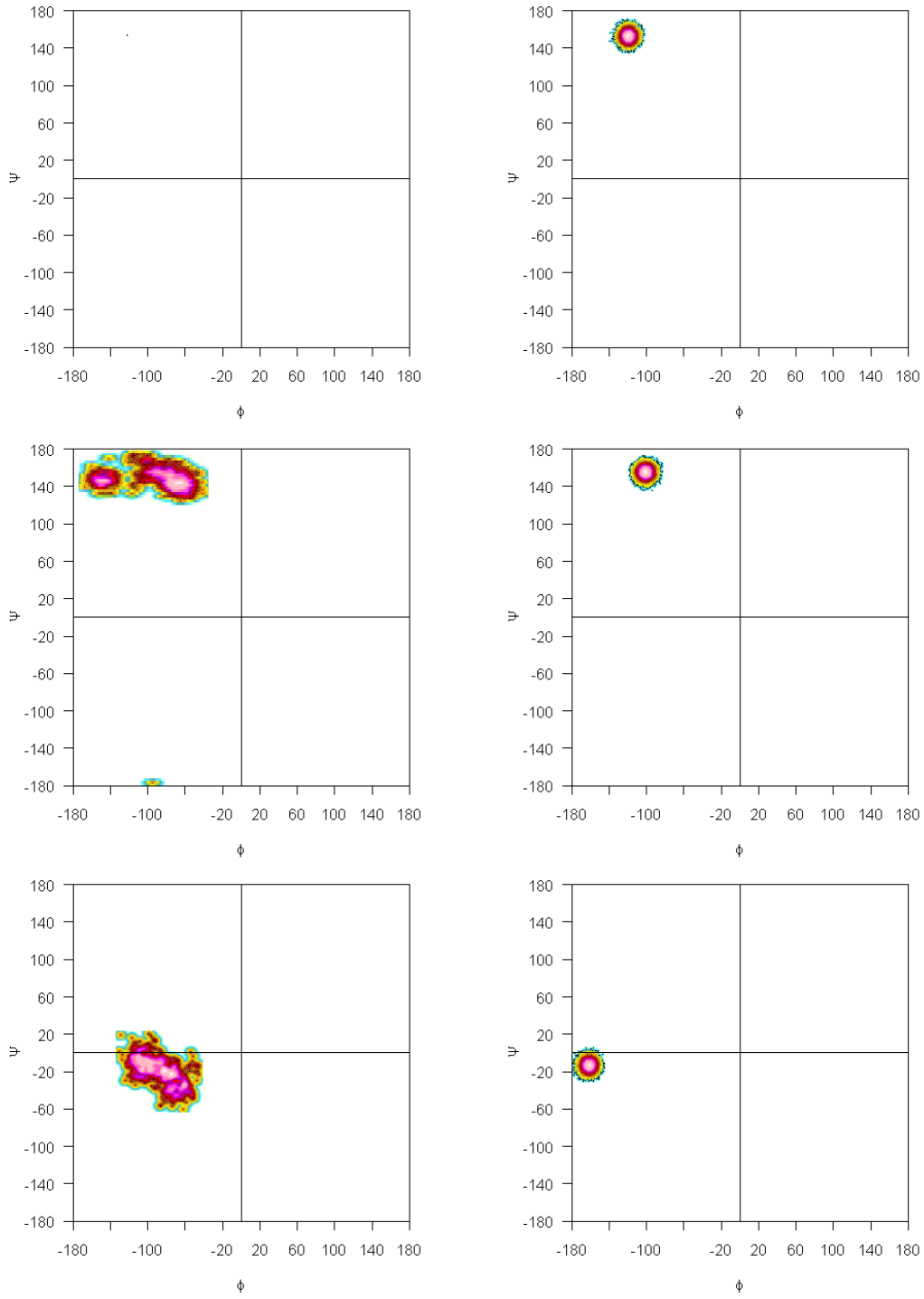

Figure 3.24 Unfolded sampling is only a subset of the complete NP space (residues 5, 6 and 7). The resultant space from *de novo* sampling (left) starting from the full conformational space gives a more complete sampling compared to the cryo-EM unfolded space (right), which is only a small set of all possible structures inside the tunnel, based on its geometry. These represent Ramachandran maps for residues 5, 6 and 7 for the first, second and third rows, respectively.

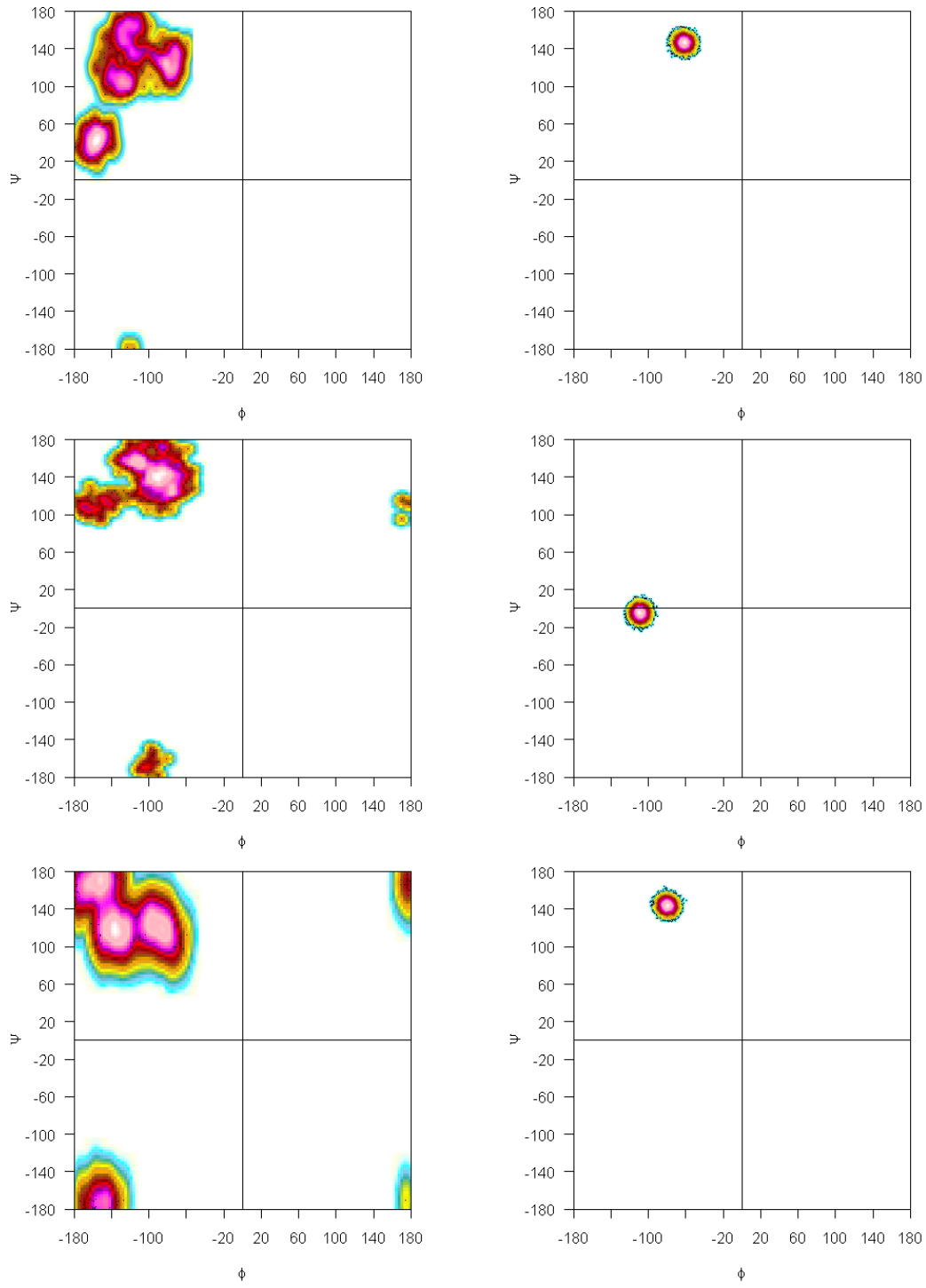

Figure 3.25 Unfolded sampling is only a subset of the complete NP space (residues 8, 9 and 10). The resultant space from *de novo* sampling (left) starting from the full conformational space gives a more complete sampling compared to the cryo-EM unfolded space (right), which is only a small set of all possible structures inside the tunnel, based on its geometry. These represent Ramachandran maps for residues 8, 9 and 10 for the first, second and third rows, respectively.

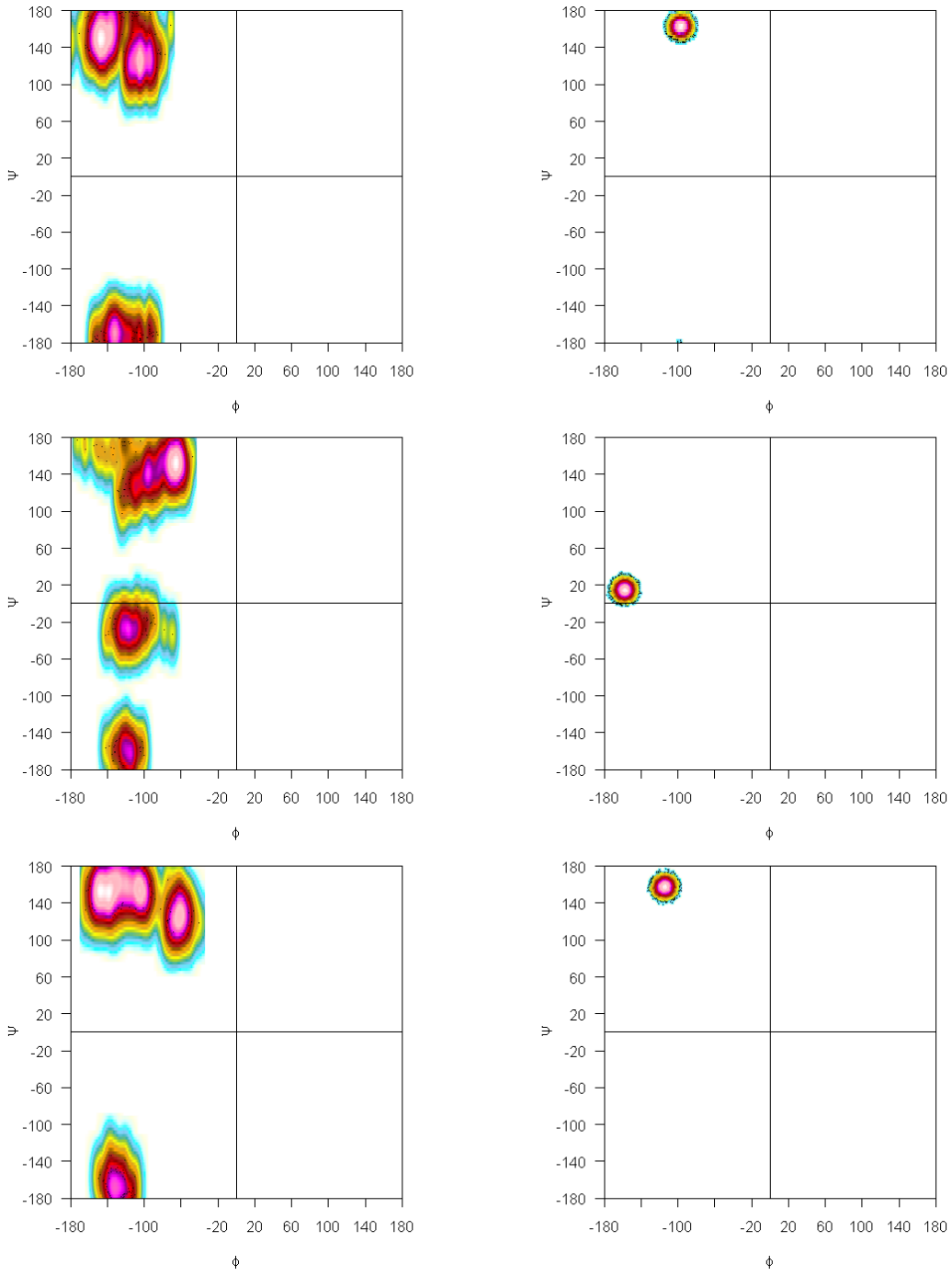

Figure 3.26 Unfolded sampling is only a subset of the complete NP space (residues 11, 12 and 13). Unfolded sampling is only a subset of the complete NP space. The resultant space from *de novo* sampling (left) starting from the full conformational space gives a more complete sampling compared to the cryo-EM unfolded space (right), which is only a small set of all possible structures inside the tunnel, based on its geometry. These represent Ramachandran maps for residues 11, 12 and 13 for the first, second and third rows, respectively.

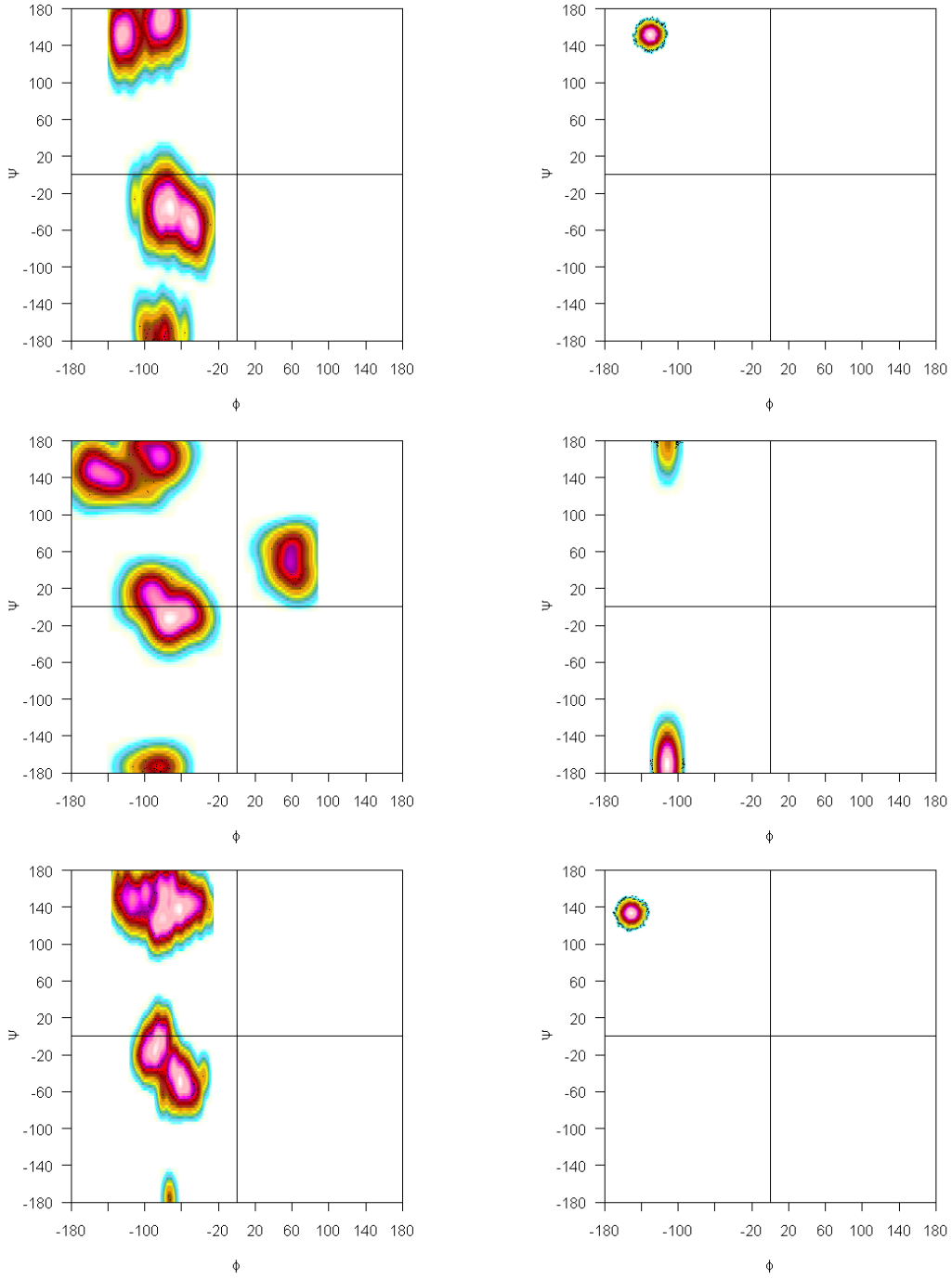

Figure 3.27 Unfolded sampling is only a subset of the complete NP space (residues 14, 15 and 16). The resultant space from *de novo* sampling (left) starting from the full conformational space gives a more complete sampling compared to the cryo-EM unfolded space (right), which is only a small set of all possible structures inside the tunnel, based on its geometry. These represent Ramachandran maps for residues 14, 15 and 16 for the first, second and third rows, respectively.

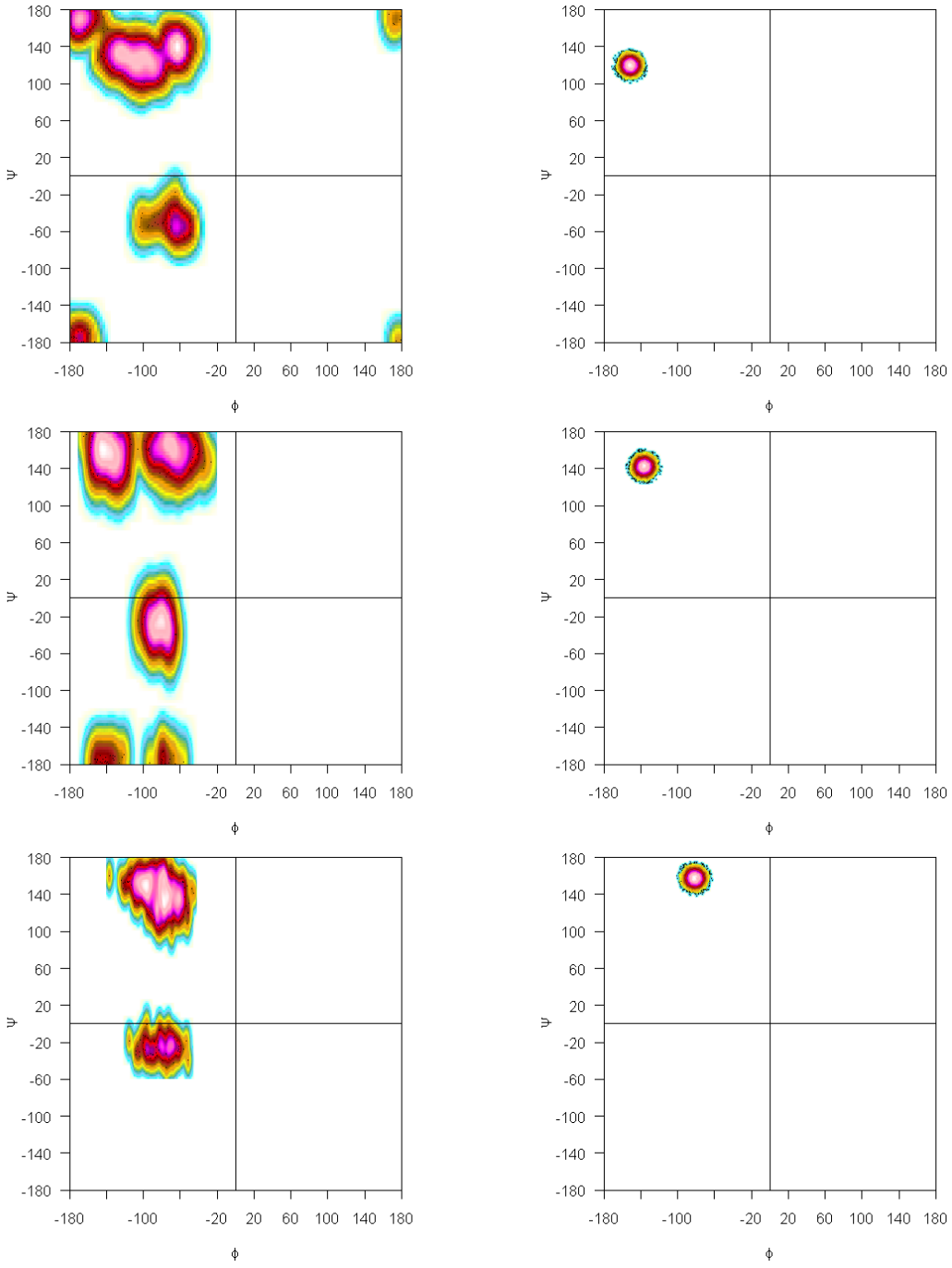

Figure 3.28 Unfolded sampling is only a subset of the complete NP space (residues 17, 18 and 19). Unfolded sampling is only a subset of the complete NP space. The resultant space from *de novo* sampling (left) starting from the full conformational space gives a more complete sampling compared to the cryo-EM unfolded space (right), which is only a small set of all possible structures inside the tunnel, based on its geometry. These represent Ramachandran maps for residues 17, 18 and 19 for the first, second and third rows, respectively.

3.4.4 Spatial thresholds

The spatial threshold can be understood to be a specific part of the tunnel after which the tunnel is least required to act as a constraining surface. The tunnel provides a passage for the nascent polypeptides to traverse the ribosome. It is known that the tunnel is very narrow closer to the PTC and wider closer to the exit [71] and is irregular. Hence, it can be expected to contain some narrow restricting regions followed by regions with a lot of spatial freedom.

There are two points of discussion based on the docking results and the filtered structure results. The docking results show a steady increase in the number of structures that fit inside the tunnel. In the absence of thresholds, a steady increase is expected. This is due to the fewer residues that need to be fit in the tunnel. Compared to the threshold at 15 Å, the threshold at 25 Å is expected to give more structures that can fit in the tunnel as there are fewer remaining residues that need to be sampled. There is also increasing freedom in the tunnel as we get closer to the exit. In sampling terms, small deviations in conformations of residues farther away from the tunnel result in much poorer structure than those closer to the exit. In other words, a small conformation change in residue 5 will have a more pronounced effect than residue 15 since residue 5 influences the direction of 15 remaining residues while residue 15 influences only 5 remaining residues. The biggest advantage is that all of these properties are common to the whole model and their collective result is the increase in the number or structures that passes docking. On the other hand, the identification of a threshold is made easier due to its deviation from the normal increase observed because of the above properties.

In the docking results, the tunnel is absolutely necessary until residue 9 since sampling 2 million structures does not give rise to any structures that can completely traverse and fit inside the tunnel. Statistically, the probability of a structure that can exit the ribosome successfully, if only the first part of the tunnel acts as a constricting

volume, is almost non-existent. There is a constant increase after residue 9 and goes until residue 20, except residue 15. To understand what happens when residue 15 was constrained, the structures that passed the constraint and those that did not were analysed with respect to the tunnel. The tunnel space is not completely accessible to the peptide and most of the conformers generated fall outside the tunnel definition at that position. Looking at the *de novo* sampling Ramachandran map, it is clear that the tunnel also accommodates left-handed alpha helical conformation due to its geometry. Although this left-handed alpha-helical conformation is possible at that tunnel position, it also affects the conformation of further residues. It is seen that the remaining residues turn at different angles inside the tunnel which cause them to not fit inside (Figure 3.29). The reason for the deviation of residue 15 from the increasing pattern is not because of constraining geometry but rather due to additional freedom. This is also reiterated by the Ramachandran maps where the following residues 16-20 still follow the usual extended conformations. So, a helical conformation with one residue in a line of extended residues will result in deviating from the tunnel and hard sphere clashes with the tunnel residues.

