Proteins orchestrate a number of cellular processes either alone or in concert with other biomolecules like nucleic acids, carbohydrates, and lipids. They exhibit an intrinsic ability to fold *de novo* to their functional states. The three-dimensional structure of a protein, dependent on its amino acid sequence, is important for its function. Understanding this sequence-structure-function relationship has become one of the primary goals in biophysics. Various experimental techniques like X-ray crystallography, Nuclear Magnetic Resonance (NMR), and site-directed mutagenesis have been used extensively towards this goal. Computational studies include mainly sequence based, and structure based approaches. The sequence based approaches such as sequence alignments, phylogenetic analysis, domain identification, statistical coupling analysis etc., aim at deriving meaningful information from the primary sequence of the protein. The structure based approaches, on the other hand, use structures of folded proteins. Recent advances in structure determination and efforts by various structural consortia have resulted in an enormous amount of structures available for analysis. Innumerable observations such as the allowed and disallowed regions in the conformations of a peptide unit, hydrophobic core in globular proteins, existence of regular secondary structures like helices, sheets, and turns and a limited fold space have been landmarks in understanding the characteristics of protein structures.

The uniqueness of protein structure is attained through non-covalent interactions among the constituent amino acids. Analyses of protein structures show that different types of non-covalent interactions like hydrophobic interactions, hydrogen bonding, salt bridges, aromatic stacking, cation–π interactions, and solvent interactions hold protein structures together. Although such structure analyses have provided a wealth of information, they have largely been performed at a pair-wise level and an investigation involving such pair-wise interactions alone is not sufficient to capture all the determinants of protein structures, since they happen at a global level. This consideration has led to the development of graphs/networks for proteins.

Graphs or Networks are a collection of nodes connected by edges. Protein Structure Networks (PSNs) can be constructed using various definitions of nodes and edges. Nodes may vary from atoms to secondary structures in
proteins, and the edges can range from simple atom–atom distances to
distance between secondary structures. To study the interplay of amino acids
in structure formation, the most commonly used PSNs consider amino acids as
nodes. The criterion for edge definition, however, varies. PSNs can be
constructed at a course grain level by considering the distances between
Cα/Cβ atoms, any side–chain atoms, or the centroids of the amino acids. At a
finer level, PSNs can be constructed using atomic details by considering the
interaction types or by computing the extent of interaction between amino
acids.

Representation of proteins as networks and their analyses has given us a
unique perspective on various aspects such as protein structure organization,
stability, folding, function, oligomerization and so on. A variety of network
properties like the degree distribution, clustering coefficient, characteristic
path lengths, clusters, and hubs have been investigated. Most of these studies
are carried out on protein structures alone. However, the interaction of
proteins with other biopolymers like nucleic acids is vital for many crucial
biological processes like transcription and translation. In this thesis, we have
attempted to address this problem by constructing and analyzing combined
graphs of the structures of protein and DNA. Also, in almost all of the PSN
studies, the connections have been made solely on the basis of geometric
criteria. In the later part of the thesis, we have taken PSN a step further by
defining the non–covalent connections based on chemical considerations in the
form of the energies of interactions.

The thesis contains two sections. The first part mainly involves the
construction and application of PSNs to study DNA binding proteins. The
DNA binding proteins are involved in several high fidelity processes like DNA
recombination, DNA replication, and transcription. Although the protein–
DNA interfaces have been extensively analyzed using pair–wise interactions,
we gain additional global perspective from network approach. Furthermore,
most of the earlier investigations have been carried out from the protein point
of view (protein centric) and the present network approach aims to combine
both the protein centric and the DNA centric view points by construction and
analyses of protein–DNA graphs. These studies are described in Chapters 3
and 4. The second part of the thesis discusses the development,
characterization, and application of protein structure networks based on non–
covalent interaction energies. The investigations are presented in chapters 5
and 6.
Chapter 3 discusses the development of Protein–DNA Graphs (PDGs) where the protein–DNA interfaces are represented as networks. PDG is a bipartite network in which amino acids form a set of nodes and the nucleotides form the other set. The extent of interaction between the two diverse types of biopolymers is normalized to define the strength of interaction. Edges are then constructed based on the interaction strength between amino acids and nucleotides. Such a representation, reported here for the first time, provides a holistic view of the interacting surface.