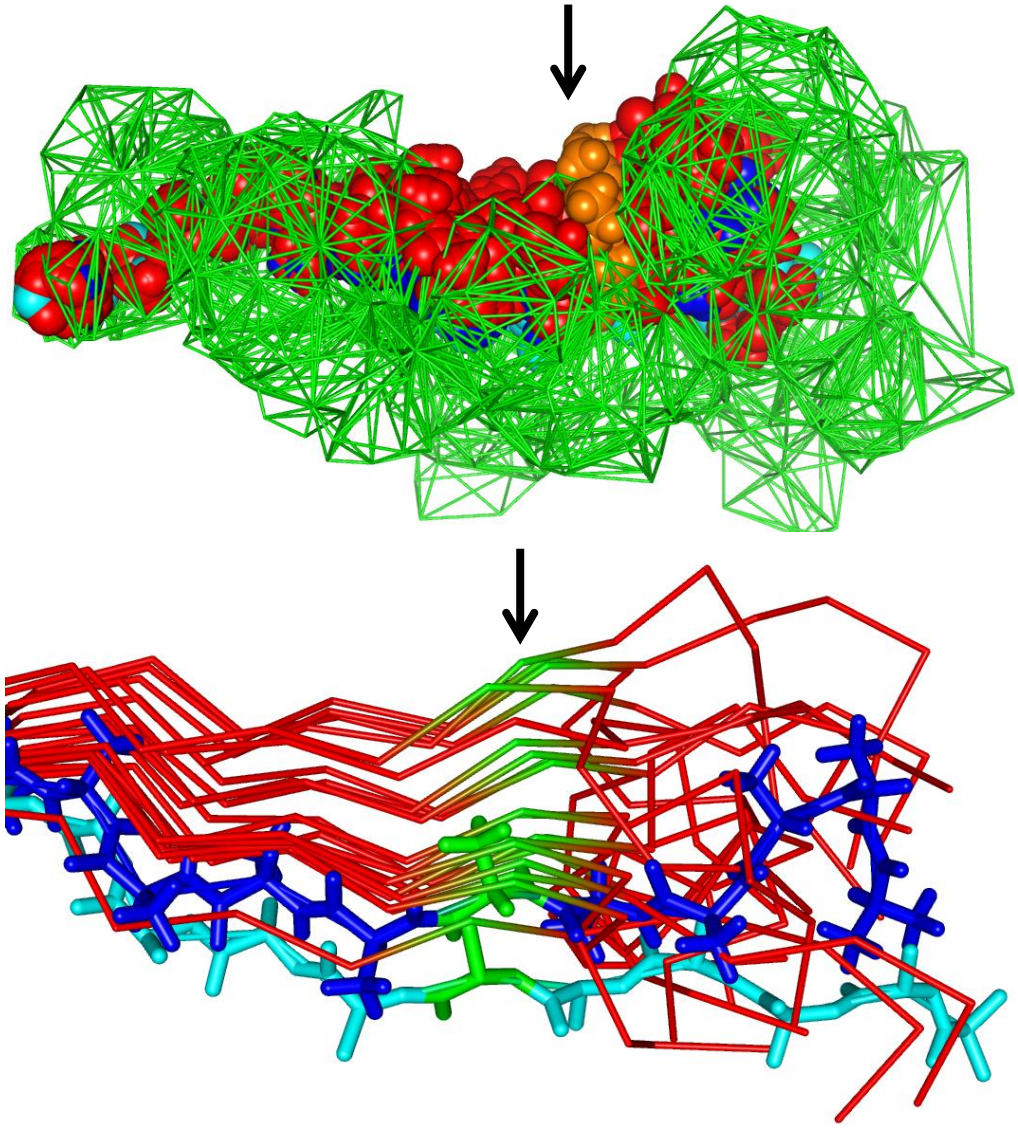

Figure 3.29 Threshold at residue 15 analysed by structures generated by constraining residues 1-15. At the top, the tunnel (green) is shown in superposition with structures that pass the filters (blue) and those that do not (red). Residue 15 is highlighted in orange and the cryo-EM structure is shown in cyan. The bottom illustration shows the same structures in ribbon representation with residue 15 in green. Once passed the constraint at residue 15, the structures fold within the tunnel and fail the filters.

Results from NC and CC filters show that residue 18 has a decrease in number of passed structures. Interestingly, this decrease is not mirrored in the docking results. Visualising the filtered out structures with the passed structures, we can conclude that the failed structures have conformations in one particular direction and has collisions with the tunnel surface.

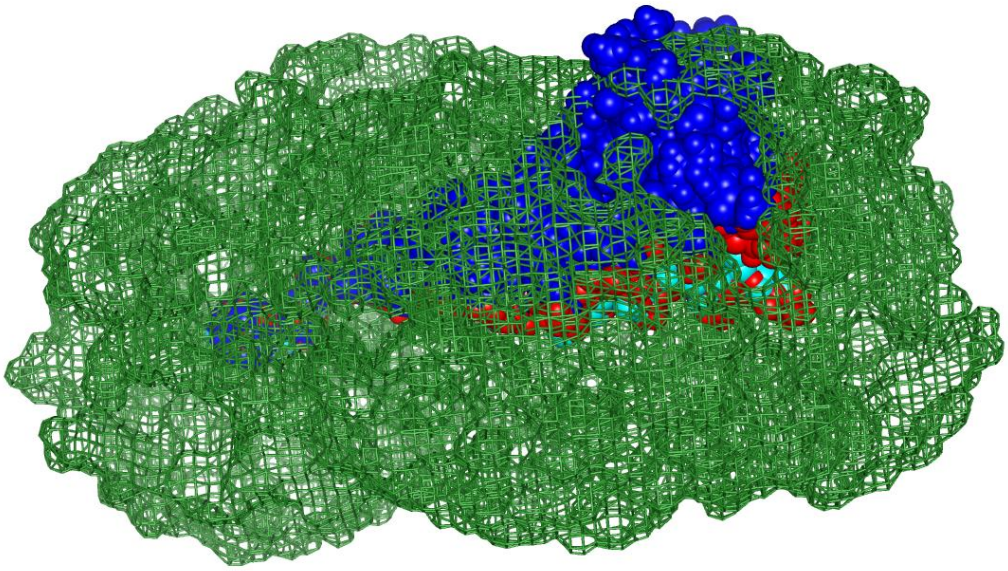

Figure 3.30 Tunnel position at residue 18. Constraining this gives rise to structures (blue and red) that are simulated from the cryo-EM structure (cyan) in the tunnel (green). The structures that pass the CC filter (red) are those that are at the end of the structure bouquet and similar to the cryo-EM structure. The structures that are filtered out (blue) would eventually fail during docking. The CC filter removes such grossly inaccurate structure prior to the computationally intensive docking step.

The distribution of conformations is not around the cryo-EM structure but away from the cryo-EM structure in one direction. So, one end of the distribution contains structures that pass the NC and CC filters and pass docking while the other end has structures that fail even the CC filters since they stray away from the tunnel exit. This gives rise to lower number of structures that pass the filters but still have structures in them that can fit in the tunnel. In other words, the structures that pass the filters will be distributed around the tunnel exit and only those that fit in the tunnel are selected by docking. But, in this case, one end of the distribution of structures that would not fit in the tunnel is missing and this decrease is shown in the NC and CC filter results.

The NC and CC filters do not have any biological implications and are used only to describe the tunnel geometry. *In vivo*, the peptide structures that traverse the tunnel would never crash into the tunnel, but rather would squeeze into a conformation that can help them traverse that segment of the tunnel. The absence of atomic structural and dynamic data showing the movement of the polypeptide and the actual geometry

of the tunnel necessitate such different approaches to understand tunnel geometry. This is an alternative to building the peptide inside the tunnel with constrained freedom. The advantage of this method is the sampling coverage and time efficiency. In this whole study alone, more than two billion nascent polypeptide all-atom structures with hydrogens have been sampled. This extensive sampling would have covered a large portion of the twenty residue long nascent polypeptide space. As more and more experimental data become available, the sampling can be redefined according to the data and further narrow down the possibilities. *De novo* sampling beginning from alanine's complete dihedral space has provided the exact conformation of the different residues at different positions of the tunnel. The biggest advantage of this method is the availability of all-atom structural ensemble that corresponds to every result that has been discussed.

3.5 Conclusion

The study of nascent polypeptide and ribosome tunnel geometry is aimed at a deeper understanding of the dynamic interactions that happen inside the tunnel. The nascent polypeptide and the tunnel are difficult to experimentally study due to being buried in one of the largest biomolecular machinery inside the cell. Ensemble sampling and dock by superposition provide an excellent tool to probe the geometry of such systems. MD and Delaunay triangulation demonstrate tunnel convolution due to the presence of the nascent peptide, by an increase in surface area and decrease in tunnel volume. The structure of nascent polypeptide established by structural studies can only contribute a small set of conformations that the tunnel can accommodate. *De novo* sampling from a complete conformational set shows that the upper to central ribosome tunnel accommodates different conformations at different segments of the tunnel. As expected, conformational freedom increases as the peptides move towards the exit. The tunnel segment at residue fifteen of the peptide contains considerable freedom to allow extended (beta and PPII) and helical (right- and left-handed)

conformations for the peptides. This freedom proves to be a disadvantage since it gives rise to helical properties in peptides that cannot be maintained further along the tunnel.

In summary, the MD simulations and Delaunay triangulation provide a tool to study the dynamics of cavities inside proteins. Ensemble sampling and dock by superposition describe the complete conformational space of the tunnel and also contribute all-atom structural representations of different thresholds and segments in the tunnel.

4. Intrinsically Disordered Proteins

4.1 Introduction

The long-standing debate of the existence of a unique low energy structure for every protein sequences was put to rest in early 2000 by the initial reports of intrinsically disordered proteins [82,87,89]. Further, the significance of the finding was increased when it was found that a large subset of the proteome has disordered regions, even up to 51% in some eukaryotes [278]. IDPs perform various important functions in the cell such as modulating cell division and regulation of macromolecular assembly among others [279-281]. A considerable fraction of transcription factors [282] and 66% of signal transduction proteins [283] have been predicted to be intrinsically disordered or contain segments of disorder. Disorder seems to play a major part in cancer, as 79% of human cancer-associated proteins (HCAPs) have been predicted to contain at least 30 residue long disordered segments [283]. A third of the eukaryotic proteins are estimated to contain stretches of at least 30 continuous disordered residues [284]. It is clear that disordered proteins are a crucial part of the signal transduction process, and is explained by their capacity to bind multiple targets [285]. This property of having multiple binding partners makes them critical points in the transduction pathways and understandably implicated in various disorders. They are also reported to act as linkers between folded domains in proteins [286]. In addition, they function as repulsive spacers in neurofilaments (entropic bristles) [287]. Titin, an IDP, is known to function as a spring and is related to function of inducing passive tension in muscle filaments [288]. Intrinsically disordered F-G domains in nucleoporins are also known to act as semi-permeable barriers in the nuclear pore complex [289].

IDPs are implicated in various diseases and disorders [97] and make it essential to study their functional mechanisms. These proteins are implicated in diseases such as

in Parkinson's and Alzheimer's due to their tendency to aggregate [102,290-292]. These motivate studies to elucidate and understand unfolded structure in aggregation. The structure of IDPs in physiological conditions has non-random diverse set of structures. These structures have been known to be essential in modulating their functions [293-295]. Elucidating the structure of unfolded proteins in detail could be the way forward for development of new therapeutics by structure-based drug design [296] .

IDPs are distinct in their amino acid compositions and vary markedly from folded proteins in this aspect. IDPs are reported have low sequence complexity compared to natively-folded proteins and are deficient in hydrophobic amino acids and enriched in hydrophilic amino acids [297]. This unique sequence complexity of disordered regions is the main characteristic used to predict and identify disordered regions from protein sequences [297-299].

Binding multiple partners is an inherent feature of IDPs and is explained by the large diversity in their structures. Structures of IDPs are highly dynamic and heterogeneous and this complicates their structural characterisation. IDPs cannot be represented as single structure but rather as an ensemble of structures. As a result, most of the experimental techniques applied to the disordered proteins can describe only the global ensemble-averaged properties and not individual dynamics of sets of atoms. NMR and SAXS are the most popular experimental techniques used for understanding the secondary structures and general dimensions of the ensemble. Obtaining detailed structural and dynamic information is very challenging. A comprehensive summary of the techniques is provided in table below.

Technique	Type of Data Obtained	Suitable for	Not Suited for
SAXS	Dimensions of the protein state including Rgyr, shape and size of the ensemble	Very easy to obtain basic information on disordered state	No structural or atomistic data and sensitive to aggregation
NMR secondary chemical shifts	Residue specific secondary structural data	Understanding mechanism of IDPs based on helical or beta conformational data	Global parameters based on ensemble average
NMR NOEs	Hydrogen-hydrogen distance information for protons closer than ~5 Å in distance.	Distance data for pairs of residues in sequence	Highly sensitive to dynamics and populations of conformers within ensemble
NMR PREs	Distance distribution functions of unpaired electron and observed spin	Studying intermediate tertiary contacts	Non-specific binding of spin label and incomplete spin labelling of protein
NMR RDCs	Orientation information of dipole-dipole interaction vectors in a common reference frame.	Reveals native-like topologies in disordered proteins; restraints in MD and MC sampling	Alignment frames differ between proteins; predicted alignment frames is sensitive to small sets of conformers
Hydrogen/deuterium Exchange	Change in deuteration due to binding and folding of disordered domains.	Studying coupled folding and binding and disorder-order transition	Secondary structural information or disorder information cannot directly be inferred.
PFG NMR	Translational diffusion coefficient, ensemble-averaged hydrodynamic radius	Dimensions of disordered ensemble; amount of disorder	Solvation is not constant among different proteins

Table 4.1 Experimental tools used to study IDPs. The current methods applied for study of IDPs are considered along with the information they provide, their advantages and their disadvantages.

Experimental data are sparsely available for unfolded proteins and the challenge to solve unfolded protein state still remains. The unfolded proteins sample a large conformational space [300,301] and have less persistent structures than native proteins. NMR is the most popular experimental technique used to study the unfolded states. In NMR, the unfolded states display poorly dispersed resonances since the chemical environment of the spins are highly degenerated [302]. This makes it difficult to interpret the spectra due to overlapping peaks. Another disadvantage of information obtained from NMR studies is that it is averaged over the complete

unfolded state and poses various methodological challenges. The distance information from NMR are weightage averages and are biased towards close contacts due to their dependence on the r^{-6} variable. These unfolded states cannot therefore be interpreted as a single structure to apply these experimentally-determined restraints [49,302]. The unfolded states are hence represented as an ensemble of structures that can fulfil experimentally-averaged data but there is no systematic approach to determine such an ensemble. These average distances obtained from experiments cannot be directly applied to refine the ensemble but rather the biased averaging need to be explicitly taken into account.

Various computational approaches have been used to circumvent this problem to study the unfolded state. The disordered state shares characteristic challenges and properties with the denatured state due to its diverse structural conformations and therefore derives from methods used to study the denatured state [12,49,303]. All these methods revolve around fitting and matching average experimental data to average calculated ensemble data. The idea is to generate a set of structures that agree with experimental data. This can be applied using different approaches. A straightforward approach is to generate structures of multiple conformations and assign weights to individual conformations so that their average matches experimental constraints. This is comparatively simpler for native states since they require only fewer structures. Unfolded states require large ensembles and more data to obtain meaningful sets of structures. Ensemble modelling has been used successfully to characterise the disordered state [100,119,304,305]. Ensemble modelling is applied by using different search algorithms to generate representative structures that can fulfill sets of experimental constraints.

Filtering and selecting the right ensemble is done by selection tools which use experimental data and find agreeable ensembles. ENSEMBLE [306] is one of the most popular Monte Carlo-based selection processes which has been used to study

IDPs [52,302,307]. Evolutionary algorithms such as ASTEROIDS [308] search for conformations that match the given experimental constraints [116,118].

There are methods available that can generate ensembles for such disordered proteins. Restrained MD or Monte Carlo simulations have been used to restrict ensembles of conformers rather than working with a single structure. They can utilise distance information obtained from NOE and PRE to restrict and filter the ensemble structures.

Irrespective of the method and approach, the initial candidate pool of structure is paramount to obtaining an accurate ensemble. The approaches on sampling structures fall into two main categories. The first is to generate a large pool of candidate structures and to select a subset of structures that agree with the observed experimental data. The second approach is to use theoretical constraints to generate only a very specific ensemble of structures [309,310]. The second approach is more popular, since it can incorporate considerable amount of constraints and because the first approach is computationally complex. Generating the computational ensemble from experimental data is still complicated by the amount of experimental parameters available against the ensemble structures. In practice, the experimental data obtained is not enough to uniquely define the ensemble. For a given experimental data, there are possibly different ensembles that can match the parameters. This leads to degenerate ensembles, all of which match the parameters. It is clear that the number of experimental parameters and the type of information will dictate if it is possible to find a detailed ensemble. Replica exchange molecular dynamics and ensemble modelling are the most popular approaches for structural characterisation of IDPs.

Molecular dynamics simulations have been used in providing detailed structural insights on different unfolded states [311,312]. Restrained MD simulations modify the potential energy functions to direct the trajectory to sample conformations that agree with specific experimental constraints. This is different from steered MD,

where only the final structure is of consequence. In restrained MD, the restraints are applied to the whole trajectory and the ensemble rather than just the final structure [53]. Replica exchange MD is a popular tool for applying such restraints. Multiple replicas of the protein are simulated in parallel and restraints and potentials are calculated by averaging over multiple replicas [313]. The amount of restraints required to describe the ensemble are complicated. PRE restraints alone may not be sufficient in describing the ensemble and more than four PRE restraints are required per replica for good performance [117]. Interestingly, in addition to these PRE restraints, simple Rgyr information has been able to provide an independent parameter to model the ensemble [53]. These clearly show that, given a few experimental constraints, it is possible to generate a representative ensemble of the disordered state. Based on sparse NMR data, partially-unfolded states of a protein have been described [314].

Monte Carlo-based statistical sampling is an alternative for generating the initial pool of structures. Monte Carlo's extensive sampling proves highly advantageous to cover a large protein backbone space. Although MD and MC are known to be equally effective in determining native states, MD is more suited for trajectory-based studies where the simulation always heads to a final state. MC is more than twice as fast as MD in native state sampling of proteins [315]. The comparatively flat energy landscape of disordered proteins further favours such random MC sampling rather than MD. Although MD has been modified to suit IDPs to overcome some of these disadvantages, time efficiency and comprehensive sampling of MC can be expected to provide better results. TraDES has been very successful and popular in generating an initial pool of structures, which can be refined by adding experimental parameters. The previous version of TraDES was limited in its ability to generate protein structures with solvent layers. It also had outdated conformational and rotamer libraries and its dictionaries were not clearly defined for specific sampling. To this

effect, TraDES 2.0 has been released with considerable updates. These include updated libraries, support for newer NCBI Asn.1 file types, capability to add multiple solvent layers and different ratios of conformational sampling, among others. Although the simulation engine remains the same, it is imperative to run it through its paces. One of the new additions include a specific random-coil sampling ratio to model disordered proteins. Generating disordered ensembles is a challenge and very few approaches can boast of successfully modelling them. Dimensions of disordered proteins from SAXS and random-coil sampling from Flexible Meccano are used to demonstrate its ability to generate ensembles for disordered proteins.

4.2 Methods

The updated TraDES 2.0 package (http://trades.blueprint.org) was used to generate the ensembles for a set of disordered proteins. A set of thirty-nine disordered proteins are chosen to be described by TraDES. These proteins have been experimentally characterised by SAXS by determining their radius of gyration (Rgyr). These proteins have also been used to verify FM's ability to sample disordered proteins and hence, the Rgyr data from FM is also available for comparison.

4.2.1 Disordered protein set

Thirty-nine protein sequences were used to generate their ensembles. Tau proteins are natively unfolded microtubule-associated proteins which occur mainly in neurons [316,317]. A set of 16 variants of tau protein were included in the sample set. These proteins have been earlier studied using SAXS and their dimensions are available [86]. These have also been described by FM [305]. In addition, nine other disordered proteins which have been studied by SAXS and FM were considered. Another set of fifteen disordered proteins were also included from a study of characterising IDPs using SAXS. The complete list of proteins is provided in Table 4.2.

4.2.2 TraDES

The TraDES 2.0 package `TraDES-2-20120612-CentOS5_5_x86_64.tar` was used to generate the ensembles. TraDES has had several updates from its previously-released version in 2002 [154,155]. The conformational libraries which were earlier based on 834 structures have been modified to include 7030 high-resolution structures from the PDB. A new sampling ratio with full random-coil sampling has been introduced in the `seq2trj` and `str2trj` programs. This makes use of random-coil libraries that are known to model disordered protein [305,318]. The sequences of the proteins consistent with their SAXS reports were obtained from UniProt [319]. `seq2trj` program was used to generate a trajectory distribution from the sequence with the option `-c T` for simulating disordered sampling. The trajectory distributions are used as the input for the program `trades`, which performs the actual sampling and generates structures. No solvent layer was added to the structures since only dimensions are analysed and not the energies. 100,000 structures were generated for every protein. `trades` generates log files that contain Rgyr values in addition to other dimensions and energies. `TraDES_R_Analysis_Package.r` was used to analyse the log files and provide Rgyr mean values.

4.3 Results

4.3.1 TraDES vs SAXS

TraDES was used to generate ensembles of 39 proteins that were previously analysed by SAXS. The average Rgyr of the TraDES structures were calculated and compared to those generated by SAXS. The specific random-coil libraries were used to generate the ensembles. The comparison of Rgyr of the 39 proteins between TraDES and SAXS are shown in Figure 4.1 and Table 4.2. They have a positive correlation of

0.8189 and are almost equally distributed on either side of the best fit line. These sets of proteins have already been studied by FM and are known to similar distributions.

Protein	**N-Expt**	**Expt Rgyr (**Å)	**TraDES Rgyr (**Å)
MeCP2 [320]	486	62.5	62.721
Msh6 N-term [321]	304	56	50.16357
Ki-1/57 [322]	292	47.5	48.38408
MeCP2 (78-305) [320]	228	37	42.53614
Synthetic resilin [323]	185	50	38.43794
Hrpo [324]	147	35	34.35366
II-1 [325]	141	41	31.51284
α-synuclein [326]	140	30	32.28426
N-tail nucleoprotein MV [327]	139	27.2	32.93899
β-synulcein [328]	137	49	31.90512
NHE1 cdt [329]	131	37.1	32.2762
NHE1 cdt [330]	131	35.3	32.2762
ERM transactivation domain [331]	130	39.6	32.23916
Neuroligin 3 [332]	118	31.5	30.59764
Prothymosin [333]	109	27.6	27.83687
paNHE1 cdt [329]	107	32.8	28.58348
paNHE1 cdt [330]	107	32.9	28.58348
FEZ1 monomer [334]	103	36	29.05648
HIV-1 tat [335]	101	33	28.37137
p53 [336]	93	28.7	27.02935
IB5 [325]	73	27.9	22.1678
N-term VS [337]	68	26	22.52805
pir[338]	75	26.5	23.75349
tauk18p301l [86]	130	35	31.73816

tauk19 [86]	99	35	27.49615
tauk17 [86]	143	36	34.02773
tauk27 [86]	171	37	36.50113
tauk18 [86]	130	38	31.83154
tauk16 [86]	174	39	37.5748
tauk10 [86]	167	40	36.62879
tauk32at8at100 [86]	202	41	40.01009
tauk25 [86]	185	41	37.52426
tauk32 [86]	202	42	39.88872
tauK23 [86]	254	49	45.1878
tauk44 [86]	283	52	47.40171
tauht23at8at100 [86]	352	52	53.20303
tauht23 [86]	352	53	53.25188
tauht23s214e [86]	352	54	53.12282
tauht40 [86]	441	65	59.6971

Table 4.2 Disordered protein set simulated by TraDES. Thirty-nine proteins which have experimental SAXS Rgyr are studied using ensemble modelling. The TraDES Rgyrs are the mean values calculated from ensembles of 5000 structures, generated by coil sampling.

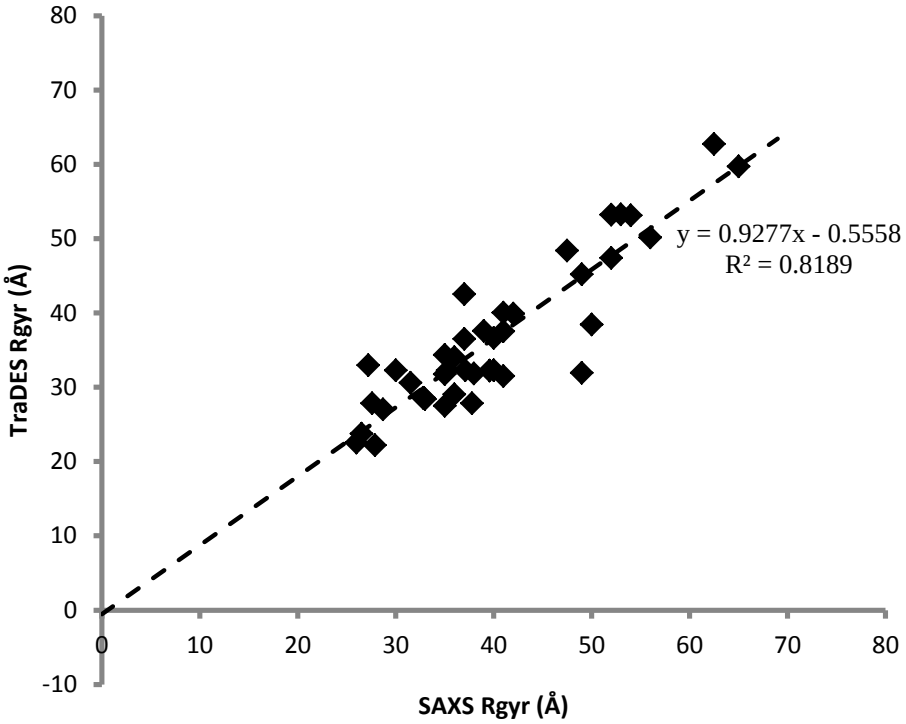

Figure 4.1 Comparing the Rgyr of IDPs calculated by TraDES and SAXS. The slope is 0.9266 and R^2 value is 0.8189. The correlation coefficient between TraDES and SAXS is 0.911732. The Rgyr values of TraDES are mean values calculated from an ensemble of 5000 structures.

4.3.2 TraDES vs FM

TraDES has been updated to include the latest set of sampling libraries. FM has been shown to provide initial ensembles that can be filtered to represent disordered protein ensembles. TraDES has shown similar distributions of Rgyr compared to FM and they are compared in Figure 4.2. The comparison shows remarkable similarities between TraDES and FM. The dimensions of disordered proteins generated by TraDES and FM have a correlation of 0.9968.

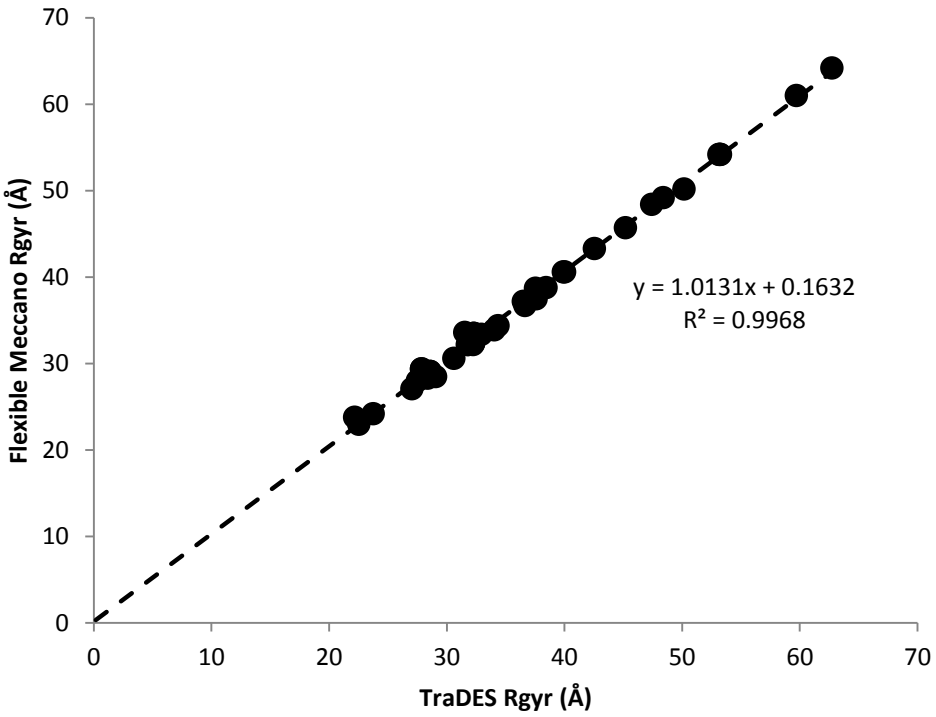

Figure 4.2 Comparing the Rgyr values from random-coil sampling by TraDES and FM. This shows excellent correlation with slope of 1.0131 and a R2 coefficient of 0.9968. The TraDES Rgyr values are mean values calculated from an ensemble of 5000 structures and FM Rgyr were obtained from [177].

4.4 Discussion

From the simulation of a set of thirty-nine proteins, it is shown that the updated TraDES package is capable of generating disordered ensembles that have good correlation with their SAXS dimensions. It is evident that TraDES is not able to completely capture the Rgyr of the proteins. Different proteins have different amounts of disorder in them. Some have long stretches of disorder between ordered regions, while others are completely disordered. Disordered proteins have been classified into four groups based on their compactness [40]. The molten globule-type proteins are expected to have smaller hydrodynamic radii. Deviations from the ensemble Rgyr have been observed in earlier studies due to different structural phenomena. Even in the case of tau proteins, it has been noted that smaller proteins have larger Rgyr values while larger proteins have smaller Rgyr values than the predicted random-coil values [86].

Another reason is that the Rgyr values calculated from ensembles are averaged over the whole ensemble. All the ensemble conformations may not always prevail in solution and the small subset of sampled conformations from the ensemble results in the discrepancy. This is common in disordered proteins due to their flexibility and is one of the reasons for their roles as interacting hubs that bind multiple partners. Their conformations change based on the environment and hence experimental constraints are required to filter the ensemble to get an accurate representation.

This diversity in disordered proteins poses a significant challenge in studying them. Most of classical single structure studies are not applicable to these proteins. However, simulations alone also do not accurately represent these proteins and cannot be used to describe them. A mixture of simulation and experimental data has been the most successful approach to tackle the challenges that exist. Experimental data provide distance and secondary structural data that can be used to improve the model. These experimental data require an excellent representation of the diverse conformations in the ensemble. A poorly represented ensemble can lead to misleading conclusions even if experimental data is available. A thorough sampling of the disordered state is a necessity if meaningful analyses have to be carried out. TraDES produces all-atom structural ensembles that completely sample the conformational space. It is time efficient and can generate a large number of structures which increases sampling coverage.

Comparison between TraDES and FM show excellent correlation in predicting the dimensions of proteins. Since both uses the random-coil libraries to sample the disordered space, their similarities are not surprising. It is expected that both these tools sample the disordered space thoroughly and provide excellent initial ensemble pool of structures that can be filtered using experimental constraints. TraDES has the added advantage of generating structures with complete side-chains while FM has spherical model which is weighed based on the amino acid type. FM, on the other

hand, can predict some of the structural parameters from the ensemble which can be compared to experimental data. However, the prediction of structural parameters in not built-in; it is possible to add such packages independently [339]. Although studying IDPs poses theoretical challenges, use of experimental data with updated sampling tools such as TraDES can be expected provide significant insights on the disordered state.

4.5 Conclusion

TraDES has been updated with latest conformation and rotamer libraries for sampling different protein states. Disordered protein space is modelled by sampling specific random-coil libraries using sequences of disordered proteins. Comparison of radius of gyration values with SAXS shows good correlation of Rgyr values, but deviations are observed in many proteins. The deviations are due to structures present in different amounts in the various proteins. Comparisons with Flexible Meccano show near-perfect agreement over the set of thirty-nine disordered proteins. This substantiates TraDES's ability to predict random-coil dimensions of disordered proteins and the accuracy of newly-added conformational libraries. In addition to Rgyr, TraDES also calculates ensemble properties such as end-to-end distance and provides all-atom structural representations of the disordered state for the application of experimental constraints.

In summary, this chapter demonstrates that TraDES can provide structural representations of disordered proteins in relation to their random-coil structure. These ensembles can be used to describe the disordered protein functions and mechanism by filtering them with experimental constraints.

5. Energetic Interactions in Folded Proteins

"The elucidation of the structure of haemoglobin has, if anything, made this problem more challenging…"

-Monod et al on Allostery, 1965

5.1 Introduction

Communication between residues in protein structure is crucial to its function and is responsible for determining its role in cellular processes. These residue-residue interactions can occur between distant sites in the protein and a change in one site can modulate the other. This control is one of the most fundamental elements of protein function and exists in inter-protein and intra-protein interactions. Such protein interactions and modulation involves the binding of an external protein or a ligand which can activate another distal site on the protein, without a direct interaction between the sites. This is termed allostery, where two distant sites can modulate their functions by conformational change, without directly interacting with each other [340].

Residue-wise interactions in a given protein structure, as we know it, are completely dependent on the sequence and the structure it folds into. Although solving the structure helps in understanding of the functionality of the protein, it does not provide any information on its inherent energetic complexity. In barnase, it was noted that three of the six major energetic interactions were not identified by the examination of the structure [341]. The protein's main evolutionary selection criterion is to carry out its function and this is directly dependent on its structure [342]. So, the protein can be expected to evolve its sequence in such a way that its structure is still functional and conserved. Thus, it becomes imperative to study the structure of the protein, independent of its current sequence to understand its complete structural make-up and

its properties. This can provide important details that have not been characterized yet, since protein structure has always been studied only as a function of its sequence. It is observed that various sequences can fold into similar structures [343,344], making it clear that a sequence is just a variable that can be modified as long as it fits the structural.

It has been clear that different positions in the protein act cooperatively to influence the function of distant sites on the protein surface [345]. The propagation of information through the protein structure is believed to be through complicated energetics [346,347]. The protein communicates within itself by energetic perturbation across its structure and is responsible for the specific changes in protein function [348,349]. But, it is not yet clear how the energy propagation happens within the protein. Studies on the binding energy of growth hormone and potassium channel pores show that only a few residues on the surface account for most of the free energy change during binding [350-352]. Protein sectors have been mapped for the PDZ domains and residues that contribute to energy transfer have been grouped using amino-acid coevolution [136]. These studies show that energy propagates through the structure in a seemingly arbitrary manner with no common mechanism and can be expected to vary depending on the tertiary structure of the protein. The PDZ has been shown to contain a network of co-evolved residues that are energetically coupled. This “sparse network” of clusters of residues has been confirmed by experiments to affect ligand binding and stability of the PDZ domain. Studies on the S1A serine proteases have defined different energetic clusters of amino acids responsible for different properties of the protein function, such as catalytic power and thermal stability [137]. It is clear that proteins have a network of interacting residues, which communicate with each other and this communication is paramount to its function. Characterising and predicting such interactions can help in engineering proteins for a wider range of applications.

The most comprehensive method to determine and understand this protein energy distribution is to completely map the energetic coupling between clusters of residues. Thermodynamic mutant cycle analysis or double mutant cycle analysis is a systematic way to probe and quantify the energy changes that occur due to binding and conformational change. If two residue positions are energetically or allosterically dependent, then energetic difference observed by mutating one of these positions will depend on the amino-acid present in the other position [353]. The change in energetic parameters such as binding energy, etc. is analysed by mutating the two residues separately and together.

$$\Delta\Delta G_{interaction} = \Delta\Delta G_{X\rightarrow A,Y\rightarrow B} - \Delta\Delta G_{X\rightarrow A} - \Delta\Delta G_{Y\rightarrow B} \quad \text{Equation 5}$$

If the residue positions are independent, mutating them separately will have the same change in energy as mutating them together. On the other hand, if they are dependent, then mutating them together will show a difference in comparison to the individual mutations (Equation 5).

In terms of binding energy, the change in free energy in a double mutant is the sum of those energies for the two individual mutations. If the residues are coupled, then the free energy change for the double mutant is different from the sum of the individual mutants [341]. It is also possible to deduce the interaction energy between the two residues by calculating the free energy change [354,355]. This can be analysed by calculating a coupling coefficient between the two positions.

This requires systematic mutation of the amino acids at different positions and to calculate the energy change. Mapping a whole protein is practically impossible due to the large set of mutants that need to be developed and the complexity of the possible amino acid permutations.

For example, only a few positive mutants for 10 positions have been established in the past 40 years for the aspartokinase from *Corynebacterium glutamicum* [356]. The biggest challenge is the practical impossibility of obtaining a significant subset of the mutants. Mutations can lead to non-functional proteins or non-stable proteins while lethal mutations can kill whole cells. A complete characterisation is hence impossible due to these challenges.

A number of methods, mostly using multiple sequence alignment and perturbation models, have been shown to provide an alternative approach to map the energetics of the protein [357-359]. However, the opinions on the information provided by MSA are not all positive [140,360] since the methods are very sensitive to the sequences and they do not consider the inherent evolutionary noise. All these methods are based on the fact that structurally-important residues will evolve together.

There have been similar methods that utilise mutual information and claim to eliminate such phylogenetic noise [139,361]. But, any method that is based on MSA would suffer the same disadvantages, albeit less or more. Another issue is every sequence in the MSA is at a different state of evolution and thereby under differing evolutionary pressures. For example, a single domain protein is under far less pressure than the same domain present in a multimeric protein. This difference would be reflected in their sequence but does not correlate with the fact that the domains still share the same energetics. Most of the co-variation tools are based on an initial multiple sequence alignment (MSA) of the sequence set and, as expected, their results are influenced by the quality of the MSA chosen. Uneven sequence representations, insufficient evolutionary divergence and presence of gaps cannot always be avoided and can lead to a non-optimal MSA and can adversely affect the results [362].

Protein structure determines its function and hence their structural integrity has been preserved throughout their sequence evolution. An initial step towards understanding the structure is to find out its interactions, independent of its current sequence. This is

possible by mutating different sequences and mapping the changes in interaction and energetics of the resultant structures. This would provide us the interaction patterns and amino acid preferences that are inherent in the structure. Experimentally, the method is very challenging given that solving a single structure takes a couple of years, on average, and to generate enough mutant structures to be statistically significant would be next to impossible. Other computational tools such as molecular dynamics (MD) simulations also cannot handle such large mutant sets of protein structures since, they are highly demanding in terms of computational complexity. Statistical tools are a very good option, but statistical tools are, in general, very simple and are easily pushed to their limits as they are based on the present data and can rarely be extrapolated. On the other hand, physical methods are complex as mentioned above. Tools that are a mixture of statistical and physical models, which can reasonably account for all the properties, have been few and far between.

Statistical mechanics provides us with one such tool, cluster expansion (CE), which has been popular in alloy theory for predicting properties of alloys. CE has been successfully extended to protein structures and in calculating protein energies [363]. Cluster expansion is a statistical mechanics tool that is able to capture a system's property, which can be broken down to discrete variables [364]. In the case of modelling energy of a system, CE uses discrete variables that give the amino acid occupancy of each point in the structure and is shown to calculate energies of a few protein domains [365]. This method is extensively used in searching for low-energy crystal structures in alloy theory and in predicting alloy phase diagrams [366,367]. We use CE to model the energy of the WW domain as a function of its sequence. CE is defined in terms of states, functions and variables. For the protein sequence data, every function is defined as the occupancy of a single position or sets of positions and the states are defined as the set of amino acids. This provides us with a

straightforward method to calculate any property of a system by linear summation of all functions in all possible states.

This approach is first tested in a small beta hairpin to check if the interactions are properly represented. Then, CE is used to represent the WW domain as a set of point and pair functions representing various positions. This model is analysed to calculate the energetic contribution of individual and pairs of positions and thereby their energetic coupling.

The first step is to describe the application of this tool to find out the sequence-independent interactions in a simple model of an anti-parallel β-strand. The model was based on the structural template of the β-strand from the ribosomal large subunit protein L22. β-strands have interactions that are not close in sequence but in structure. The interactions in such simple structures are straightforward and are verified easily. This also provides a small simple model to showcase the different capabilities of the CE model.

The L22 β-strand can provide a model to verify the approach of CE but there are no specific biological inferences that can be obtained. The ultimate goal is to calculate detailed residue interaction patterns which can provide important clues on the energetic network inside the protein structure. It is clear that such energetically interacting residues are often found in clusters, referred to as hot-spots [352], which play important functional roles such as stability, binding, catalysis, etc. Identifying these interaction maps can facilitate drug design. These interacting clusters are sites which mediate allostery and could be direct targets for inhibition or modulation. Even otherwise, changes in these residues could adversely affect protein stability and reduce its half-life. This would also help understand the mechanism and change binding preferences for inter-protein interactions [368].

To deduce energetic interactions in the WW domain structure, CE is used to map the positional energetic interactions between residues in the WW domain. The WW

domain is a compact folded structure 34-35 residues in length and contains two conserved Trp residues and a conserved Proline residue [369,370]. Due to its small size and fast folding capability, the WW domain is well studied. A significant amount of mutants are also available to cross-check and verify any interaction data obtained from CE. CE characterises the β-strand well and shows diagonal interactions due to side-chains, and identifies a network of residues that interact energetically in the WW-domain that contribute to its stability and binding.

5.2 Methods

To establish the sequence independent interaction pattern in the structure, a diverse set of sequences are threaded onto the backbone of the structure and their energy is modelled and calculated using CE. The sequences are generated at random using a given set of amino acids. The number of sequences generated is proportional to the number of positions that are modelled and the amino acid set being used. Once the sequence set is generated, it is threaded onto the structure and their individual energies are calculated using physical potentials. Cluster expansion is used to define functions that represent the presence of amino acids in various positions. The model is tested by calculating the energy of the structure and is evaluated using a training and a test set. The RMSE (root-mean-squared error) of the predicted and the calculated energies is used to measure the accuracy of the model. If the model is able to predict the energy accurately, then the variables are used to provide interaction data. The various steps are detailed below.

5.2.1 Cluster expansion

CE is a statistical mechanics tool to discretise any property of system into orthogonal functions Minimally, CE is an expansion of the energy using a set of orthogonal functions that encompass all the possibilities of the sequence space i.e. all amino acid sequences on the protein backbone. The CE model was built using alanine as the reference amino-acid, as described in [365]. They have used CE to model the energy

of a given structure and, using a training set of sequences, calculate the energy of any sequence in that structure. We apply a similar model to represent structure and have extended the method to find the interaction patterns within the structure.

The focus is to first to apply the cluster expansion methodology to protein sequence data and create a model that will completely represent the data. Creating a model that cannot be biologically or physically interpretable, is not useful. So, it is very important to define the variables, or rather the coefficients, in such a way that they could be related to some biological or physically-defined property, that can be measured. For a protein ten residues long, the number of variables and coefficients could exponentially increase as higher order terms are included. Thus, it is essential to come up with a scoring system which would be able to detect and differentiate which variables are important and which are not. This is not an easy task as some of the variables may be important but the information they could provide may already be in the model, i.e. variables need to be chosen not just based on their information but based on the information they provide versus the information already in the model.

5.2.2 Theory

CE essentially expresses the energy of a protein which has been folded to a particular conformation, using functions taken from its sequence. CE takes a system of interacting variables and breaks it down to linearly independent basis functions that could be summed up to calculate any property of the system. CE divides the energy in terms of the defined variables that provide the occupancy of each position in the protein structure. CE is defined in terms of states, functions and variables. For the protein sequence data, every function is defined as the occupancy of a single position or sets of positions and the states are defined as the set of amino acids. So, this provides us with a straightforward method to calculate any property of a system by linear summation of all functions in all the possible states.

Here, f(s) is the property of the system and φs are the linearly independent basis functions. So, if energy is considered as the property and the functions are its residue-occupancy, then CE translates to the energy of the system being equal to the sum of all the amino acids in all the positions. The functions here are defined using binary variables, 1 if the amino acids are present and 0 if not. This, then, gives an equation to the energy of the system in terms of the occupancy of amino acids in each of those positions.

$$f(x) = \sum_{i=1}^{n-1} (J_i \varphi_i(x)) \qquad \text{Equation 6}$$

For a set of protein sequences, Equation 6 gives the energy, function and coefficients. From Equation 6, we know that energy for every sequence is proportional to the functions. So, when these are written in matrix form, we can see that the energy, E, is proportional to the occupancy function (φs), and the coefficients (or weighted sums) in the equations are the Js. So, from the sequence data we can determine the φ functions. Using training set of sequences and energies, we can fit the model and find the coefficients that minimise the error from the Equation 7.

$$\begin{matrix} E(x_1) \\ \dots \\ E(x_k) \end{matrix} = \begin{matrix} 1 & \dots & \Phi_i(x_1) \\ \dots & \dots & \dots \\ 1 & \dots & \Phi_i(x_k) \end{matrix} \cdot \begin{matrix} J_0 \\ \dots \\ J_i \end{matrix} \qquad \text{Equation 7}$$

If there are more sequences than the functions, then Equation 7 becomes over-determined, and it is possible to use least-squares fitting to find the optimal values of J's.

5.2.3 Scoring

For a system of a few thousand different variables and a much larger sample size, the fitting needs to be robust and stable with minimum error. Since the model used is a linear and is over-determined (number of samples is greater than the number of variables), least-squares fitting was used to estimate the coefficients [371].

Variable selection is a key issue since there are a large number of variables and not all variables may be important. The scoring system used is therefore a cross-validation (CV) score with respect to the variable set. This system has been used in [365] to determine if a function needs to be included in the model. But, in this case, it has been extended to find the actual contribution of the variables or sets of variables. In this system, a variable is left out of the fitting and the model is trained, tested and the rootmeansquared error is calculated. If taking the variable out of the fitting increases the error, we know that the variable is important to the fitting. But as the error is calculated on the test set, this is the true error rather than over-fitting. Hence a CV-score for every variable is calculated, that directly relates to the contribution of the variable to the fit. When more functions are included in the fit, the RMSE score decreases whereas the CV score might increase, if the functions are not relevant due to over-fitting. The CV score of the variables is therefore also of biological significance as it tells us which of the functions are important.

5.2.4 Finding interaction patterns

Each value is the RMSE of the model when a pair of positions is included. As interaction positions provide useful information to the model, the RMSE decreases when they are included, whereas the non-important positions lead to over-fitting and higher error. Two types of interactions and their preferences are analysed here. Positional preferences are based on interactions of pairs of positions. To compute this, we take the set of function variables that denote the various pairs of positions encompassing all the amino acid possibilities, and calculate their CV score by leaving them out of fitting. For example, to calculate the CV score of a pair at positions 1 and 2, we take the entire set of pair functions (over all combinations of amino-acids) that denote positions 1 and 2. I.e. function (Ala at 1 and Ala at 2), function (Ala at 1 and Cys at 2) etc, function (Tyr at 1 and Tyr at 2). If an ‘n’ amino acid alphabet set used to generate sequences, then there are a total of `'n*(n-1)/2'` functions that

represent the pair of positions 1 and 2. This entire set is added and its CV score is calculated. This shows the sequence independent importance of the pair of positions 1 and 2 in the protein structure. This can be used not only to calculate positions preferences but residue preferences as well. By including all functions that represent the pair of amino acids Ala-Ala (in various pairs of positions), we can show which amino acids interactions are important to the structure and is the compositional preference of the structure.

5.2.5 Modelling the L22-beta strand

The model was applied to residues 76 to 101 in the protein L22, which form the anti-parallel β-strand. The model is built using a subset of 9 amino acids, which are preferred in beta structures [372-374]. In the β-strand, 10 positions that hydrogen bond with each other were chosen as functions. Those 10 positions are allowed take up any of the 9 beta-specific amino acids as their states. The 9 amino acids used are Ile, Leu, Met, Phe, Thr, Trp, Val, Cys and Tyr. Only single point functions, which capture the contribution of an amino acid at a position, and pair functions, which capture the contribution of pairs of amino acids at two positions, are considered. So the entire expansion is done using a constant function, 90 point functions (10 functions that can be in 9 states i.e. 10 positions can have any of the 9 amino acids) and 3645 pair functions (10*9/2 functions can be in any of the 9*9 states i.e. any of the 45 different pairs of positions can have any of the 81 pairs of amino acids).

50,000 sequences and their energies are used to fit and validate the system containing 3736 variables. In this model, triplet functions do not contribute well to the fitting due to the earlier addition of pair-wise functions and hence have not been included in the model. In simple models, triplet functions mostly contain the same information as individual pair functions combined and thus lead to over-fitting, if added. Pair functions can be expressed in fewer basis functions than triplet functions and also have a better CV score and so, only the pair functions were included in the fit.