The developed PDGs are further analyzed in terms of clusters of interacting residues and identification of highly connected residues, known as hubs, along the protein–DNA interface and discussed in terms of their interacting motifs. Important clusters have been identified in a set of protein–DNA complexes, where the amino acids interact with different chemical components of DNA such as phosphate, deoxyribose and base with varying degrees of connectivity. An analysis of such fragment based PDGs provided insights into the nature of protein–DNA interaction, which could not have been obtained by conventional pair–wise analysis. The predominance of deoxyribose–amino acid clusters in beta–sheet proteins, distinction of the interface clusters in helix–turn–helix and the zipper type proteins are some of the new findings from the analysis of PDGs. Additionally, a potential classification scheme has been proposed for protein–DNA complexes on the basis of their interface clusters. This classification scheme gives a general idea of how the proteins interact with different components of DNA in various complexes. The present graph–based method has provided a deeper insight into the analysis of the protein–DNA recognition mechanisms from both protein and DNA view points, thus throwing more light on the nature and specificity of these interactions (Sathyapriya, Vijayabaskar et al. 2008).

Chapter 4 delineates the application of PSN to an important problem in molecular biology. An analysis of interface clusters from multimeric proteins provides a clue to the important residues contributing to the stability of the oligomers. One such prediction was made on the DNA binding protein under starvation from Mycobacterium smegmatis (Ms–Dps) using PSNs. Two types of trimers, Trimer A (tA) and Trimer B (tB) can be derived from the dodecamer because of the inherent three fold symmetry of the spherical crystal structure. The irreversible dodecamerization of these native Ms–Dps trimers, in vitro, is known to be directly associated with the bimodal function (DNA binding and iron storage) of this protein. Interface clusters which were
identified from the PSNs of the derived trimers, allowed us to convincingly predict the residues E146 and F47 for mutation studies. The prediction was followed up by our experimental collaborators (Rakhi PC and Dipankar Chatterji), which led to the elucidation of the molecular mechanism behind the *in vitro* oligomerization of *Ms–Dps*. The F47E mutant was impaired in dodecamerization, and the double mutant (E146AF47E) was a native monomer in solution. These two observations suggested that the two trimers are important for dodecamerization and that the residues selected are important for the structural stability of the protein *in vitro*. From the structural and functional characterizations of the mutants, we have proposed an oligomerization pathway of *Ms–Dps* (Chowdhury, Vijayabaskar *et al*. 2008).

The second part of the thesis involves the development, characterization (Chapter 5) and application (Chapter 6) of Protein Energy Networks (PENs). As mentioned above, the PSNs constructed on the geometric basis efficiently capture the topology and associated properties at the level of atom–atom contact. The chemistry, however, is not completely captured by these network representations, and a wealth of information can be extracted by incorporating the details of chemical interactions. This study is an advancement over the existing PSNs, in terms of edges being defined on the basis of interaction energies among the amino acids. This interaction energy is the resultant of various types of interactions within a protein. Use of such realistic interaction energies in a weighted network captures all the essential features responsible for maintaining the protein structure.

The methodology involved in representing proteins as interaction energy weighted networks, with realistic edge weights obtained from standard force fields is described in Chapter 5. The interaction energies were derived from equilibrium ensembles (obtained using molecular dynamics simulations) to account for the structural plasticity, which is essential for function elucidation. The suitability of this method to study single static structures was validated by obtaining interaction energies on minimized crystal structures of proteins. The PENs were then characterized using network parameters like edge weight distributions, clusters, hubs, and shortest paths. The PENs exhibited three distinct behaviors in terms of the size of the largest connected cluster as a function of interaction energy; namely, the pre–transition, transition, and post transition regions, irrespective of the topology of the proteins. The pre–transition region (energies<–20 kJ/mol) comprises smaller clusters with mainly charged and polar residues as hubs. Crucial topological changes take
place in the transition region (∼10 to −20 kJ/mol), where the smaller clusters aggregate, through low energy van der Waals interactions, to form a single large cluster in the post–transition region (energies > −10 kJ/mol). These behaviors reinforce the concept that hydrophobic interactions hold together local clusters of highly interacting residues, keeping the protein topology intact (Vijayabaskar and Vishveshwara 2010).