5.2.6 WW domain energetic linked clusters

The WW domain structure (PDB ID: 1PIN) [375] was used as the structural scaffold upon which the different sequences were threaded onto. All possible combinations of sequence space give 1.7 X 10^{44} possibilities. To reduce the sequence space, a MSA was constructed from the SMART [376] database and the sequence space were reduced to 9 X 10^{9} possibilities. 200,000 random sequences were generated based on the MSA and were threaded onto the WW domain structure using Rosetta fixed backbone design [377]. The energies and the positional preferences were similarly analysed using RMSE changes. Positive changes in pair-wise interaction changes were noted and an interaction map was manually inferred from the pairs of interacting residues.

The coefficients in the fit are either point functions, that describe the contribution of the specific amino-acid at that position, or pair functions, which describe the contribution of a pair of given amino-acids at a pair of positions. It has been shown that the coefficients for pair functions indeed correspond to the double mutant coupling energy [354,365]. This is similar to the way double mutant coupling energy is calculated [341]. A point mutation energy is not the energy of the side-chain of given X residue at a position but is rather the relative energy of one amino-acid with respect to another in that position. A more accurate way to quantify the energetic contribution is to mutate single and pair of positions [378]. In double mutant cycle experiments, the pairs of positions are energetically independent if the sum of the point mutation change in energy is the double mutation change in energy [379]. In case of CE, if two positions are independent, then their pair positional functions would add no extra information compared to the individual point functions, and would lead to over-fitting and reduce the CV score. Inversely, the reduction of the CV score by a pair function corresponds to them being energetically interacting, with the CV score providing a quantitative measure of the interaction energy. This is

exactly how the CE is able to correlate the double mutation energy since it represents the energy of a structure into discrete point and pair functions. In extension, the energetic contribution of a given position is the complete contribution of all the available individual amino acids at the position. Similarly, the energetics of a pair of positions is the overall contribution of all possible pairs of amino acids at those pairs of positions. This contribution is calculated by using the decrease in error of the model when these specific positions are added to the fit. In simple words, there are a group of functions (pair and point) that additively can provide the energy of any sequence in a proteins structure. The basis for the method is that if two positions are energetically connected, then the group of functions that represent these positions will have greater contribution to the overall energy of the structure. If, by adding such clusters of functions, the overall model is able to predict the energy better, the positions those clusters correspond to are energetically important to the protein. This contribution is quantified by using the CV-RMSE score, which gives the decrease in error of the model when the functions are added. Thus, more important positions have higher single and double mutant energies as their coefficients and as a result there would be a larger decrease in the error of the model when their functions are added.

5.2.7 Protein energy calculation

For the β-strand model, random protein sequences are generated using the nine amino acids for 10 positions. 50,000 such sequences are generated and are threaded onto the backbone of the crystal structure. Among the 50,000 sequences, 35,000 were used as training sequences and the coefficients were estimated using the method of least squares. These coefficients were then used to predict the energies of the remaining 15,000 sequences in the test set. These sequences are used as inputs to CHARMM [380] which generates the side-chain rotamers by avoiding any bad contacts. This is used as an initial structure to search for the optimal rotamers with lowest energy. The above were done using scripts written in PERL and UNIX Shell. The algorithm to

search through the rotamer space is based on dead-end elimination [381] and A-star search algorithms implemented for protein design [382-384]. The protein design code calculates E_{fold}-E_{repack} of the sequence in that structure. This is essentially the energy of the folded conformation disregarding the energy required to repack it. In essence, this is the energy of the folded protein after removing the individual contributions of the amino acids and is the folding energy of the sequence in that structure. This can include interactions within the backbone, side-chains and between them. The energy calculated is a sum of different energies such as van der Waals, electrostatics, surface accessible solvent area, etc.

The WW domain was modelled in Rosetta using the fixed backbone modelling algorithm. The energy obtained was the $E_{fold} - E_{repack}$. 200,000 sequences were threaded onto the WW domain structure using the `fixbb` module in Rosetta and their energies were calculated [146,377]. These were used to define and calculate the coefficients of the point vectors in the CE. 50,000 sequences were similarly used for calculation of coefficients for the pair and point functions and their RMSE and correlations were noted.

5.3 Results

5.3.1 The L22-beta Strand

The sequences are threaded onto the backbone of the structure and their energies were calculated as described in methods. The energies were normally distributed with a mean of -49.4 kcal/mol and range of -70 to -29 kcal/mol.

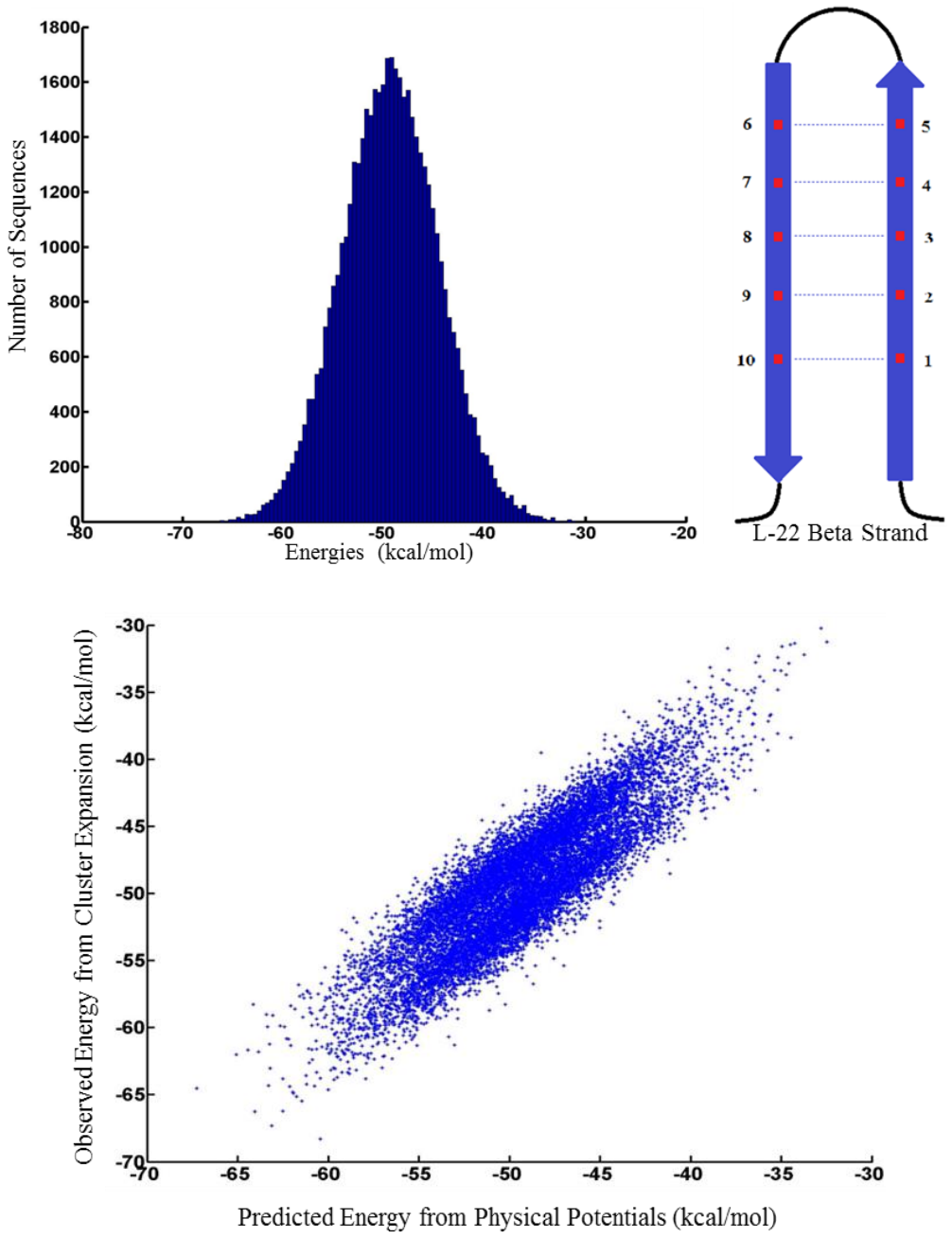

Figure 5.1 Distribution of energies of the sequences, model of the L22-β-strand and scatter plot of observed and predicted energies. This shows excellent correlation between the energy calculated from physical potentials and CE. CE accurately calculates the energy of any sequence in the L22 beta strand structure.

This provides an estimate of the randomness of the sequence and energy space (Figure 5.1). The energies are normally distributed without any skew-ness, which resembles a random model.

Coefficients calculated from the 35,000 training sequences were used to predict the energies of the remaining 15,000 sequences in the test set. This gave a good positive

correlation of +0.8614 and rootmeansquared error of 2.455 kcal/mol for the test sequences, which is very good considering the large sequence space. The energy was predicted with a mean error of 2.1137 kcal/mol, which is 5% of the given range of the energies. Hence, the model is able to predict the energy of the structure with good accuracy.

5.3.1.1 Interaction patterns for positional preferences

The CV score for all the variables were calculated iteratively as mentioned earlier. The CV score was used to determine how important the variables were. Figure 5.2 is a plot of the change in RMSE versus the pairs of positions. Each value is the RMSE of the model when a pair of positions is included. As interaction positions provide useful information to the model, the RMSE decreases when they are included, whereas the non-important positions lead to over-fitting and higher error. This CV score plot shows that some pairs of positions have reduced the CV score while some do not have a significant effect on the error as discussed earlier. So, the variables that have no effect are not important to the fitting and the variables that decrease are important. Here, there are eight pairs of variables that decrease the CV score while the rest do not contribute significantly. When these are compared with respect to the structure, it shows that these in fact represent the interactions that have been experimentally verified and predicted. This is represented using an impression of the β-strand (Figure 5.2 inset) with respective interactions in the same colour as the bars in the graph. Interestingly, this shows very negligible interactions between residues 2 and 9 (dotted line in Figure 5.2), which form a hydrogen bond.

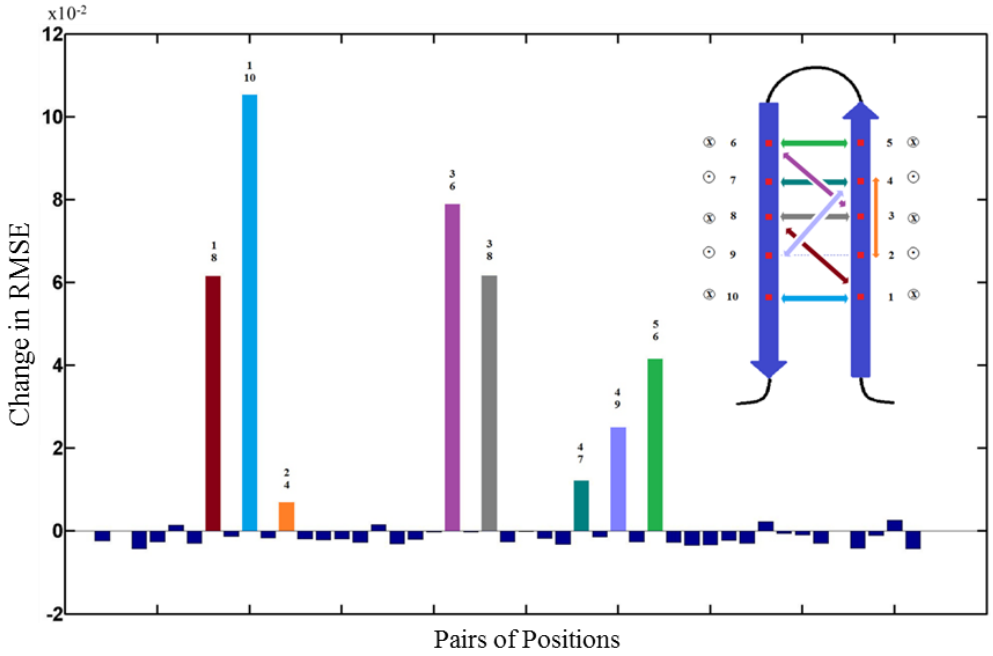

Figure 5.2 Positional preferences of residue positions in the L22 β-strand. The main pairs that show large reduction in RMSE are the hydrogen bonding pairs of amino-acids. Positions that are diagonally opposite to each other also show energetic coupling since their side-chains are pointed in the same side with respect to the plane of the beta-strand

5.3.1.2 Interaction patterns for compositional preferences

Similar to the positional preferences method above, compositional preferences were also analysed for which pairs of amino acids provide more information to the model. Figure 5.3 shows the change in RMSE versus the pairs of amino acids. Every bar in the graph corresponds to the change in RMSE when a pair of amino acids is included in the model. If the interaction between the included residue pair is important, the RMSE change is significant (red bars). The pairs of positions are averaged over all pairs of positions.

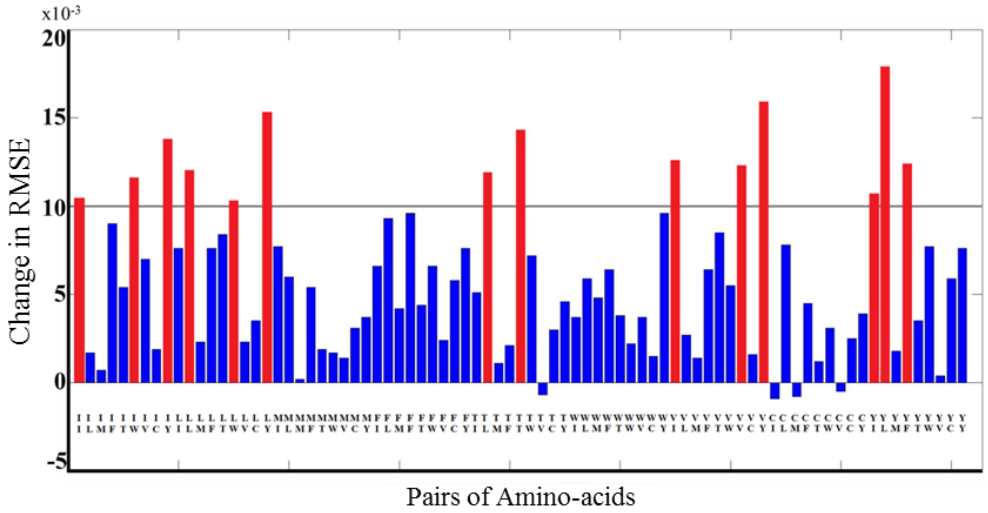

Figure 5.3 Compositional preferences of amino acids. Comparing different pairs of amino-acids averaged over all sets of positions show that tyrosine, isoleucine and leucine show reductions in RMSE when paired with each other. In addition, tyrosine, and threonine also are shown to be important.

For the threshold, any value above half of the total range was considered important. The amino acid pairs that are important, in no particular order, are Tyr-Leu, Val-Val, Val-Tyr, Leu-Tyr, Ile-Tyr, Tyr-Phe, Thr-Thr, Val-Ile, Leu-Leu, Ile-Trp, Thr-Leu, Tyr-Ile, Ile-Ile, Leu-Trp. These are acceptable preferences which re-emphasise that van der Waals and side-chain packing play an important role in determining the energy of this structure.

5.3.1.3 Searching for low energy sequences

The main advantage of this model is the fact that it is faster and considerably more accurate in calculating the energy of a sequence in the structure. Apart from the 50,000 sequences, which were used as training and test set, an additional 100,000 sequences were randomly generated and their energies were calculated using the model. The energies for those lowest energy sequences were again calculated using physical potentials to provide a comparison of the two methods. Out of all the sequences, the lowest 50 from both the methods are plotted in the graph below (Figure 5.4).

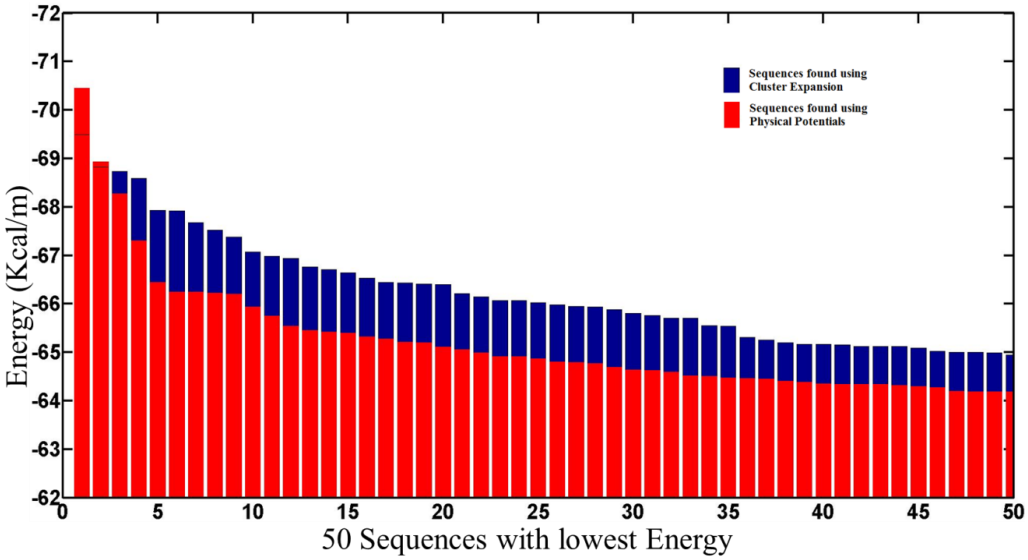

Figure 5.4 CE vs. physical potentials, in searching low energy sequences. The comparison of top 50 sequences with lowest energy between CE and physical potentials shows that CE is able to effectively search through the sequence space and find low energy sequences in the L22-beta strand structure.

From the above graph it becomes clear that the model is able to provide sequences with lower energies compared to physical potentials due to its time efficiency and its ability to search through large sequence spaces. To search through 50,000 sequences for a given structure using physical potentials, dead-end elimination and A-star search algorithms take a little over 500 days in a single processor. While in the above case, double the sequence space was searched in less than 2 minutes, which is around a million times faster than physical potentials.

5.3.1.4 Energy specific interactions

All the above results are derived based on the total energy computed from various types of energies such as van der Waals, electrostatics, etc. But the functions and the results inferred are based on the energy model that is being used. Extending the above analysis, this model has been applied to the structure using different energy inputs such as only electrostatic energy, only van der Waals energy and total energy without electrostatics, to understand the various amino acid preferences with respect to various energies. These results are represented as heat maps in Figure 5.5.

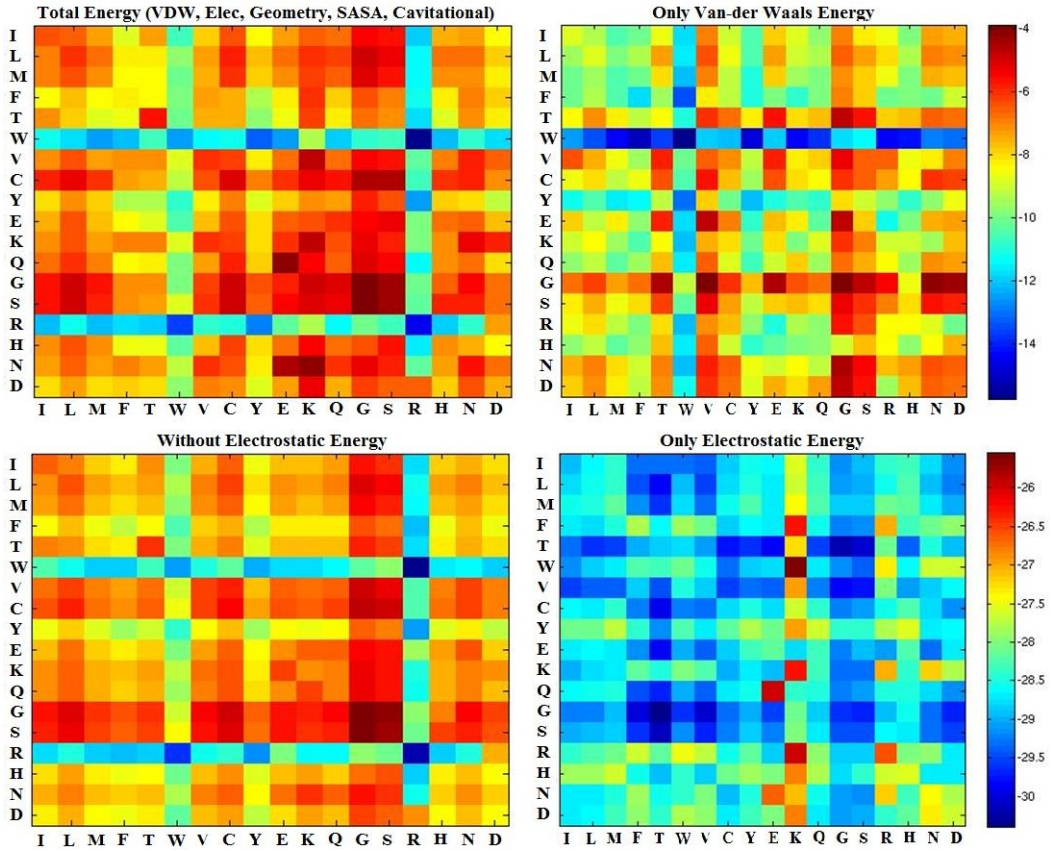

Figure 5.5 Heat-map for amino acid preferences using different energy terms. Individual values are the RMSE changes when the residue pair is added to the model. Red to Blue shows increasing change in RMSE and corresponds to greater contribution.

The individual energy-based models have been used to study the compositional preference of the amino acid pairs across all positions (Figure 5.5). The tryptophan interactions stand out in the van der Waals energy while smaller amino acids such as glycine and valine do not cause much of a change. Similarly, arginine and tryptophan are important when the electrostatic energy is not considered. When only electrostatic energy is used, lysine and arginine are different from the other amino acids. But, both lysine and arginine cause a negative effect compared to other residues while tryptophan is not important.

5.3.2 WW domain energetic mapping

The WW domain was modelled with 95 point and pair functions. All the point functions were included and the pair functions that reduced the error were added to

the model. The pair wise functions which reduced the error in the overall model were noted down. Figure 5.6 shows the progressive reduction of RMSE as the point functions are added to the model. As more information is included in the model, the model is able to better predict the energy of the sequences.

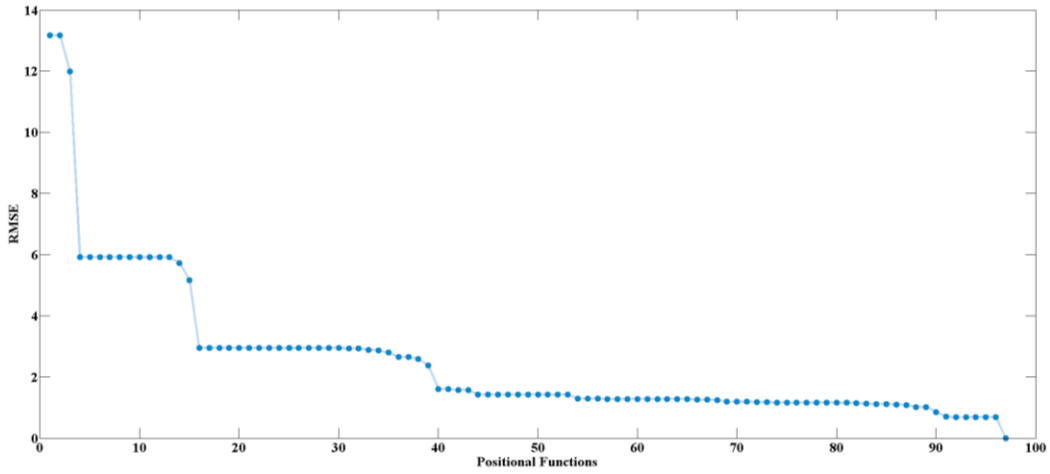

Figure 5.6 Change in RMSE by addition of point functions. Addition of point functions adds information to the model and decreases the RMSE. The initial point functions show large decreases in RMSE since there is little information in the model to begin with. As more functions are added, the change is more subtle since earlier functions have contributed.

The total 200,000 sequences were split as 140,000 training and 60,000 test sequences. The RMSE was calculated from the test sequences to overcome over-fitting issues. The 60,000 test sequences were used to check the validity of the model. The CE model, trained on the 140,000 sequences, was used to predict the energies of the 60,000 test sequences. The point vectors alone provided an excellent model with RMSE of 0.6892 and a correlation coefficient of 0.9987. The scatter plot of the predicted and the actual energy from physical potentials are drawn in Figure 5.7. The CE with the point functions alone provides a good model to predict the energies of the sequences in the WW domain structure. The energies vary from -37 kcal/mol to +18 kcal/mol and the range has a spread of 55 kcal/mol. Considering the range of 55 kcal/mol, the root-mean-squared error is only 0.6892 kcal/mol.

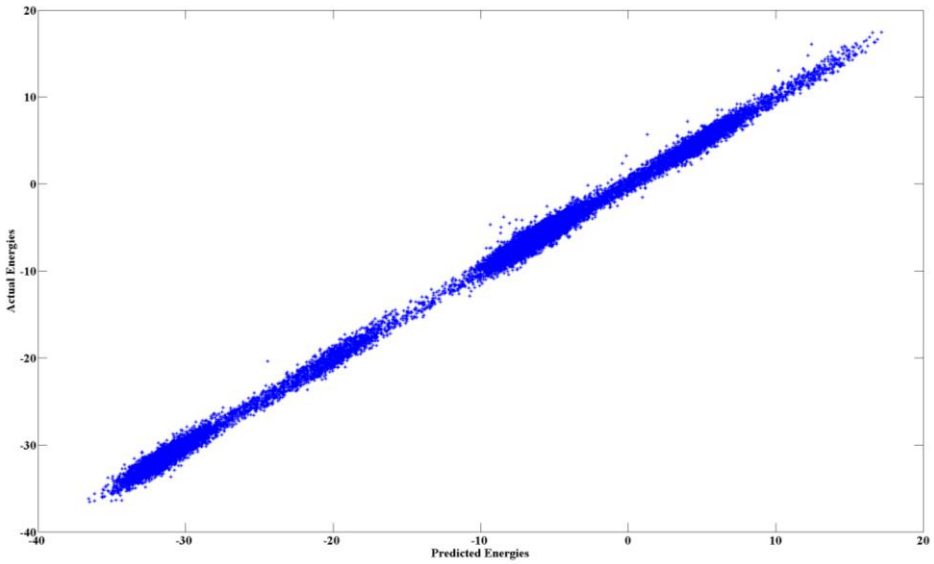

Figure 5.7 Scatter plot of predicted and actual Energies from point functions. The predicted energies are generated from CE while the actual energies are derived from Rosetta. This model is generated by including only point functions and shows excellent correlation

50,000 sequences were used to generate pair-wise interaction data and were divided as 35,000 sequences for training and 15,000 sequences for testing. The test sequences gave a RMSE of 0.6811 for point vectors and a correlation coefficient of 0.9987. The complete points and pair-wise model consisting of 95 point functions, 4253 pair functions and 1 constant function were modelled and tested. The pair vectors, when added to the model, gave a RMSE of 0.2929 and a correlation of 0.9998 (Figure 5.8). Although the point vectors alone gave a good fit for the predicted and actual energies, the addition of pair functions further reduced the root-mean-squared error. The RMSE of 0.2929 kcal/mol for the energy range of 55 kcal/mol confirms the accuracy of cluster expansion to model the protein interaction and energetics.

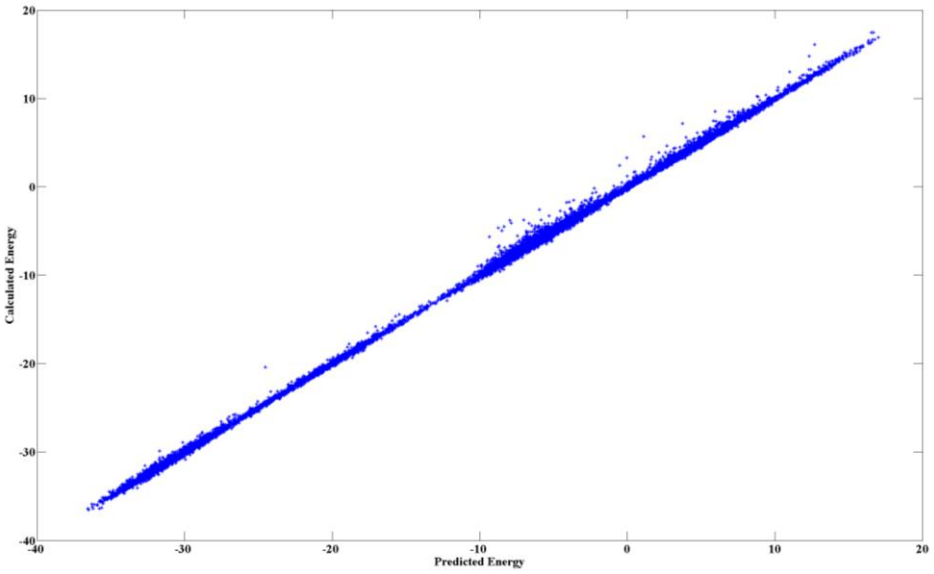

Figure 5.8 Scatter plot of predicted and actual energies from point and pair functions. The predicted energies are generated from CE while the actual energies are derived from Rosetta. This model is generated by including point and pair functions from CE.

Once the whole model was tested and its accuracy measured, the pair-wise interacting data was calculated. By using the point vectors alone, every set of pair functions that correspond to interaction of two positions were added and the RMSE was noted. A pair of positions had several cluster functions associated to it, depending on the mutability of the position. For example, position 2 had an amino-acid set of isoleucine, leucine and valine, while position 16 was varied between amino acids histidine, lysine and arginine. Hence, the pair positions 2 and 16 in the WW domain consisted of 9 cluster functions. The amino-acid set at different positions were derived from the MSA built from the SMART database [376]. The pair-wise interaction functions of positions 2 and 16 would consist of all possible individual pair-wise functions isoleucine at position 2 and histidine at position 16, leucine at 2 and histidine at position 16, etc. These functions are added to the model and the change in the RMSE is calculated. The pairs of positions which reduce the RMSE were tabulated and the bases for their interactions are listed in Table 5.1.

Position 1	Position 2	Interaction Basis
2	31	Close in Structure, N-C interaction 3.5 Å apart
7	20	Across the β-strand, H-bond 2.9 Å apart
7	22	Side-chain diagonal relationship across the β-strand
8	19	Across the β-strand but no H-bond
16	30	Not close in structure 12.1 Å
18	20	1-3 side-chain interaction on the same β-strand
18	27	Side-chain diagonal relationship across the β-strand
19	31	Close in space, side-chain interaction
20	22	1-3 SC interaction, but not in β-strand
20	25	Side-chain diagonal relationship across the β-strand
20	27	Across the β-strand but no H-bond
21	23	1-3 interaction, end of b-ladder and turn residues
28	30	1-3 interaction, end of β-ladder, weak H-bond
30	31	Consecutive turn residues

Table 5.1 Pairs of positions that show energetic coupling. The positions that show coupling are shown along with possible interactions between them based on the WW-domain structure.

From the residue interaction data from Table 5.1, we can group these residues into interacting clusters. Residues 2, 8, 16, 19, 28, 30 and 31 interact between each other and cluster together. Similarly, residues 7, 18, 20, 22, 25 and 27 form another cluster of interacting residues. Residues 21 and 23 have an interaction that cannot be grouped with either of the clusters. These clusters can be visualised in the WW domain structure in Figure 5.9

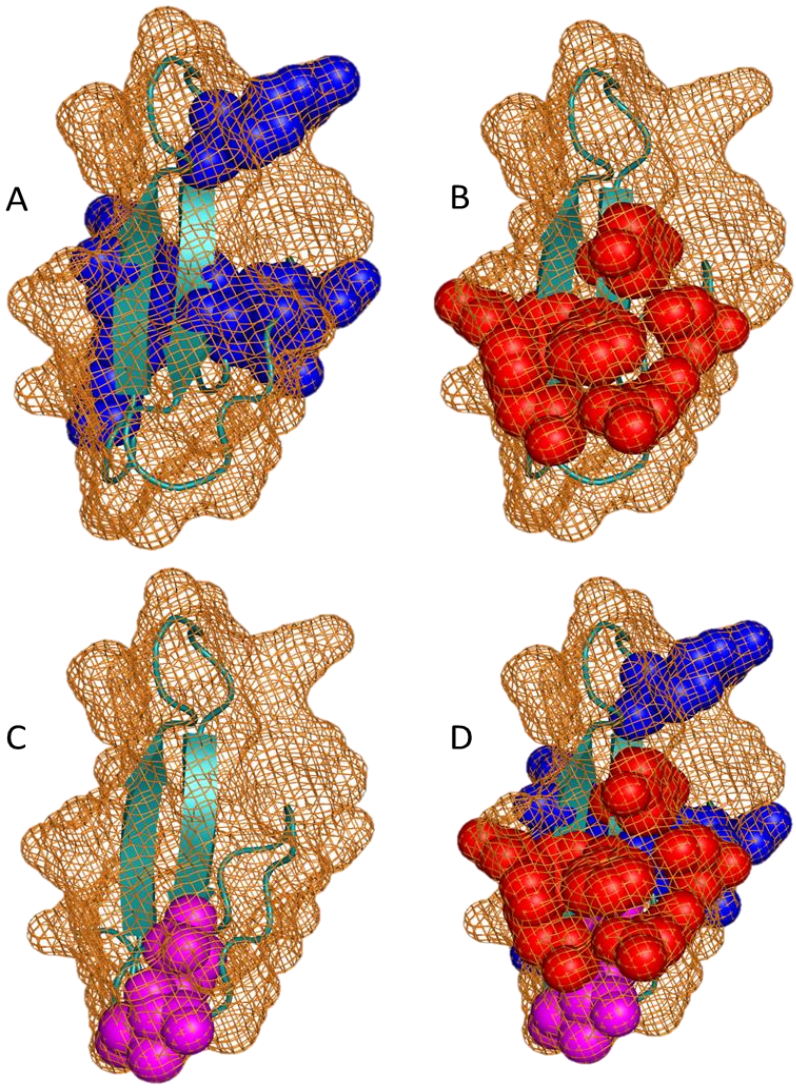

Figure 5.9 Clusters of interacting residues in the WW-domain. Cluster A (blue) comprising residues 2, 8, 16, 19, 28, 30 and 31; cluster B(red) with residues 7, 18, 20, 22, 25 and 27; and cluster C(pink) with residues 21 and 23 are shown in A, B and C, respectively. The structure of the WW domain (cyan and orange) is shown with all three clusters in D.

5.4 Discussion

5.4.1 CE and residue interactions

A new method to quantitatively evaluate the various interaction patterns in protein structure using CE is applied to the simple model of L22 β-strand. The RMSE for the energy of the model is extremely low. The β-strand is a simple model for which interaction patterns are easily verified. CE picks out the hydrogen bonding patterns and there is a lot of importance for side-chain interactions. This is due to the

repacking that happens with multiple amino acids. If the structure originally has small chained amino acids in some positions, replacing them with amino acids with bulkier side-chains can influence the interaction data. This is evident in the pair interaction between residues 2 and 9, which results in the high interaction values. Similar effects are seen in diagonal interactions and 1-3 interactions. Diagonal interactions between residues 3 and 6, residues 1 and 8 occur due to their side-chains being projected in the same direction and are an effect of the close proximity of the set of residues. The diagonal interaction might be an artefact of threading large amino acids in place of a smaller amino acid but showcases the structural makeup of the positions. The CE can be adopted to use a smaller and more specific set of residues at different positions according to the protein being studied. Also, the positional interaction data are averaged over all possible sets of amino acid permutations. As an example, for positions 1 and 10, the interaction data are an average of the whole set of functions like alanine at residue 1 and alanine at residue 10, alanine at residue 1 and valine at residue 10 (through all possible combinations). A thorough breakdown of every pair-wise position with respect of its amino acids is also possible. In the case of closely packed structures, bulkier amino acids might play a more important part while core residues might have preferences in terms of hydrophobicity. Such details can also be derived from the model. Similarly, compositional preferences are averaged over the sets of positions for a given pair of amino-acids and they can also be used to calculate preferences of specific sets of amino-acids at specific positions.

For the model to represent the protein, it has to be able to predict the energy of any sequence in the structure. Only after the CE is able to model the protein sequence energy, it can be broken down to calculate interaction details. Choosing amino acids that represent each position is important to obtain a good model. A multiple sequence alignment (MSA) is one of the best methods to choose the amino acid set for each position. A larger set of residues will increase detail but including non-suitable amino

acids can cause over-fitting errors. Select positions can also be varied, while leaving the rest with one amino acid choice for those positions. Larger set of amino acids and more positions provide additional detail at computation cost. As we go to pair and triplet functions for large proteins with the full residue set, the CE exponentially grows to beyond current computational powers.

This is an information model. CE is the dissection of the energy of the protein to a set of orthogonal functions, where every function provides information on the interactions. Including all functions is not a requirement since information can be provided by a subset alone. In case of the β-strand, the information provided by pair functions at position 1, 10 and 1, 8 can be related to the pair function at positions 8 and 10. This redundant information is calculated in the form of CE over-fitting and is removed. Due to this redundancy between functions, an approach similar to double mutant cycle is used to calculate pair-wise interactions [341]. The point functions that correspond to independent mutations are included in the model and every pair function is systematically added. If the two positions are energetically independent, the pair function would contain only redundant information and over-fit the model. This is determined by calculating the RMSE of the model using the test sequences. Over-fitting happens when the model becomes more complex than the property it actually represents and often due to having too many parameters and can be identified by the error in the training set compared to the test set. Any reduction of error in the training set which is not reflected in the test set is an evidence of over-fitting. So, by calculating the RMSE of the test set, the redundancy of the pair function can be accessed. Any pair of positions that can reduce the RMSE are energetically interacting and the measure of their interaction is the magnitude of reduction. A pair of positions that are highly dependent will have pair functions that greatly reduce the RMSE. This reduction is different for every model and is a relative measurement. Applying the concept to the β-strand, positions 1, 10 are energetically more

interacting than residues 3, 8. To avoid information overlap between different pairs of positions, each pair functions is individually added to the set of point functions and RMSE is calculated separately.

All current models of finding interactions do not account for sequence as they study the structure and the sequence as a whole unit. CE has been used to find interaction patterns inherent in the L22 β-strand and the WW domain structure independent of its sequence. This provides us with structural properties of the protein for which we can design sequences and understand energetic pathways between residues. Using the compositional preferences, we can understand which amino acids are more preferred and the positional preferences show us which positions are important. More importantly, we have analysed particular sets of functions to arrive at these conclusions. The model is so precise that it will be able to quantitatively show the effect of specific amino acids in their positions. For example, it is possible to calculate the preference of Arg in position 2 over Glu in position 2, or how Arg in position 2 is more important that Arg in positions 3, or how Leu in position 4 contributes to the structure compared to Ile in position 5. Furthermore, we can extend them to pairs of positions, such as preference of Tyr in position 3 and Trp in position 4 against Arg in position 3 and Lys in position 4. These would provide valuable starting points for mutational studies and for various structural experiments that look at important functional positions.

Using the same principles, the CE model was developed for the WW domain and the results are tabulated. A different of set of amino acids was used at every position, giving rise to 413 point functions. The point functions alone, with 300,000 training and 100,000 test sequences, gave a high RMSE of 189.1971 kcal/mol. The model had too many parameters and was too complex. The options were to increase the number of sequences (computationally intensive) or decrease the complexity of the model. MSA was used and the amino acid set was reduced to 95 point functions and these

gave an error of 0.6892 with only 140,000 training sequences and 60,000 test sequences. The amino acid set used does not have a large influence on the positional interaction data since the interactions of pairs of positions are carried out by considering all possible amino acid combinations in the model (refer Methods). Only after the model predicts the energy of the protein with excellent accuracy, the pair functions are added and the change in RMSE is noted. The positions that decrease the RMSE are tabulated (Table 5.1).

In the WW domain, the set of positions 2, 8, 16, 19 28, 30 and 31 form a cluster of energetically interacting positions. There are a few mutants of the WW domain that have been characterised and this experimental data can be used to verify these clusters. Position 19 is implicated in proline binding property of the WW domain and mutating it to a cysteine affects its ability to bind proline. The position 19 is part of the hydrophobic core [385] that is disrupted by mutating it [386]. Side-chain of position 2 participates in the hydrophobic core formation and mutating it to alanine significantly destabilises the structure at physiological temperature [385]. Similar destabilisation is evident when residue 24 is mutated to an Alanine, and shows that position 24 is important to the protein. Mutating position 19 to Alanine results in the protein becoming completely denatured [385].

The second set of positions 7, 18, 20, 22, 25 and 27 also forms another cluster that have similar functions. The positions 27 and 18 are known to form the XP groove in the protein, which is the binding surface for proline-rich ligands [387]. Similarly, positions 7, 18 and 20 are part of another binding surface, the XP2 groove of the protein [388], and can be expected to be energetically linked in the protein structure. Mutation at position 18 has also been shown to substantially unfold the protein [385].

A smaller cluster of two positions, 21 and 23 also have indispensable functions. Position 21 serves as an H-bond donor and acceptor and interacts with position 23.

The position 21 is also part of the hydrophobic core of the protein and mutating it to an alanine unfolds the protein [385].

In terms of clusters, cluster A consists of positions 2, 8, 16, 19, 28, 30 and 31, while cluster B consists of positions 7, 18, 20, 22, 25 and 27. Summarising the mutational data, positions 2 and 19 in Cluster A form hydrophobic core of the protein, which is destabilised when they are mutated. Positions 7, 18 and 20 form the XP2 groove of the protein, while positions 18 and 27 form the XP groove and hence contribute to proline binding property of the WW-domain. From these results, it is possible that these two clusters A and B correspond respectively to protein stability and binding of the WW-domain, although more mutational data is required to confirm these results.

It can be noted that the important positions 6, 29 and 32, which represent the two characteristic tryptophan residues and the important proline residue, have not been identified in the CE model. The positions were not altered in the model and hence their contributions are captured by the constant function in the expansion. Mutating these residues would compromise the identity of the domain, although it is possible to develop the model by varying them.

This current model was developed with the repacking energy and is limited by it. But the expansion is robust, and can be modified by the use of different energy calculations. This is demonstrated by using four different types of energies (total energy, van der Waals, electrostatics, and excluding electrostatics) in the β-strand model. These different energies are able to produce characteristic results. The use of only van der Waals energy should give importance to the bulky side-chained amino acids and shows that tryptophan is the major contributor to the van der Waals energy and so is tyrosine, to a certain extent, as well, which is expected. If electrostatics is removed, then there is not much difference and this may be because the structure itself is not electrostatically dependent. This would also explain the negative heat map for only electrostatic energy and the non-favourable lysine and arginine residue in it.

This type of breakdown of energy-based interactions is able to provide details that cannot be gathered from basic distance-based interaction analysis. This also provides a list of possible sequences with their total energies that can be used to understand how protein sequences affect energetics in a structure. In addition, there are tools like the Design Structure Matrix (DSM) that can use specific properties of the system to cluster and provide additional details on the system.

In theory, all these interactions provide information on the protein structure, but their practical use has to be tempered with caution. The model is only as good as the starting positions and amino acid sets. Mutating residues to non-specific amino acids might provide some hypothetical interactions that cannot happen *in vivo*. The structure, onto which the sequences are threaded, is also a very important consideration. A crystal structure of a protein in a different solvent or in the presence of a ligand can give rise to different results and can be easily misconstrued. On the other hand, the difference in pattern between the two structures can provide us important clues on how each of them are stabilised. Discerning the two different patterns might not be straightforward and requires careful data interpretation and experimental data such as binding energy calculated with specific mutations.

5.5 Conclusion

This study of protein structure and its energetics can further of our understanding of protein interactions and their mechanisms. The aim of this study is to construct an energetic interaction map for residues in a protein structure. Cluster expansion was applied to describe energetic interaction between residues, independent of its sequence. By breaking down complex interaction networks to orthogonal functions, compositional and positional interaction data is obtained. Specific coefficients in the expansion also predict double-mutant coupling energies between pairs of positions. The results show that energetic information inherent to the structure can be extracted by systematically mutating residues and measuring their energies using physical

potentials. The results identify energetically dependent clusters of amino acid positions, which are sequence independent structural features in proteins. The study presented here implies the possibility of computationally predicting protein interaction maps for a given structure. These results are noteworthy due to their statistical efficiency and energetic basis. The clusters of positions, identified by CE, correlate with experimental data on WW domain mutants. This approach is straightforward and applicable to other protein structures.

In summary, energetic coupling between different residual positions in proteins have been identified by representing the energy of the protein using cluster expansion. This energetic coupling is used to describe interacting network of positions in proteins and is verified by experimental mutant data in the WW domain.

6. Conclusion and Future Directions

The primary aims of the thesis are fourfold and revolve around the central theme of obtaining structural insights into different protein states using ensemble sampling and cluster expansion. These computational approaches deepen our understanding of these states and mechanisms, which pose methodological challenges to current experimental techniques. Ensemble modelling is used to describe the denatured, unfolded and nascent protein states while energetic interactions in folded protein structures are analysed by cluster expansion.

1. **Identifying residual structures in chemically denatured proteins**

In Chapter 2, the chemically denatured state of protein was described by ensemble modelling and its residual structures were identified. Ensemble sampling with 50% extended and 50% coil conformations accurately models the denatured Rgyr. The predictive model also follows Flory's relationship, which is a test of self-consistency for unstructured protein structures. Native-like residual structures are identified and tabulated for eighteen proteins. Short structural motifs, like beta-turn-helix, and secondary structural elements, like the β-turn, are considerably populated in the DSEs. Experimentally determined residual structures are largely shown to be in agreement with this model.

The residual structure identified by DSEs can be verified by amide deuterium-exchange mass spectroscopy and insights could help in refining the ensemble model. Studying polar characteristics of residual structures and their effect on nucleation could also provide important clues on how residual structures affect folding thermodynamics.

2. Understanding ribosome tunnel and nascent polypeptide interactions

Chapter 3 addressed the question of nascent polypeptide-induced tunnel geometry and the effect of spatial constraints of the exit tunnel on the nascent polypeptide conformation.

Molecular dynamics simulations of ribosome in the presence and absence of the nascent polypeptide reveal an increase in tunnel convolution due to the polypeptide. TraDES *de novo* sampling, with tunnel surface as a geometric constraint, shows that the tunnel is capable of accommodating different conformations of amino acids in its different segments. The tunnel segment corresponding to residue 15 shows increased spatial freedom and results in structures that fold inside the tunnel. The method of ensemble sampling with surface constraints and analysing them with dock by superposition has also been demonstrated to provide structural ensembles of the peptide that fit in the tunnel.

This method can be extended to use multiple snapshots of the tunnel, when it is expanded or collapsed. The difference in resultant polypeptide geometry and conformation would show how change in tunnel geometry affects the nascent polypeptide. Random expulsion molecular dynamics can be alternatively used to study tunnel geometry during peptide expulsion.

3. Modelling intrinsic disordered proteins using ensemble sampling

In Chapter 4, an updated version of TraDES is used to generate structural representations of disordered proteins by sampling the random-coil regions of protein conformational space. Great agreement with other sampling tools and good agreement with experimental data is shown by comparing the Rgyr of the proteins.

4. Unravelling energetic interactions in proteins using cluster expansion

Chapter 5 demonstrates the use of cluster expansion to model energetic interactions in protein structure. The ability of cluster expansion to calculate energy of different

sequences in a protein structure is demonstrated using a simple model of an anti-parallel β-strand. Positional and compositional interaction data are also determined from the expansion. This CE model is used to find pair of positions that are energetically coupled in the WW domain. This establishes the presence of interacting clusters of residual positions in the protein structure, which can correspond to different functions of the protein such as stability and binding. Mutation data from experiments also verify the important residues identified by CE.

Using molecular dynamics and *in silico* mutants, the importance of positions and their effects on binding and stability could be verified. Analysing triplet functions and deriving clusters based on them should also improve the accuracy of CE.

7. References

1. Tanford C: **Protein denaturation.** *Adv Protein Chem* 1968, **23**:121-282.

2. Flory PJ: **Principles of Polymer Chemistry.** *Cornell University Press, Ithaca, NY.* 1953,

3. Anson ML, Mirsky AE: **The Effect of Denaturation on the Viscosity of Protein Systems.** *J Gen Physiol* 1932, **15(3)**:341-350.