The applications of PENs in studying protein organization, allosteric communication, thermophilic stability and the structural relation of remote homologues of TIM barrel families have been outlined in Chapter 6.

In the first case, the weighted networks were used to identify stabilization regions in protein structures and hierarchical organization in the folded proteins, which may provide some insights into the general mechanism of protein folding and stabilization (Vijayabaskar and Vishveshwara 2010). In the second case the features of communication paths in proteins were elucidated from PENs, and specific paths have been extensively discussed in the case of PDZ domain, which is known to bring together protein partners, mediating various cellular processes. Changes in PEN upon ligand binding, resulting in alterations of the shortest paths (energetically most favorable paths) for a small fraction of residues, indicated that allosteric communication is anisotropic in PDZ. The observations also establish that the shortest paths between functionally important sites traverse through key residues in PDZ2 domain. Furthermore, shortest paths in PENs provide us the exact pathways of communication between residues. Although the communication in PDZ has been extensively investigated, detailed information of pathways at the energy level has emerged for the first time from the present study from PEN analysis (Vijayabaskar and Vishveshwara 2010). In the third case, a set of thermophilic and mesophilic proteins were compared to determine the factors responsible for their thermal stability from a network perspective using PENs. The sub–graph parameters such as cluster population, hubs and cliques were the prominent contributing factors for thermal stability. Also, the thermophilic proteins have a better–packed hydrophobic core. The property of thermophilic protein to increase stability by increasing the connectivity but retain conformational flexibility is discussed from a cliques and communities (higher order inter–connection of residues) perspective (Vijayabaskar and Vishveshwara 2010). Finally, the remote homologues from the TIM barrel fold have been analyzed using PENs to identify the interactions responsible for the maintenance of the fold despite low sequence similarity. A study of conserved
interactions in family specific PENs reveals that the formation of the central beta barrel is vital for the TIM barrel formation. The beta barrel is being formed by either conserved long range electrostatic interactions or by tandem arrangement of low energy hydrophobic interactions. The contributions of helix–sheet and helix–helix interactions are not conserved in the families. This study suggests that the sequentially near residues forming the helix–sheet interactions are common in many folds and hence formed despite non–conservation, whereas formation of beta barrel requires long range interactions, thus more conserved within the families.

The thesis also consists of an appendix in which a web–tool, developed to express proteins as networks and analyze these networks using different network parameters is discussed. The web based program–GraProStr allows us to represent proteins as structure graphs/networks by considering the amino acid residues as nodes and representing non–covalent interactions among them as edges. The different networks (classified based on edge definition) which can be obtained using GraProStr are Protein Side–chain Networks (PScNs), Cα/Cβ distance based networks (PcNs) and Protein–Ligand Networks (PLNs). The parameters which can be generated include clusters, hubs, cliques (rigid regions in proteins) and communities (group of cliques). It is also possible to differentiate the above mentioned parameters for monomers and interfaces in multimeric proteins. The well tested tool is now made available to the scientific community for the first time. GraProStr is available online and can be accessed from http://vishgraph.mbu.iisc.ernet.in/GraProStr/index.html. With a variety of structure networks, and a set of easily interpretable network parameters GraProStr can be useful is analyzing protein structures from a global paradigm (Vijayabaskar, Vidya et al. 2010).

In summary, we have extensively studied DNA binding proteins using side–chain based protein structure networks and by integrating the DNA molecule into the network. Also, we have upgraded the existing methodology of generating structure networks, by representing both the geometry and the chemistry of residues as interaction energies among them. Using this energy based network we have studied diverse problems like protein structure formation, stabilization, and allosteric communication in detail. The above mentioned methodologies are a considerable advancement over existing structure network representations and have been shown in this thesis to shed more light on the structural features of proteins.