4. Pauling L, Corey RB: **Atomic coordinates and structure factors for two helical configurations of polypeptide chains.** *Proc Natl Acad Sci U S A* 1951, **37(5)**:235-240.

5. Neurath H, Greenstein JP, Putnam FW, Erickson JA: **The Chemistry of Protein Denaturation.** *Chem Rev* 1944, **34(2)**:157-265.

6. Ptitsyn OB, Pain RH, Semisotnov GV, Zerovnik E, Razgulyaev OI: **Evidence for a molten globule state as a general intermediate in protein folding.** *FEBS Lett* 1990, **262(1)**:20-24.

7. Ptitsyn OB, Uversky VN: **The molten globule is a third thermodynamical state of protein molecules.** *FEBS Lett* 1994, **341(1)**:15-18.

8. Ptitsyn OB: **Molten globule and protein folding.** *Adv Protein Chem* 1995, **47**:83-229.

9. Schweers O, Schönbrunn-Hanebeck E, Marx A, Mandelkow E: **Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure.** *J Biol Chem* 1994, **269(39)**:24290-24297.

10. Weinreb PH, Zhen W, Poon AW, Conway KA, Lansbury PT: **NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded.** *Biochemistry* 1996, **35(43)**:13709-13715.

11. Shortle D: **Staphylococcal nuclease: a showcase of m-value effects.** *Adv Protein Chem* 1995, **46**:217-247.

12. Neri D, Billeter M, Wider G, Wüthrich K: **NMR determination of residual structure in a urea-denatured protein, the 434-repressor.** *Science* 1992, **257(5076)**:1559-1563.

13. Langmuir I: **Protein Denaturation.** *Cold Spring Harb. Symp. quant. Biol* 1938, **6(159)**

14. Foster JF, Samsa EG: **Streaming Orientation Studies on Denatured Proteins. I. Heat Denaturation of Ovalbumin in Acid Media1.** *J Am Chem Soc* 1951, **73(7)**:3187-3190.

15. Timasheff SN, Gibbs RJ: **The state of plasma albumin in acid pH.** *Arch Biochem Biophys* 1957, **70(2)**:547-560.

16. Lumry R, Eyring H: **Conformation Changes of Proteins.** *J Phys Chem* 1954, **58(2)**:110-120.

17. Pauling L, Corey RB, Branson HR: **The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain.** *Proc Natl Acad Sci U S A* 1951, **37(4)**:205-211.

18. Mirsky AE, Pauling L: **On the Structure of Native, Denatured, and Coagulated Proteins.** *Proc Natl Acad Sci U S A* 1936, **22(7)**:439-447.

19. Anfinsen CB: **Principles that govern the folding of protein chains.** *Science* 1973, **181(4096)**:223-230.

20. WHITE FH: **Regeneration of native secondary and tertiary structures by air oxidation of reduced ribonuclease.** *J Biol Chem* 1961, **236**:1353-1360.

21. Epstein HF, Schechter AN, Chen RF, Anfinsen CB: **Folding of staphylococcal nuclease: kinetic studies of two processes in acid renaturation.** *J Mol Biol* 1971, **60(3)**:499-508.

22. ANFINSEN CB, HABER E: **Studies on the reduction and re-formation of protein disulfide bonds.** *J Biol Chem* 1961, **236**:1361-1363.

23. Tanford C, Kawahara K, Lapanje S: **Proteins in 6-M guanidine hydrochloride. Demonstration of random coil behavior.** *J Biol Chem* 1966, **241(8)**:1921-1923.

24. Yao J, Chung J, Eliezer D, Wright PE, Dyson HJ: **NMR structural and dynamic characterization of the acid-unfolded state of apomyoglobin provides insights into the early events in protein folding.** *Biochemistry* 2001, **40(12)**:3561-3571.

25. Garcia P, Serrano L, Durand D, Rico M, Bruix M: **NMR and SAXS characterization of the denatured state of the chemotactic protein CheY: implications for protein folding initiation.** *Protein Sci* 2001, **10(6)**:1100-1112.

26. Sari N, Alexander P, Bryan PN, Orban J: **Structure and dynamics of an acid-denatured protein G mutant.** *Biochemistry* 2000, **39(5)**:965-977.

27. Zhang X, Xu Y, Zhang J, Wu J, Shi Y: **Structural and dynamic characterization of the acid-unfolded state of hUBF HMG box 1 provides clues for the early events in protein folding.** *Biochemistry* 2005, **44(22)**:8117-8125.

28. Kumar A, Srivastava S, Mishra RK, Mittal R, Hosur RV: **Local structural preferences and dynamics restrictions in the urea-denatured state of SUMO-1: NMR characterization.** *Biophys J* 2006, **90(7)**:2498-2509.

29. Modig K, Jürgensen VW, Lindorff-Larsen K, Fieber W, Bohr HG, Poulsen FM: **Detection of initiation sites in protein folding of the four helix bundle ACBP by chemical shift analysis.** *FEBS Lett* 2007, **581(25)**:4965-4971.

30. Kazmirski SL, Wong KB, Freund SM, Tan YJ, Fersht AR, Daggett V: **Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution.** *Proc Natl Acad Sci U S A* 2001, **98(8)**:4349-4354.

31. Robic S, Guzman-Casado M, Sanchez-Ruiz JM, Marqusee S: **Role of residual structure in the unfolded state of a thermophilic protein.** *Proc Natl Acad Sci U S A* 2003, **100(20)**:11345-11349.

32. Morrone A, McCully ME, Bryan PN, Brunori M, Daggett V, Gianni S, Travaglini-Allocatelli C: **The denatured state dictates the topology of two proteins with almost identical sequence but different native structure and function.** *J Biol Chem* 2011, **286(5)**:3863-3872.

33. Bruun S: **Cooperative formation of native-like tertiary contacts in the ensemble of unfolded states of a four-helix protein.** *Proceedings of the National Academy of Sciences* 2010, **107(30)**:13306-13311.

34. Lietzow MA, Jamin M, Dyson HJ, Wright PE: **Mapping long-range contacts in a highly unfolded protein.** *J Mol Biol* 2002, **322(4)**:655-662.

35. Shortle D, Ackerman MS: **Persistence of native-like topology in a denatured protein in 8 M urea.** *Science* 2001, **293(5529)**:487-489.

36. Huang JR, Gabel F, Jensen MR, Grzesiek S, Blackledge M: **Sequence-specific mapping of the interaction between urea and unfolded ubiquitin from ensemble analysis of NMR and small angle scattering data.** *J Am Chem Soc* 2012, **134(9)**:4429-4436.

37. Berteotti A, Barducci A, Parrinello M: **Effect of urea on the β-hairpin conformational ensemble and protein denaturation mechanism.** *J Am Chem Soc* 2011, **133(43)**:17200-17206.

38. Eberini I, Emerson A, Sensi C, Ragona L, Ricchiuto P, Pedretti A, Gianazza E, Tramontano A: **Simulation of urea-induced protein unfolding: a lesson from bovine β-lactoglobulin.** *J Mol Graph Model* 2011, **30**:24-30.

39. Dill KA, Shortle D: **Denatured states of proteins.** *Annu Rev Biochem* 1991, **60**:795-825.

40. Uversky VN: **Natively unfolded proteins: a point where biology waits for physics.** *Protein Sci* 2002, **11(4)**:739-756.

41. Pace CN, Laurents DV, Erickson RE: **Urea denaturation of barnase: pH dependence and characterization of the unfolded state.** *Biochemistry* 1992, **31(10)**:2728-2734.

42. Lindman S, Linse S, Mulder FAA, André I: **pK(a) values for side-chain carboxyl groups of a PGB1 variant explain salt and pH-dependent stability.** *Biophys J* 2007, **92(1)**:257-266.

43. Kuhlman B, Luisi DL, Young P, Raleigh DP: **pKa values and the pH dependent stability of the N-terminal domain of L9 as probes of electrostatic interactions in the denatured state. Differentiation between local and nonlocal interactions.** *Biochemistry* 1999, **38(15)**:4896-4903.

44. Cho JH, Sato S, Raleigh DP: **Thermodynamics and kinetics of non-native interactions in protein folding: a single point mutant significantly stabilizes the N-terminal domain of L9 by modulating non-native interactions in the denatured state.** *J Mol Biol* 2004, **338(4)**:827-837.

45. Bowler BE: **Residual structure in unfolded proteins. (2).** *Curr Opin Struct Biol* 2012, **22(1)**:4-13.

46. Whitten ST, García-Moreno E B: **pH dependence of stability of staphylococcal nuclease: evidence of substantial electrostatic interactions in the denatured state.** *Biochemistry* 2000, **39(46)**:14292-14304.

47. Cho JH, Sato S, Horng JC, Anil B, Raleigh DP: **Electrostatic interactions in the denatured state ensemble: their effect upon protein folding and protein stability.** *Arch Biochem Biophys* 2008, **469(1)**:20-28.

48. Aune KC, Salahuddin A, Zarlengo MH, Tanford C: **Evidence for residual structure in acid- and heat-denatured proteins.** *J Biol Chem* 1967, **242(19)**:4486-4489.

49. Gillespie JR, Shortle D: **Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels.** *J Mol Biol* 1997, **268(1)**:158-169.

50. Millett I, Doniach S, Plaxco K: **Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins. (1).** *Adv Protein Chem* 2002, **62**:241-262.

51. Bowler BE: **Thermodynamics of protein denatured states.** *Mol Biosyst* 2007, **3(2)**:88-99.

52. Marsh JA, Forman-Kay JD: **Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints.** *J Mol Biol* 2009, **391(2)**:359-374.

53. Allison JR, Varnai P, Dobson CM, Vendruscolo M: **Determination of the free energy landscape of alpha-synuclein using spin label nuclear magnetic resonance measurements.** *J Am Chem Soc* 2009, **131(51)**:18314-18326.

54. Rösner HI, Poulsen FM: **Residue-specific description of non-native transient structures in the ensemble of acid-denatured structures of the all-beta protein c-src SH3.** *Biochemistry* 2010, **49(15)**:3246-3253.

55. Dedmon MM, Christodoulou J, Wilson MR, Dobson CM: **Heat shock protein 70 inhibits alpha-synuclein fibril formation via preferential binding to prefibrillar species.** *J Biol Chem* 2005, **280(15)**:14733-14740.

56. Vise P, Baral B, Stancik A, Lowry DF, Daughdrill GW: **Identifying long-range structure in the intrinsically unstructured transactivation domain of p53.** *Proteins* 2007, **67(3)**:526-530.

57. Bertoncini CW, Jung YS, Fernandez CO, Hoyer W, Griesinger C, Jovin TM, Zweckstetter M: **Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein.** *Proc Natl Acad Sci U S A* 2005, **102(5)**:1430-1435.

58. Bernadó P, Bertoncini CW, Griesinger C, Zweckstetter M, Blackledge M: **Defining long-range order and local disorder in native alpha-synuclein using residual dipolar couplings.** *J Am Chem Soc* 2005, **127(51)**:17968-17969.

59. Louhivuori M, Pääkkönen K, Fredriksson K, Permi P, Lounila J, Annila A: **On the origin of residual dipolar couplings from denatured proteins.** *J Am Chem Soc* 2003, **125(50)**:15647-15650.

60. Fernández CO, Hoyer W, Zweckstetter M, Jares-Erijman EA, Subramaniam V, Griesinger C, Jovin TM: **NMR of alpha-synuclein-polyamine complexes elucidates the mechanism and kinetics of induced aggregation.** *EMBO J* 2004, **23(10)**:2039-2046.

61. Kolb V, Makeyev E, Spirin A: **Folding of firefly luciferase during translation in a cell-free system.** *EMBO J* 1994, **13(15)**:3631-3637.

62. Hartl FU, Hayer-Hartl M: **Molecular chaperones in the cytosol: from nascent chain to folded protein.** *Science* 2002, **295(5561)**:1852-1858.

63. Ugrinov KG, Clark PL: **Cotranslational folding increases GFP folding yield.** *Biophys J* 2010, **98(7)**:1312-1320.

64. Svetlov M, Kommer A, Kolb V, Spirin A: **Effective cotranslational folding of firefly luciferase without chaperones of the Hsp70 family.** *Protein Sci* 2006, **15(2)**:242-247.

65. Kolb VA: **Co-translational folding of an eukaryotic multidomain protein in a prokaryotic translation system.** *J Biol Chem* 2000, **275(22)**:16597-16601.

66. Kolb VA, Makeyev EV, Spirin AS: **Co-translational folding of an eukaryotic multidomain protein in a prokaryotic translation system.** *J Biol Chem* 2000, **275(22)**:16597-16601.

67. Manning M, Colón W: **Structural basis of protein kinetic stability: resistance to sodium dodecyl sulfate suggests a central role for rigidity and a bias toward beta-sheet structure.** *Biochemistry* 2004, **43(35)**:11248-11254.

68. Bryngelson JD, Wolynes PG: **Spin glasses and the statistical mechanics of protein folding.** *Proc Natl Acad Sci U S A* 1987, **84(21)**:7524-7528.

69. Lu J, Deutsch C: **Folding zones inside the ribosomal exit tunnel.** *Nature Structural & Molecular Biology* 2005, **12(12)**:1123-1129.

70. Kosolapov A, Deutsch C: **Tertiary interactions within the ribosomal exit tunnel.** *Nat Struct Mol Biol* 2009, **16(4)**:405-411.

71. Voss N, Gerstein M, Steitz T, Moore P: **The geometry of the ribosomal polypeptide exit tunnel. (1).** *J Mol Biol* 2006, **360(4)**:893-906.

72. Gilbert RJC, Fucini P, Connell S, Fuller SD, Nierhaus KH, Robinson CV, Dobson CM, Stuart DI: **Three-dimensional structures of translating ribosomes by Cryo-EM.** *Mol Cell* 2004, **14(1)**:57-66.

73. Gabashvili IS, Gregory ST, Valle M, Grassucci R, Worbs M, Wahl MC, Dahlberg AE, Frank J: **The polypeptide tunnel system in the ribosome and its gating in erythromycin resistance mutants of L4 and L22.** *Mol Cell* 2001, **8(1)**:181-188.

74. Makeyev EV, Kolb VA, Spirin AS: **Enzymatic activity of the ribosome-bound nascent polypeptide.** *FEBS Lett* 1996, **378(2)**:166-170.

75. Kudlicki W, Chirgwin J, Kramer G, Hardesty B: **Folding of an enzyme into an active conformation while bound as peptidyl-tRNA to the ribosome.** *Biochemistry* 1995, **34(44)**:14284-14287.

76. Kolb VA, Makeyev EV, Kommer A, Spirin AS: **Cotranslational folding of proteins.** *Biochem Cell Biol* 1995, **73(11-12)**:1217-1220.

77. Fischer E: **Einfluss der configuration auf die wirkung der enzyme.** *Ber Dt Chem Ges 27 p* 1894, :2985-2993.

78. Kendrew J: **A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.** *Nature p* 1958, **181(4610)**:662-666.

79. Kendrew J: **Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A.** *resolution Nature p* 1960, **185(4711)**:422-427.

80. Koshland D: *The Key-Lock Theory and the Induced Fit Theory.* 1994.

81. Uversky VN, Dunker AK: **Understanding protein non-folding.** *Biochim Biophys Acta* 2010, **1804(6)**:1231-1264.

82. Tompa P: **Intrinsically unstructured proteins.** *Trends Biochem Sci* 2002, **27(10)**:527-533.

83. Dyson HJ, Wright PE: **Intrinsically unstructured proteins and their functions.** *Nat Rev Mol Cell Biol* 2005, **6(3)**:197-208.

84. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z: **Intrinsically disordered protein.** *J Mol Graph Model* 2001, **19(1)**:26-59.

85. Friedler A, Veprintsev DB, Freund SMV, von Glos KI, Fersht AR: **Modulation of binding of DNA to the C-terminal domain of p53 by acetylation.** *Structure* 2005, **13(4)**:629-636.

86. Mylonas E, Hascher A, Bernadó P, Blackledge M, Mandelkow E, Svergun DI: **Domain conformation of tau protein studied by solution small-angle X-ray scattering.** *Biochemistry* 2008, **47(39)**:10345-10353.

87. Uversky VN: **What does it mean to be natively unfolded?** *Eur J Biochem* 2002, **269(1)**:2-12.

88. Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE: **Protein disorder and the evolution of molecular recognition: theory, predictions and observations.** *Pac Symp Biocomput* 1998, :473-484.

89. Uversky VN, Gillespie JR, Fink AL: **Why are "natively unfolded" proteins unstructured under physiologic conditions?** *Proteins* 2000, **41(3)**:415-427.

90. Tompa P: **The interplay between structure and function in intrinsically unstructured proteins.** *FEBS Lett* 2005, **579(15)**:3346-3354.

91. Uversky VN: **Intrinsically disordered proteins from A to Z.** *Int J Biochem Cell Biol* 2011, :.

92. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM: **Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes.** *PLoS Comput Biol* 2006, **2(8)**:e100.

93. Patil A, Nakamura H: **Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks.** *FEBS Lett* 2006, **580(8)**:2041-2045.

94. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN: **Flexible nets. The roles of intrinsic disorder in protein interaction networks.** *FEBS J* 2005, **272(20)**:5129-5148.

95. Uversky VN, Oldfield CJ, Dunker AK: **Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling.** *J Mol Recognit* 2005, **18(5)**:343-384.

96. Uversky VN, Oldfield CJ, Dunker AK: **Intrinsically disordered proteins in human diseases: introducing the D2 concept.** *Annual review of biophysics* 2008, **37**:215-246.

97. Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN: **Protein disorder in the human diseasome: unfoldomics of human genetic diseases.** *BMC Genomics* 2009, **10 Suppl 1**:S12.

98. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN: **Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins.** *J Proteome Res* 2007, **6(5)**:1917-1932.

99. Dunker AK, Uversky VN: **Drugs for 'protein clouds': targeting intrinsically disordered transcription factors.** *Curr Opin Pharmacol* 2010, **10(6)**:782-788.

100. Mittag T, Forman-Kay JD: **Atomic-level characterization of disordered protein ensembles.** *Curr Opin Struct Biol* 2007, **17(1)**:3-14.

101. McCarney ER, Kohn JE, Plaxco KW: **Is there or isn't there? The case for (and against) residual structure in chemically denatured proteins.** *Crit Rev Biochem Mol Biol* 2005, **40(4)**:181-189.

102. Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM: **Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations.** *J Am Chem Soc* 2005, **127(2)**:476-477.

103. Fuxreiter M, Simon I, Friedrich P, Tompa P: **Preformed structural elements feature in partner recognition by intrinsically unstructured proteins.** *J Mol Biol* 2004, **338(5)**:1015-1026.

104. Loris R, Marianovsky I, Lah J, Laeremans T, Engelberg-Kulka H, Glaser G, Muyldermans S, Wyns L: **Crystal structure of the intrinsically flexible addiction antidote MazE.** *J Biol Chem* 2003, **278(30)**:28252-28257.

105. Buck M: **Crystallography: embracing conformational flexibility in proteins.** *Structure* 2003, **11(7)**:735-736.

106. Zhan Y, Song X, Zhou GW: **Structural analysis of regulatory protein domains using GST-fusion proteins.** *Gene* 2001, **281(1-2)**:1-9.

107. Glatter O, Angle X: **O. Small ray Scattering.** *England London* 1982,

108. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI: **Structural characterization of flexible proteins using small-angle X-ray scattering.** *J Am Chem Soc* 2007, **129(17)**:5656-5664.

109. Doniach S, Bascle J, Garel T, Orland H: **Partially folded states of proteins: characterization by X-ray scattering.** *J Mol Biol* 1995, **254(5)**:960-967.

110. Barbar E: **NMR characterization of partially folded and unfolded conformational ensembles of proteins.** *Biopolymers* 1999, **51(3)**:191-207.

111. Bracken C: **NMR spin relaxation methods for characterization of disorder and folding in proteins.** *J Mol Graph Model* 2001, **19(1)**:3-12.

112. Receveur-Bréchot V, Bourhis JM, Uversky VN, Canard B, Longhi S: **Assessing protein disorder and induced folding.** *Proteins* 2006, **62(1)**:24-45.

113. Wishart DS, Sykes BD, Richards FM: **The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy.** *Biochemistry* 1992, **31(6)**:1647-1651.

114. Wishart DS, Sykes BD: **Chemical shifts as a tool for structure determination.** *Methods Enzymol* 1994, **239**:363-392.

115. Wishart DS, Sykes BD, Richards FM: **Relationship between nuclear magnetic resonance chemical shift and protein secondary structure.** *J Mol Biol* 1991, **222(2)**:311-333.

116. Jensen MR, Salmon L, Nodet G, Blackledge M: **Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts.** *J Am Chem Soc* 2010, **132(4)**:1270-1272.

117. Ganguly D, Chen J: **Structural interpretation of paramagnetic relaxation enhancement-derived distances for disordered protein states.** *J Mol Biol* 2009, **390(3)**:467-477.

118. Nodet G, Salmon L, Ozenne V, Meier S, Jensen MR, Blackledge M: **Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings.** *J Am Chem Soc* 2009, **131(49)**:17908-17918.

119. Jensen MR, Markwick PRL, Meier S, Griesinger C, Zweckstetter M, Grzesiek S, Bernadó P, Blackledge M: **Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings.** *Structure* 2009, **17(9)**:1169-1185.

120. Dyson HJ, Wright PE: **Unfolded proteins and protein folding studied by NMR.** *Chem Rev* 2004, **104(8)**:3607-3622.

121. Huang F, Rajagopalan S, Settanni G, Marsh RJ, Armoogum DA, Nicolaou N, Bain AJ, Lerner E, Haas E, Ying L, Fersht AR: **Multiple conformations of full-length p53 detected with single-molecule fluorescence resonance energy transfer.** *Proc Natl Acad Sci U S A* 2009, **106(49)**:20758-20763.

122. Haas E: **Ensemble FRET methods in studies of intrinsically disordered proteins.** *Methods Mol Biol* 2012, **895**:467-498.

123. Ohashi T, Galiacy SD, Briscoe G, Erickson HP: **An experimental study of GFP-based FRET, with application to intrinsically unstructured proteins.** *Protein Sci* 2007, **16(7)**:1429-1438.

124. Schuler B, Müller-Späth S, Soranno A, Nettels D: **Application of confocal single-molecule FRET to intrinsically disordered proteins.** *Methods Mol Biol* 2012, **896**:21-45.

125. Bernadó P, Blanchard L, Timmins P, Marion D, Ruigrok RWH, Blackledge M: **A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering.** *Proc Natl Acad Sci U S A* 2005, **102(47)**:17002-17007.

126. Mendoza C, Figueirido F, Tasayco ML: **DSC studies of a family of natively disordered fragments from Escherichia coli thioredoxin: surface burial in intrinsic coils.** *Biochemistry* 2003, **42(11)**:3349-3358.

127. Fisher CK, Huang A, Stultz CM: **Modeling intrinsically disordered proteins with bayesian statistics.** *J Am Chem Soc* 2010, **132(42)**:14919-14927.

128. Ferreon ACM, Moran CR, Gambin Y, Deniz AA: **Single-molecule fluorescence studies of intrinsically disordered proteins.** *Methods Enzymol* 2010, **472**:179-204.

129. Luque I, Leavitt SA, Freire E: **The linkage between protein folding and functional cooperativity: two sides of the same coin?** *Annu Rev Biophys Biomol Struct* 2002, **31**:235-256.

130. Smock RG, Gierasch LM: **Sending signals dynamically.** *Science* 2009, **324(5924)**:198-203.

131. Cui Q, Karplus M: **Allostery and cooperativity revisited.** *Protein Sci* 2008, **17(8)**:1295-1307.

132. Koshland DE, Némethy G, Filmer D: **Comparison of experimental binding data and theoretical models in proteins containing subunits.** *Biochemistry* 1966, **5(1)**:365-385.

133. Monod J, Wyman J, Changeux JP: **On the nature of allosteric transitions: A plausible model.** *J Mol Biol* 1965, **12**:88-118.

134. Volkman BF, Lipson D, Wemmer DE, Kern D: **Two-state allosteric behavior in a single-domain signaling protein.** *Science* 2001, **291(5512)**:2429-2433.

135. Cooper A, Dryden DT: **Allostery without conformational change. A plausible model.** *Eur Biophys J* 1984, **11(2)**:103-109.

136. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science* 1999, **286(5438)**:295-299.

137. Halabi N, Rivoire O, Leibler S, Ranganathan R: **Protein sectors: evolutionary units of three-dimensional structure.** *Cell* 2009, **138(4)**:774-786.

138. Reynolds KA, McLaughlin RN, Ranganathan R: **Hot spots for allosteric regulation on protein surfaces.** *Cell* 2011, **147(7)**:1564-1575.

139. Lee BC, Kim D: **A new method for revealing correlated mutations under the structural and functional constraints in proteins.** *Bioinformatics* 2009, **25(19)**:2506-2513.

140. Chi CN, Elfström L, Shi Y, Snäll T, Engström A, Jemth P: **Reassessing a sparse energetic network within a single protein domain.** *Proc Natl Acad Sci U S A* 2008, **105(12)**:4679-4684.

141. **So much more to know.** *Science* 2005, **309(5731)**:78-102.

142. Dobson CM: **Protein folding and misfolding.** *Nature* 2003, **426(6968)**:884-890.

143. Daggett V, Fersht A: **The present view of the mechanism of protein folding.** *Nat Rev Mol Cell Biol* 2003, **4(6)**:497-502.

144. Levinthal C: **Are there pathways for protein folding.** *J. Chim. phys* 1968, **65(1)**:44-45.

145. Srinivasan R, Rose GD: **Ab initio prediction of protein structure using LINUS.** *Proteins* 2002, **47(4)**:489-495.

146. 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: **Structure prediction for CASP8 with all-atom refinement using Rosetta.** *Proteins* 2009, **77 Suppl 9**:89-99.

147. Toxvaerd S: **Stability of molecular dynamics simulations of classical systems.** *J Chem Phys* 2012, **137(21",)**:.

148. Schiffer CA, Caldwell JW, Stroud RM, Kollman PA: **Inclusion of solvation free energy with molecular mechanics energy: alanyl dipeptide as a test case.** *Protein Sci* 1992, **1(3)**:396-400.

149. Arnold GE, Ornstein RL: **An evaluation of implicit and explicit solvent model systems for the molecular dynamics simulation of bacteriophage T4 lysozyme.** *Proteins* 1994, **18(1)**:19-33.

150. Lazaridis T, Karplus M: **Effective energy function for proteins in solution.** *Proteins* 1999, **35(2)**:133-152.

151. Eisenberg D, McLachlan AD: **Solvation energy in protein folding and binding.** *Nature* 1986, **319(6050)**:199-203.

152. Wesson L, Eisenberg D: **Atomic solvation parameters applied to molecular dynamics of proteins in solution.** *Protein Sci* 1992, **1(2)**:227-235.

153. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML: **Comparison of simple potential functions for simulating liquid water.** *J Chem Phys* 1983, **79(2",)**:926-935.

154. Feldman HJ, Hogue CWV: **Probabilistic sampling of protein conformations: new hope for brute force?** *Proteins* 2002, **46(1)**:8-23.

155. Feldman HJ, Hogue CW: **A fast method to sample real protein conformational space.** *Proteins* 2000, **39(2)**:112-131.

156. Lyklema JW, Kremer K: **The growing self avoiding walk.** *Journal of Physics A: Mathematical and General* 1984, **17(13)**:L691.

157. Majid I, Jan N, Coniglio A, Stanley HE: **Kinetic Growth Walk: A New Model for Linear Polymers.** *Phys. Rev. Lett.* 1984, **52**:1257-1260.

158. Shapovalov MV, Dunbrack RL: **A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions.** *Structure* 2011, **19(6)**:844-858.

159. Plaxco KW, Millett IS, Segel DJ, Doniach S, Baker D: **Chain collapse can occur concomitantly with the rate-limiting step in protein folding.** *Nat Struct Biol* 1999, **6(6)**:554-556.

160. Millett IS, Doniach S, Plaxco KW: **Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins.** *Adv Protein Chem* 2002, **62**:241-262.

161. Bowler BE: **Residual structure in unfolded proteins.** *Curr Opin Struct Biol* 2012, **22(1)**:4-13.

162. Chang JY: **Structural heterogeneity of 6 M GdmCl-denatured proteins: implications for the mechanism of protein folding.** *Biochemistry* 2009, **48(40)**:9340-9346.

163. Wildes D, Anderson LM, Sabogal A, Marqusee S: **Native state energetics of the Src SH2 domain: evidence for a partially structured state in the denatured ensemble.** *Protein Sci* 2006, **15(7)**:1769-1779.

164. Muff S, Caflisch A: **Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein.** *Proteins: Structure, Function, and Bioinformatics* 2008, **70(4)**:1185-1195.

165. Smith C, Bu Z, Engelman D, Regan L, Anderson K, Sturtevant J: **Surface point mutations that significantly alter the structure and stability of a protein's denatured state.** *Protein science* 1996, **5(10)**:2009-2019.

166. Shortle D: **The denatured state (the other half of the folding equation) and its role in protein stability.** *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 1996, **10(1)**:27-34.

167. Choy WY, Mulder FA, Crowhurst KA, Muhandiram D, Millett IS, Doniach S, Forman-Kay JD, Kay LE: **Distribution of molecular size within an unfolded state ensemble using small-angle X-ray scattering and pulse field gradient NMR techniques.** *J Mol Biol* 2002, **316(1)**:101-112.

168. Hodsdon ME, Frieden C: **Intestinal fatty acid binding protein: the folding mechanism as determined by NMR studies.** *Biochemistry* 2001, **40(3)**:732-742.

169. Chugh J, Sharma S, Hosur RV: **Pockets of short-range transient order and restricted topological heterogeneity in the guanidine-denatured state ensemble of GED of dynamin.** *Biochemistry* 2007, **46(42)**:11819-11832.

170. Merchant KA, Best RB, Louis JM, Gopich IV, Eaton WA: **Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations.** *Proc Natl Acad Sci U S A* 2007, **104(5)**:1528-1533.

171. Matsuo K, Sakurada Y, Yonehara R, Kataoka M, Gekko K: **Secondary-structure analysis of denatured proteins by vacuum-ultraviolet circular dichroism spectroscopy.** *Biophys J* 2007, **92(11)**:4088-4096.

172. Wang HM, Yu C: **Investigating the refolding pathway of human acidic fibroblast growth factor (hFGF-1) from the residual structure(s) obtained by denatured-state hydrogen/deuterium exchange.** *Biophys J* 2011, **100(1)**:154-164.

173. Meier S, Grzesiek S, Blackledge M: **Mapping the conformational landscape of urea-denatured ubiquitin using residual dipolar couplings.** *J Am Chem Soc* 2007, **129(31)**:9799-9807.

174. Hu KN, Havlin RH, Yau WM, Tycko R: **Quantitative determination of site-specific conformational distributions in an unfolded protein by solid-state nuclear magnetic resonance.** *J Mol Biol* 2009, **392(4)**:1055-1073.

175. Wang Z, Plaxco KW, Makarov DE: **Influence of local and residual structures on the scaling behavior and dimensions of unfolded proteins.** *Biopolymers* 2007, **86(4)**:321-328.

176. Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR, Hasan MZ, Pande VS, Ruczinski I, Doniach S, Plaxco KW: **Random-coil behavior and the dimensions of chemically unfolded proteins.** *Proc Natl Acad Sci U S A* 2004, **101(34)**:12491-12496.

177. Bernadó P, Blackledge M: **A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering.** *Biophys J* 2009, **97(10)**:2839-2845.

178. Wang Y, Shortle D: **Residual helical and turn structure in the denatured state of staphylococcal nuclease: analysis of peptide fragments.** *Fold Des* 1997, **2(2)**:93-100.

179. Ding F, Jha RK, Dokholyan NV: **Scaling behavior and structure of denatured proteins.** *Structure* 2005, **13(7)**:1047-1054.

180. Fieber W, Kristjansdottir S, Poulsen FM: **Short-range, long-range and transition state interactions in the denatured state of ACBP from residual dipolar couplings.** *J Mol Biol* 2004, **339(5)**:1191-1199.

181. Shan B, Bhattacharya S, Eliezer D, Raleigh DP: **The low-pH unfolded state of the C-terminal domain of the ribosomal protein L9 contains significant secondary structure in the absence of denaturant but is no more compact than the low-pH urea unfolded state.** *Biochemistry* 2008, **47(36)**:9565-9573.

182. Baldwin R, Rose G: **Is protein folding hierarchic? II. Folding intermediates and transition states.** *Trends Biochem Sci* 1999, **24(2)**:77-83.

183. Smith L, Fiebig K, Schwalbe H, Dobson C: **The concept of a random coil. Residual structure in peptides and denatured proteins.** *Fold Des* 1996, **1(5)**:R95-106.

184. Yoo TY, Meisburger SP, Hinshaw J, Pollack L, Haran G, Sosnick TR, Plaxco K: **Small-angle X-ray scattering and single-molecule FRET spectroscopy produce highly divergent views of the low-denaturant unfolded state.** *J Mol Biol* 2012, **418(3-4)**:226-236.

185. Kräutler V, Hiller S, Hünenberger PH: **Residual structure in a peptide fragment of the outer membrane protein X under denaturing conditions: a molecular dynamics study.** *Eur Biophys J* 2010, **39(10)**:1421-1432.

186. Liu C, Yao M, Hogue CWV: **Near-membrane ensemble elongation in the proline-rich LRP6 intracellular domain may explain the mysterious initiation of the Wnt signaling pathway.** *BMC Bioinformatics* 2011, **12 Suppl 13**:S13.

187. Mohana-Borges R, Goto NK, Kroon GJA, Dyson HJ, Wright PE: **Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings.** *J Mol Biol* 2004, **340(5)**:1131-1142.

188. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based archival file for macromolecular structures.** *J Mol Biol* 1977, **112(3)**:535-542.

189. Garnier J, Gibrat JF, Robson B: **GOR method for predicting protein secondary structure from amino acid sequence.** *Methods Enzymol* 1996, **266**:540-553.

190. Garnier J, Osguthorpe DJ, Robson B: **Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.** *J Mol Biol* 1978, **120(1)**:97-120.

191. Pierre-Gilles Gennes: **Scaling Concepts in Polymer Physics.** *Cornell University Press, Ithaca, NY.* 1979,

192. Mor A, Haran G, Levy Y: **Characterization of the unfolded state of repeat proteins.** *HFSP journal* 2008, **2(6)**:405-415.

193. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22(12)**:2577-2637.

194. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G: **A series of PDB related databases for everyday needs.** *Nucleic Acids Res* 2011, **39(Database issue)**:D411-D419.

195. Saito R, Sato T, Ikai A, Tanaka N: **Structure of bovine carbonic anhydrase II at 1.95 A resolution.** *Acta Crystallogr D Biol Crystallogr* 2004, **60(Pt 4)**:792-795.

196. Cho HS, Lee SY, Yan D, Pan X, Parkinson JS, Kustu S, Wemmer DE, Pelton JG: **NMR structure of activated CheY.** *J Mol Biol* 2000, **297(3)**:543-551.

197. Thunnissen MM, Taddei N, Liguri G, Ramponi G, Nordlund P: **Crystal structure of common type acylphosphatase from bovine testis.** *Structure* 1997, **5(1)**:69-79.

198. Tsukihara T, Shimokata K, Katayama Y, Shimada H, Muramoto K, Aoyama H, Mochizuki M, Shinzawa-Itoh K, Yamashita E, Yao M, Ishimura Y, Yoshikawa S: **The low-spin heme of cytochrome c oxidase as the driving element of the proton-pumping process.** *Proc Natl Acad Sci U S A* 2003, **100(26)**:15304-15309.

199. Dunbar J, Yennawar HP, Banerjee S, Luo J, Farber GK: **The effect of denaturants on protein structure.** *Protein Sci* 1997, **6(8)**:1727-1733.

200. Warren MS, Brown KA, Farnum MF, Howell EE, Kraut J: **Investigation of the functional role of tryptophan-22 in Escherichia coli dihydrofolate reductase by site-directed mutagenesis.** *Biochemistry* 1991, **30(46)**:11092-11103.

201. Bezsonova I, Singer A, Choy WY, Tollinger M, Forman-Kay JD: **Structural comparison of the unstable drkN SH3 domain and a stable mutant.** *Biochemistry* 2005, **44(47)**:15550-15560.

202. Donahue JP, Patel H, Anderson WF, Hawiger J: **Three-dimensional structure of the platelet integrin recognition segment of the fibrinogen gamma chain obtained by carrier protein-driven crystallization.** *Proc Natl Acad Sci U S A* 1994, **91(25)**:12178-12182.

203. Michaux C, Pouyez J, Wouters J, Privé GG: **Protecting role of cosolvents in protein denaturation by SDS: a structural study.** *BMC Struct Biol* 2008, **8**:29.

204. Pastore A, Saudek V, Ramponi G, Williams RJ: **Three-dimensional structure of acylphosphatase. Refinement and structure analysis.** *J Mol Biol* 1992, **224(2)**:427-440.

205. Hubbard SR, Hendrickson WA, Lambright DG, Boxer SG: **X-ray crystal structure of a recombinant human myoglobin mutant at 2.8 A resolution.** *J Mol Biol* 1990, **213(2)**:215-218.

206. Li H, Dunn JJ, Luft BJ, Lawson CL: **Crystal structure of Lyme disease antigen outer surface protein A complexed with an Fab.** *Proc Natl Acad Sci U S A* 1997, **94(8)**:3584-3589.

207. Kuszewski J, Gronenborn AM, Clore GM: **Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration.** *J Am Chem Soc* 1999, **121(10)**:2337-2338.

208. Wikström M, Drakenberg T, Forsén S, Sjöbring U, Björck L: **Three-dimensional solution structure of an immunoglobulin light chain-binding domain of protein L. Comparison with the IgG-binding domains of protein G.** *Biochemistry* 1994, **33(47)**:14011-14017.

209. Williams RL, Greene SM, McPherson A: **The crystal structure of ribonuclease B at 2.5-A resolution.** *J Biol Chem* 1987, **262(33)**:16020-16031.

210. Chen J, Lu Z, Sakon J, Stites WE: **Increasing the thermostability of staphylococcal nuclease: implications for the origin of protein thermostability.** *J Mol Biol* 2000, **303(2)**:125-130.

211. Wang Q, Young P, Walters KJ: **Structure of S5a bound to monoubiquitin provides a model for polyubiquitin recognition.** *J Mol Biol* 2005, **348(3)**:727-739.

212. Alagaratnam S, van Pouderoyen G, Pijning T, Dijkstra BW, Cavazzini D, Rossi GL, Van Dongen WMAM, van Mierlo CPM, van Berkel WJH, Canters GW: **A crystallographic study of Cys69Ala flavodoxin II from Azotobacter vinelandii: structural determinants of redox potential.** *Protein Sci* 2005, **14(9)**:2284-2295.

213. Vogt J, Schulz GE: **The structure of the outer membrane protein OmpX from Escherichia coli reveals possible mechanisms of virulence.** *Structure* 1999, **7(10)**:1301-1309.

214. Xu Y, Yang W, Wu J, Shi Y: **Solution structure of the first HMG box domain in human upstream binding factor.** *Biochemistry* 2002, **41(17)**:5415-5420.

215. Le Guillou J, Zinn-Justin J: **Critical exponents for the n-vector model in three dimensions from field theory.** *Phys Rev Lett* 1977, **39(2)**:95-98.

216. Lim WK, Rosgen J, Englander SW: **Urea, but not guanidinium, destabilizes proteins by forming hydrogen bonds to the peptide group.** *Proceedings of the National Academy of Sciences* 2009, **106(8)**:2595-2600.

217. Rashid F, Sharma S, Bano B: **Comparison of guanidine hydrochloride (GdnHCl) and urea denaturation on inactivation and unfolding of human placental cystatin (HPC).** *Protein J* 2005, **24(5)**:283-292.

218. Wang S, Gu J, Larson SA, Whitten ST, Hilser VJ: **Denatured-state energy landscapes of a protein structural database reveal the energetic determinants of a framework model for folding.** *J Mol Biol* 2008, **381(5)**:1184-1201.

219. Sharpe T, Jonsson AL, Rutherford TJ, Daggett V, Fersht AR: **The role of the turn in beta-hairpin formation during WW domain folding.** *Protein Sci* 2007, **16(10)**:2233-2239.

220. Fuller AA, Du D, Liu F, Davoren JE, Bhabha G, Kroon G, Case DA, Dyson HJ, Powers ET, Wipf P, Gruebele M, Kelly JW: **Evaluating beta-turn mimics as beta-sheet folding nucleators.** *Proc Natl Acad Sci U S A* 2009, **106(27)**:11067-11072.

221. Petrovich M, Jonsson AL, Ferguson N, Daggett V, Fersht AR: **Phi-analysis at the experimental limits: mechanism of beta-hairpin formation.** *J Mol Biol* 2006, **360(4)**:865-881.

222. Martinez JC, Pisabarro MT, Serrano L: **Obligatory steps in protein folding and the conformational diversity of the transition state.** *Nat Struct Biol* 1998, **5(8)**:721-729.

223. Marcelino AMC, Gierasch LM: **Roles of beta-turns in protein folding: from peptide models to protein engineering.** *Biopolymers* 2008, **89(5)**:380-391.

224. Alexandrescu AT, Abeygunawardana C, Shortle D: **Structure and dynamics of a denatured 131-residue fragment of staphylococcal nuclease: a heteronuclear NMR study.** *Biochemistry* 1994, **33(5)**:1063-1072.

225. Bofill R, Searle MS: **Engineering stabilising beta-sheet interactions into a conformationally flexible region of the folding transition state of ubiquitin.** *J Mol Biol* 2005, **353(2)**:373-384.

226. Takano K, Yamagata Y, Yutani K: **Role of amino acid residues at turns in the conformational stability and folding of human lysozyme.** *Biochemistry* 2000, **39(29)**:8655-8665.

227. Nabuurs SM, Westphal AH, van Mierlo CPM: **Extensive formation of off-pathway species during folding of an alpha-beta parallel protein is due to docking of (non)native structure elements in unfolded molecules.** *J Am Chem Soc* 2008, **130(50)**:16914-16920.

228. Tafer H, Hiller S, Hilty C, Fernández C, Wüthrich K: **Nonrandom structure in the urea-unfolded Escherichia coli outer membrane protein X (OmpX).** *Biochemistry* 2004, **43(4)**:860-869.

229. Camilloni C, Guerini Rocco A, Eberini I, Gianazza E, Broglia R, Tiana G: **Urea and guanidinium chloride denature protein L in different ways in molecular dynamics simulations.** *Biophys J* 2008, **94(12)**:4654-4661.

230. Kuwata K, Shastry R, Cheng H, Hoshino M, Batt CA, Goto Y, Roder H: **Structural and kinetic characterization of early folding events in beta-lactoglobulin.** *Nat Struct Biol* 2001, **8(2)**:151-155.

231. Malkin LI, Rich A: **Partial resistance of nascent polypeptide chains to proteolytic digestion due to ribosomal shielding.** *J Mol Biol* 1967, **26(2)**:329-346.

232. Blobel G, Sabatini DD: **Controlled proteolysis of nascent polypeptides in rat liver cell fractions. I. Location of the polypeptides within ribosomes.** *J Cell Biol* 1970, **45(1)**:130-145.

233. Yonath A, Leonard KR, Wittmann HG: **A tunnel in the large ribosomal subunit revealed by three-dimensional image reconstruction.** *Science* 1987, **236(4803)**:813-816.

234. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA: **The structural basis of ribosome activity in peptide bond synthesis.** *Science* 2000, **289(5481)**:920-930.

235. Harms J, Schluenzen F, Zarivach R, Bashan A, Gat S, Agmon I, Bartels H, Franceschi F, Yonath A: **High resolution structure of the large ribosomal subunit from a mesophilic eubacterium.** *Cell* 2001, **107(5)**:679-688.

236. Beckmann R, Spahn CM, Eswar N, Helmers J, Penczek PA, Sali A, Frank J, Blobel G: **Architecture of the protein-conducting channel associated with the translating 80S ribosome.** *Cell* 2001, **107(3)**:361-372.

237. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA: **The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.** *Science* 2000, **289(5481)**:905-920.

238. Bernabeu C, Lake JA: **Nascent polypeptide chains emerge from the exit domain of the large ribosomal subunit: immune mapping of the nascent chain.** *Proc Natl Acad Sci U S A* 1982, **79(10)**:3111-3115.

239. Seidelt B, Innis CA, Wilson DN, Gartmann M, Armache JP, Villa E, Trabuco LG, Becker T, Mielke T, Schulten K, Steitz TA, Beckmann R: **Structural insight into nascent polypeptide chain-mediated translational stalling.** *Science* 2009, **326(5958)**:1412-1415.

240. Nakatogawa H, Ito K: **The ribosomal exit tunnel functions as a discriminating gate.** *Cell* 2002, **108(5)**:629-636.

241. Petrone PM, Snow CD, Lucent D, Pande VS: **Side-chain recognition and gating in the ribosome exit tunnel.** *Proc Natl Acad Sci U S A* 2008, **105(43)**:16549-16554.

242. Moore SD, Sauer RT: **Revisiting the mechanism of macrolide-antibiotic resistance mediated by ribosomal protein L22.** *Proc Natl Acad Sci U S A* 2008, **105(47)**:18261-18266.

243. Cabrita LD, Dobson CM, Christodoulou J: **Protein folding on the ribosome.** *Curr Opin Struct Biol* 2010, **20(1)**:33-45.

244. Elcock AH: **Molecular simulations of cotranslational protein folding: fragment stabilities, folding cooperativity, and trapping in the ribosome.** *PLoS Comput Biol* 2006, **2(7)**:e98.

245. Ziv G: **Ribosome exit tunnel can entropically stabilize -helices.** *Proceedings of the National Academy of Sciences* 2005, **102(52)**:18956-18961.

246. Lu J, Deutsch C: **Secondary structure formation of a transmembrane segment in Kv channels.** *Biochemistry* 2005, **44(23)**:8230-8243.

247. Woolhead CA, McCormick PJ, Johnson AE: **Nascent membrane and secretory proteins differ in FRET-detected folding far inside the ribosome and in their exposure to ribosomal proteins.** *Cell* 2004, **116(5)**:725-736.

248. Li C, Tang C, Liu M: **Protein dynamics elucidated by NMR technique.** *Protein & cell* 2013, **4(10)**:726-730.

249. Yu H: **Extending the size limit of protein nuclear magnetic resonance.** *Proc Natl Acad Sci U S A* 1999, **96(2)**:332-334.

250. Trylska J: **Simulating activity of the bacterial ribosome.** *Q Rev Biophys* 2009, **42(4)**:301-316.

251. Brandman R, Brandman Y, Pande VS, Zheng J: **A-site residues move independently from P-site residues in all-atom molecular dynamics simulations of the 70S bacterial ribosome.** *PloS one* 2012, **7(1)**:e29377.

252. Ishida H, Hayward S: **Path of nascent polypeptide in exit tunnel revealed by molecular dynamics simulation of ribosome.** *Biophys J* 2008, **95(12)**:5962-5973.

253. Gavezzotti A: **The calculation of molecular volumes and the use of volume analysis in the investigation of structured media and of solid-state organic reactivity.** *J Am Chem Soc* 1983, **105(16)**:5220-5220.

254. Pavan R, Ranghino G: **A method to compute the volume of a molecule.** *Comput Chem* 1982, **6(3)**:133-133.

255. Edelsbrunner H, Koehl P: **The geometry of biomolecular solvation.** *Combinatorial and computational geometry* 2005, **52**:243-243.

256. Tseng YY, Dupree C, Chen ZJ, Li WH: **SplitPocket: identification of protein functional surfaces and characterization of their spatial patterns.** *Nucleic Acids Res* 2009, **37(Web Server issue)**:W384-W389.

257. Medek P: **Computation of tunnels in protein molecules using Delaunay triangulation.** *Journal of WSCG* 2007, **15(1-3)**:107-114.

258. Bakowies D, Van Gunsteren WF: **Water in protein cavities: A procedure to identify internal water and exchange pathways and application to fatty acid-binding protein.** *Proteins* 2002, **47(4)**:534-545.

259. Lucent D, Vishal V, Pande VS: **Protein folding under confinement: a role for solvent.** *Proc Natl Acad Sci U S A* 2007, **104(25)**:10430-10434.

260. Case D, Darden T, Cheatham Iii T, Simmerling C, Wang J, Duke R, Luo R, Crowley M, Walker R, Zhang W, others: **AMBER 10.** *University of California, San Francisco* 2008, **32**

261. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P: **A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations.** *J Comput Chem* 2003, **24(16)**:1999-2012.

262. Trabuco LG, Villa E, Mitra K, Frank J, Schulten K: **Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics.** *Structure* 2008, **16(5)**:673-683.

263. Villa E, Sengupta J, Trabuco LG, LeBarron J, Baxter WT, Shaikh TR, Grassucci RA, Nissen P, Ehrenberg M, Schulten K, Frank J: **Ribosome-induced changes in**

elongation factor Tu conformation control GTP hydrolysis. *Proc Natl Acad Sci U S A* 2009, **106(4)**:1063-1068.

264. Tsui V, Case DA: **Theory and applications of the generalized Born solvation model in macromolecular simulations.** *Biopolymers* 2000, **56(4)**:275-291.

265. Anglada M: **An improved incremental algorithm for constructing restricted Delaunay triangulations.** *Computers \& Graphics* 1997, **21(2)**:215-215.

266. Edelsbrunner H, Shah N: **Incremental topological flipping works for regular triangulations.** *Algorithmica* 1996, **15(3)**:223-223.

267. Edelsbrunner H: **The union of balls and its dual shape.** *Discrete \& Computational Geometry* 1995, **13(1)**:415-415.

268. Mach P, Koehl P: **Geometric measures of large biomolecules: surface, volume, and pockets.** *J Comput Chem* 2011, **32(14)**:3023-3038.

269. Bryant R, Edelsbrunner H, Koehl P, Levitt M: **The area derivative of a space-filling diagram.** *Discrete \& Computational Geometry* 2004, **32(3)**:293-293.

270. Edelsbrunner H, Koehl P: **The weighted-volume derivative of a space-filling diagram.** *Proceedings of the National Academy of Sciences* 2003, **100(5)**:2203.

271. Humphrey W, Dalke A, Schulten K: **VMD: visual molecular dynamics.** *J Mol Graph* 1996, **14(1)**:33-8, 27.

272. Ohkawa H, Ostell J, Bryant S: **MMDB: an ASN.1 specification for macromolecular structure.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3**:259-267.

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

274. **The PyMOL Molecular Graphics System, Version 1.5.0.1 Schrödinger, LLC.**

275. **Origin (OriginLab, Northampton, MA).**

276. Wilson DN, Beckmann R: **The ribosomal tunnel as a functional environment for nascent polypeptide folding and translational stalling.** *Curr Opin Struct Biol* 2011, **21(2)**:274-282.

277. Wang K, Hu F, Xu K, Cheng H, Jiang M, Feng R, Li J, Wen T: **CASCADE_SCAN: mining signal transduction network from high-throughput data based on steepest descent method.** *BMC Bioinformatics* 2011, **12**:164.

278. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z: **Intrinsic disorder and protein function.** *Biochemistry* 2002, **41(21)**:6573-6582.

279. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN: **Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions.** *J Proteome Res* 2007, **6(5)**:1899-1916.

280. Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW: **Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits.** *Biochemistry* 2008, **47(29)**:7598-7609.

281. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z: **Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions.** *J Proteome Res* 2007, **6(5)**:1882-1898.

282. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK: **Intrinsic disorder in transcription factors.** *Biochemistry* 2006, **45(22)**:6873-6888.

283. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK: **Intrinsic disorder in cell-signaling and cancer-associated proteins.** *J Mol Biol* 2002, **323(3)**:573-584.

284. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT: **Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.** *J Mol Biol* 2004, **337(3)**:635-645.

285. Hsu WL, Oldfield CJ, Xue B, Meng J, Huang F, Romero P, Uversky VN, Dunker AK: **Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding.** *Protein Sci* 2013, **22(3)**:258-273.

286. Ayed A, Mulder FA, Yi GS, Lu Y, Kay LE, Arrowsmith CH: **Latent and active p53 are identical in conformation.** *Nat Struct Biol* 2001, **8(9)**:756-760.

287. Hoh JH: **Functional protein domains from the thermally driven motion of polypeptide chains: a proposal.** *Proteins* 1998, **32(2)**:223-228.

288. Tskhovrebova L, Trinick J: **Titin: properties and family relationships.** *Nat Rev Mol Cell Biol* 2003, **4(9)**:679-689.

289. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, Sali A, Rout MP: **The molecular architecture of the nuclear pore complex.** *Nature* 2007, **450(7170)**:695-701.

290. Mukrasch MD, Biernat J, von Bergen M, Griesinger C, Mandelkow E, Zweckstetter M: **Sites of tau important for aggregation populate {beta}-structure and bind to microtubules and polyanions.** *J Biol Chem* 2005, **280(26)**:24978-24986.

291. Marsh JA, Singh VK, Jia Z, Forman-Kay JD: **Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation.** *Protein Sci* 2006, **15(12)**:2795-2804.

292. Uversky VN, Eliezer D: **Biophysics of Parkinson's disease: structure and aggregation of alpha-synuclein.** *Curr Protein Pept Sci* 2009, **10(5)**:483-499.

293. Bienkiewicz EA, Adkins JN, Lumb KJ: **Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27(Kip1).** *Biochemistry* 2002, **41(3)**:752-759.

294. Baker JMR, Hudson RP, Kanelis V, Choy WY, Thibodeau PH, Thomas PJ, Forman-Kay JD: **CFTR regulatory region interacts with NBD1 predominantly via multiple transient helices.** *Nat Struct Mol Biol* 2007, **14(8)**:738-745.

295. Mittag T, Kay LE, Forman-Kay JD: **Protein dynamics and conformational disorder in molecular recognition.** *J Mol Recognit* 2010, **23(2)**:105-116.

296. Huang SY, Zou X: **Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking.** *Proteins* 2007, **66(2)**:399-421.

297. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK: **Sequence complexity of disordered protein.** *Proteins* 2001, **42(1)**:38-48.

298. Yang ZR, Thomson R, McNeil P, Esnouf RM: **RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins.** *Bioinformatics* 2005, **21(16)**:3369-3376.

299. Dosztányi Z, Csizmók V, Tompa P, Simon I: **The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.** *J Mol Biol* 2005, **347(4)**:827-839.

300. Huang A, Stultz CM: **Finding order within disorder: elucidating the structure of proteins associated with neurodegenerative disease.** *Future medicinal chemistry* 2009, **1(3)**:467-482.

301. Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, Uversky VN: **The unfoldomics decade: an update on intrinsically disordered proteins.** *BMC Genomics* 2008, **9 Suppl 2**:S1.

302. Choy WY, Forman-Kay JD: **Calculation of ensembles of structures representing the unfolded state of an SH3 domain.** *J Mol Biol* 2001, **308(5)**:1011-1032.

303. Gillespie JR, Shortle D: **Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures.** *J Mol Biol* 1997, **268(1)**:170-184.

304. Fisher CK, Stultz CM: **Constructing ensembles for intrinsically disordered proteins.** *Curr Opin Struct Biol* 2011, **21(3)**:426-431.

305. Ozenne V, Bauer F, Salmon L, Huang JR, Jensen MR, Segard S, Bernadó P, Charavay C, Blackledge M: **Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables.** *Bioinformatics* 2012, **28(11)**:1463-1470.

306. Krzeminski M, Marsh JA, Neale C, Choy WY, Forman-Kay JD: **Characterization of disordered proteins with ENSEMBLE.** *Bioinformatics* 2013, **29(3)**:398-399.

307. Mittag T, Marsh J, Grishaev A, Orlicky S, Lin H, Sicheri F, Tyers M, Forman-Kay JD: **Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase.** *Structure* 2010, **18(4)**:494-506.

308. Salmon L, Nodet G, Ozenne V, Yin G, Jensen MR, Zweckstetter M, Blackledge M: **NMR characterization of long-range order in intrinsically disordered proteins.** *J Am Chem Soc* 2010, **132(24)**:8407-8418.

309. Mathieson SI, Penkett CJ, Smith LJ: **Characterisation of side-chain conformational preferences in a biologically active but unfolded protein.** *Pac Symp Biocomput* 1999, :542-553.

310. Fawzi NL, Phillips AH, Ruscio JZ, Doucleff M, Wemmer DE, Head-Gordon T: **Structure and dynamics of the Abeta(21-30) peptide from the interplay of NMR experiments and molecular simulations.** *J Am Chem Soc* 2008, **130(19)**:6145-6158.

311. Wen EZ, Luo R: **Interplay of secondary structures and side-chain contacts in the denatured state of BBA1.** *J Chem Phys* 2004, **121(5)**:2412-2421.

312. Wong KB, Clarke J, Bond CJ, Neira JL, Freund SM, Fersht AR, Daggett V: **Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding.** *J Mol Biol* 2000, **296(5)**:1257-1282.

313. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M: **Simultaneous determination of protein structure and dynamics.** *Nature* 2005, **433(7022)**:128-132.

314. Fuentes G, Nederveen AJ, Kaptein R, Boelens R, Bonvin AMJJ: **Describing partially unfolded states of proteins from sparse NMR data.** *J Biomol NMR* 2005, **33(3)**:175-186.

315. Ulmschneider JP, Ulmschneider MB, Di Nola A: **Monte Carlo vs molecular dynamics for all-atom polypeptide folding simulations.** *J Phys Chem B* 2006, **110(33)**:16733-16742.

316. Drubin DG, Kirschner MW: **Tau protein function in living cells.** *J Cell Biol* 1986, **103(6 Pt 2)**:2739-2746.

317. Butner KA, Kirschner MW: **Tau protein binds to microtubules through a flexible array of distributed weak sites.** *J Cell Biol* 1991, **115(3)**:717-730.

318. Das RK, Pappu RV: **Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues.** *Proceedings of the National Academy of Sciences* 2013,

319. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data.** *Database : the journal of biological databases and curation* 2011, **2011**:bar009.

320. Yang C, van der Woerd MJ, Muthurajan UM, Hansen JC, Luger K: **Biophysical analysis and small-angle X-ray scattering-derived structures of MeCP2-nucleosome complexes.** *Nucleic Acids Res* 2011, **39(10)**:4122-4135.

321. Shell SS, Putnam CD, Kolodner RD: **The N terminus of Saccharomyces cerevisiae Msh6 is an unstructured tether to PCNA.** *Mol Cell* 2007, **26(4)**:565-578.

322. Bressan GC, Silva JC, Borges JC, Dos Passos DO, Ramos CHI, Torriani IL, Kobarg J: **Human regulatory protein Ki-1/57 has characteristics of an intrinsically unstructured protein.** *J Proteome Res* 2008, **7(10)**:4465-4474.

323. Nairn KM, Lyons RE, Mulder RJ, Mudie ST, Cookson DJ, Lesieur E, Kim M, Lau D, Scholes FH, Elvin CM: **A synthetic resilin is largely unstructured.** *Biophys J* 2008, **95(7)**:3358-3365.

324. Gazi AD, Bastaki M, Charova SN, Gkougkoulia EA, Kapellios EA, Panopoulos NJ, Kokkinidis M: **Evidence for a coiled-coil interaction mode of disordered proteins from bacterial type III secretion systems.** *J Biol Chem* 2008, **283(49)**:34062-34068.

325. Boze H, Marlin T, Durand D, Pérez J, Vernhet A, Canon F, Sarni-Manchado P, Cheynier V, Cabane B: **Proline-rich salivary proteins have extended conformations.** *Biophys J* 2010, **99(2)**:656-665.

326. Li J, Uversky VN, Fink AL: **Conformational behavior of human alpha-synuclein is modulated by familial Parkinson's disease point mutations A30P and A53T.** *Neurotoxicology* 2002, **23(4-5)**:553-567.

327. Longhi S, Receveur-Bréchot V, Karlin D, Johansson K, Darbon H, Bhella D, Yeo R, Finet S, Canard B: **The C-terminal domain of the measles virus**

nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J Biol Chem* 2003, **278(20)**:18638-18648.

328. Uversky VN, Li J, Souillac P, Millett IS, Doniach S, Jakes R, Goedert M, Fink AL: **Biophysical properties of the synucleins and their propensities to fibrillate: inhibition of alpha-synuclein assembly by beta- and gamma-synucleins.** *J Biol Chem* 2002, **277(14)**:11970-11978.

329. Nørholm AB, Hendus-Altenburger R, Bjerre G, Kjaergaard M, Pedersen SF, Kragelund BB: **The intracellular distal tail of the Na+/H+ exchanger NHE1 is intrinsically disordered: implications for NHE1 trafficking.** *Biochemistry* 2011, **50(17)**:3469-3480.

330. Kjaergaard M, Nørholm AB, Hendus-Altenburger R, Pedersen SF, Poulsen FM, Kragelund BB: **Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II?** *Protein Sci* 2010, **19(8)**:1555-1564.

331. Lens Z, Dewitte F, Monté D, Baert JL, Bompard C, Sénéchal M, Van Lint C, de Launoit Y, Villeret V, Verger A: **Solution structure of the N-terminal transactivation domain of ERM modified by SUMO-1.** *Biochem Biophys Res Commun* 2010, **399(1)**:104-110.

332. Paz A, Zeev-Ben-Mordehai T, Lundqvist M, Sherman E, Mylonas E, Weiner L, Haran G, Svergun DI, Mulder FAA, Sussman JL, Silman I: **Biophysical characterization of the unstructured cytoplasmic domain of the human neuronal adhesion protein neuroligin 3.** *Biophys J* 2008, **95(4)**:1928-1944.

333. Uversky VN, Gillespie JR, Millett IS, Khodyakova AV, Vasiliev AM, Chernovskaya TV, Vasilenko RN, Kozlovskaya GD, Dolgikh DA, Fink AL, Doniach S, Abramov VM: **Natively unfolded human prothymosin alpha adopts partially folded collapsed conformation at acidic pH.** *Biochemistry* 1999, **38(45)**:15009-15016.

334. Alborghetti MR, Furlan AS, Silva JC, Paes Leme AF, Torriani ICL, Kobarg J: **Human FEZ1 protein forms a disulfide bond mediated dimer: implications for cargo transport.** *J Proteome Res* 2010, **9(9)**:4595-4603.

335. Foucault M, Mayol K, Receveur-Bréchot V, Bussat MC, Klinguer-Hamour C, Verrier B, Beck A, Haser R, Gouet P, Guillon C: **UV and X-ray structural studies of a 101-residue long Tat protein from a HIV-1 primary isolate and of its mutated, detoxified, vaccine candidate.** *Proteins* 2010, **78(6)**:1441-1456.

336. Wells M, Tidow H, Rutherford TJ, Markwick P, Jensen MR, Mylonas E, Svergun DI, Blackledge M, Fersht AR: **Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain.** *Proc Natl Acad Sci U S A* 2008, **105(15)**:5762-5767.

337. Leyrat C, Jensen MR, Ribeiro EA, Gérard FCA, Ruigrok RWH, Blackledge M, Jamin M: **The N(0)-binding region of the vesicular stomatitis virus phosphoprotein is globally disordered but contains transient α-helices.** *Protein Sci* 2011, **20(3)**:542-556.

338. Moncoq K, Broutin I, Craescu CT, Vachette P, Ducruix A, Durand D: **SAXS study of the PIR domain from the Grb14 molecular adaptor: a natively unfolded protein with a transient structure primer?** *Biophys J* 2004, **87(6)**:4056-4064.

339. Marsh JA, Forman-Kay JD: **Ensemble modeling of protein disordered states: Experimental restraint contributions and validation.** *Proteins* 2011,

340. Monod J, Changeux JP, Jacob F: **Allosteric proteins and cellular control systems.** *J Mol Biol* 1963, **6(4)**:306-329.

341. Schreiber G, Fersht AR: **Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles.** *J Mol Biol* 1995, **248(2)**:478-486.

342. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5(4)**:823-826.

343. Holm L, Sander C: **Mapping the protein universe.** *Science* 1996, **273(5275)**:595-603.

344. Rost B: **Twilight zone of protein sequence alignments.** *Protein Eng* 1999, **12(2)**:85-94.

345. Ciaccio C, Coletta A, De Sanctis G, Marini S, Coletta M: **Cooperativity and allostery in haemoglobin function.** *IUBMB Life* 2008, **60(2)**:112-123.

346. Turner GJ, Galacteros F, Doyle ML, Hedlund B, Pettigrew DW, Turner BW, Smith FR, Moo-Penn W, Rucknagel DL, Ackers GK: **Mutagenic dissection of hemoglobin cooperativity: effects of amino acid alteration on subunit assembly of oxy and deoxy tetramers.** *Proteins* 1992, **14(3)**:333-350.

347. Pettigrew DW, Romeo PH, Tsapis A, Thillet J, Smith ML, Turner BW, Ackers GK: **Probing the energetics of proteins through structural perturbation: sites of regulatory energy in human hemoglobin.** *Proc Natl Acad Sci U S A* 1982, **79(6)**:1849-1853.

348. Freire E: **The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme.** *Proc Natl Acad Sci U S A* 1999, **96(18)**:10118-10122.

349. Holt JM, Ackers GK: **The pathway of allosteric control as revealed by hemoglobin intermediate states.** *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 1995, **9(2)**:210-218.

350. Wells JA: **Binding in the growth hormone receptor complex.** *Proc Natl Acad Sci U S A* 1996, **93(1)**:1-6.

351. Atwell S, Ultsch M, De Vos AM, Wells JA: **Structural plasticity in a remodeled protein-protein interface.** *Science* 1997, **278(5340)**:1125-1128.

352. Clackson T, Wells JA: **A hot spot of binding energy in a hormone-receptor interface.** *Science* 1995, **267(5196)**:383-386.

353. Hidalgo P, MacKinnon R: **Revealing the architecture of a K+ channel pore through mutant cycles with a peptide inhibitor.** *Science* 1995, **268(5208)**:307-310.

354. Serrano L, Horovitz A, Avron B, Bycroft M, Fersht AR: **Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles.** *Biochemistry* 1990, **29(40)**:9343-9352.

355. Horovitz A, Serrano L, Avron B, Bycroft M, Fersht AR: **Strength and co-operativity of contributions of surface salt bridges to protein stability.** *J Mol Biol* 1990, **216(4)**:1031-1044.

356. Chen Z, Meyer W, Rappert S, Sun J, Zeng AP: **Coevolutionary analysis enabled rational deregulation of allosteric enzyme inhibition in *Corynebacterium glutamicum* for lysine production.** *Appl Environ Microbiol* 2011, **77(13)**:4352-4360.

357. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW: **Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis.** *Mol Biol Evol* 2000, **17(1)**:164-178.

358. Göbel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins.** *Proteins* 1994, **18(4)**:309-317.

359. Altschuh D, Lesk AM, Bloomer AC, Klug A: **Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus.** *J Mol Biol* 1987, **193(4)**:693-707.

360. Fodor AA, Aldrich RW: **Influence of conservation on calculations of amino acid covariance in multiple sequence alignments.** *Proteins* 2004, **56(2)**:211-221.

361. Dunn SD, Wahl LM, Gloor GB: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.** *Bioinformatics* 2008, **24(3)**:333-340.

362. Yip KY, Patel P, Kim PM, Engelman DM, McDermott D, Gerstein M: **An integrated system for studying residue coevolution in proteins.** *Bioinformatics* 2008, **24(2)**:290-292.

363. Zhou F, Grigoryan G, Lustig SR, Keating AE, Ceder G, Morgan D: **Coarse-graining protein energetics in sequence variables.** *Phys Rev Lett* 2005, **95(14)**:148103.

364. Sanchez J, Ducastelle F, Gratias D: **Generalized cluster description of multicomponent systems.** *Physica A: Statistical Mechanics and its Applications* 1984, **128(1–2)**:334-350.

365. Grigoryan G, Zhou F, Lustig SR, Ceder G, Morgan D, Keating AE: **Ultra-fast evaluation of protein energies directly from sequence.** *PLoS Comput Biol* 2006, **2(6)**:e63.

366. Asta M, Ozolins V, Woodward C: **A first-principles approach to modeling alloy phase equilibria.** *JOM* 2001, **53(9)**:16-19.

367. Van der Ven A, Aydinol M, Ceder G: **First-Principles Evidence For Stage Ordering In LixCoO2.** *Journal Of The Electrochemical Society* 1998, **145(6)**:2149.

368. Acuner Ozbabacan SE, Gursoy A, Keskin O, Nussinov R: **Conformational ensembles, signal transduction and residue hot spots: application to drug discovery.** *Curr Opin Drug Discov Devel* 2010, **13(5)**:527-537.

369. Hu H, Columbus J, Zhang Y, Wu D, Lian L, Yang S, Goodwin J, Luczak C, Carter M, Chen L, James M, Davis R, Sudol M, Rodwell J, Herrero JJ: **A map of WW domain family interactions.** *Proteomics* 2004, **4(3)**:643-655.

370. Sudol M: **Structure and function of the WW domain.** *Prog Biophys Mol Biol* 1996, **65(1-2)**:113-132.

371. Penrose R: **A generalized inverse for matrices.** *Mathematical Proceedings of the Cambridge Philosophical Society* 1955, **51(03)**:406-406.

372. Chou PY, Fasman GD: **Empirical predictions of protein conformation.** *Annu Rev Biochem* 1978, **47**:251-276.

373. Chou PY, Fasman GD: **Prediction of the secondary structure of proteins from their amino acid sequence.** *Adv Enzymol Relat Areas Mol Biol* 1978, **47**:45-148.

374. Chou PY, Fasman GD: **Prediction of protein conformation.** *Biochemistry* 1974, **13(2)**:222-245.

375. Ranganathan R, Lu KP, Hunter T, Noel JP: **Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent.** *Cell* 1997, **89(6)**:875-886.

376. Ponting CP, Schultz J, Milpetz F, Bork P: **SMART: identification and annotation of domains from signalling and extracellular protein sequences.** *Nucleic Acids Res* 1999, **27(1)**:229-232.

377. Leaver-Fay A, Kuhlman B, Snoeyink J: **An adaptive dynamic programming algorithm for the side chain placement problem.** *Pac Symp Biocomput* 2005, :16-27.

378. Carter PJ, Winter G, Wilkinson AJ, Fersht AR: **The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus).** *Cell* 1984, **38(3)**:835-840.

379. Horovitz A: **Non-additivity in protein-protein interactions.** *J Mol Biol* 1987, **196(3)**:733-735.

380. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M: **CHARMM: the biomolecular simulation program.** *J Comput Chem* 2009, **30(10)**:1545-1614.

381. Desmet J, De Maeyer M, Hazes B, Lasters I: **The dead-end elimination theorem and its use in protein side-chain positioning.** *Nature* 1992, **356(6369)**:539-542.

382. Altman MD, Tidor B: **MultigridPBE — Software for computation and display of electrostatic potentials.** *MIT* 2003,

383. Green DF, Kangas E, Hendsch ZS, Tidor B: **ICE — Integrated Continuum Electrostatics.** *MIT* 2000,

384. Green DF, Tidor B: **Current Protocols in Bioinformatics. Petsko GE, editor. Chapter 8.3.** *New York: John Wiley & Sons, Inc* 2003,

385. Jäger M, Dendle M, Kelly JW: **Sequence determinants of thermodynamic stability in a WW domain--an all-beta-sheet protein.** *Protein Sci* 2009, **18(8)**:1806-1813.

386. Jäger M, Zhang Y, Bieschke J, Nguyen H, Dendle M, Bowman ME, Noel JP, Gruebele M, Kelly JW: **Structure-function-folding relationship in a WW domain.** *Proc Natl Acad Sci U S A* 2006, **103(28)**:10648-10653.

387. Zarrinpar A, Lim WA: **Converging on proline: the mechanism of WW domain peptide recognition.** *Nat Struct Biol* 2000, **7(8)**:611-613.

388. Kato Y, Hino Y, Nagata K, Tanokura M: **Solution structure and binding specificity of FBP11/HYPA WW domain as Group-II/III.** *Proteins* 2006, **63(1)**:227-234.