Protein Contact Networks: A Network Description of Proteins

Nishant Singh
MS08037

A dissertation submitted for the partial fulfilment of BS-MS dual degree in Science

Indian Institute of Science Education and Research Mohali

June 18, 2013

Protein Contact Networks:
A Network Description of Proteins

Nishant Singh
MS08037

A dissertation submitted for the partial fulfillment of
BS-MS dual degree in Science

Indian Institute of Science Education and Research Mohali

June 18, 2013

Certificate of Examination

This is to certify that the dissertation titled Protein Contact Networks: A Network Description of Proteins submitted by Mr. Nishant Singh (Reg. No. MS08037) for the partial fulfillment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Kuljeet Sandhu Dr. Shashi Bhushan Pandit Prof. Somdatta Sinha
(Supervisor)

Dated: June 18^{th}, 2013

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Prof. Somdatta Sinha at the Indian Institute of Science Education and Research Mohali. This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Nishant Singh
MS08037
June 18^{th}, 2013

In my capacity as the supervisor of the candidates project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Prof. Somdatta Sinha
(Supervisor)
June 18^{th}, 2013

Acknowledgment

I am deeply grateful to Prof. Somdatta Sinha, my supervisor for her guidance and support, giving me the freedom to think and execute experiments. Her patient guidance, enthusiasm, support, critical analysis and constructive suggestions all put together is an asset that I will always cherish in my memory. Her attitude and assistance motivated me towards research.

I would also like to thank Prof. Somdatta Sinha and Department of Mathematics, IISc Bangalore for providing an opportunity to attend International School and Conference on Networks in Biology, Social Science and Engineering at Indian Institute of Science, Bangalore, July 2012 and also to present the poster (my work) at IISc International Conference on Biomolecular Forms and Functions, January 2013.

I would like to convey my deep sense of gratitude and very special thanks to Mr. Ashutosh who guided me throughout my project, and showed his patience during initial phases of learning. I am also thankful to Ms Priya and Mr. Vinay Gade for being such nice lab mates who were always ready to provide help.

I would also like to acknowledge Dr. Shashi Bhushan Pandit and Dr. Kuljeet Sandhu for their valuable suggestions about the thesis.

And finally I would like to thank the IISER community at large especially the IISER, Library and computation facility. Thank you.

Nishant Singh

Contents

List of Figures

Abstract

Proteins are biological macromolecules made up of linear chains of amino acids, and are organized into three-dimensional structures comprising of different secondary structural elements. In its functional form a protein acts like a complex network where the nodes are the constituent amino acids, and the links are the chemical interactions that hold them together with short and long-range contacts. Thus protein three-dimensional structures can be modelled using Graph Theory as complex networks of interacting amino acids. These are termed Protein Contact Networks (PCN). Since many topological properties of networks can be understood from the network parameters, we believe that it can also be a useful approach to identify the different structural classes of proteins and their influence in protein function. In this study, we have attempted to understand how network properties and attributes can be used to study - (a) the major structural classes of proteins, and (b) the relationship between structure and function in proteins, which do not show significant conformational changes in ligand-binding.

As per the Structural Classification of Proteins (SCOP), proteins are grouped into four classes, i.e. α, β, $\alpha + \beta$, and α/β based on their major secondary structural contents, which have different topologies. PCNs were developed for each class (50 proteins in each class) and different visual methods used to understand the differences among them. Several network parameters were calculated at local and global level, and their distribution studied. Average clustering coefficients showed statistically significant differences among the classes, except between $\alpha + \beta$, and α/β. The average shortest path did not show any difference among any class. The degree distribution and the number of residues having the most common degree show variation among the structural classes. Additionally, all PCNs of proteins of all classes showed **small world** nature.

In our attempt to study structure-function relationship using the PCN approach, we used the HIV-1 Reverse Transcriptase protein (apo) and its ligand-bound form (holo) as the model systems. We used three structures - 1RTJ (HIV-1 RT unbound), 1IKW (HIV-1 RT bound to EFZ, which is a non-nucleoside reverse transcriptase inhibitor), and 1FKO (HIV-1 RT with resistance mutation at K103N). We calculated the root mean square deviation (RMSD) among the three structures, which was found to differ very little. This was corroborated by insignificant variations in the global clustering coefficient and average shortest paths of the three PCN. We then used the PCNs to study the loss and gain of contacts among the three networks with different functionality. We analysed the contacts in the ligand-binding pocket and interface between the two chains, and identified few important contacts that allow the change in function in spite of the three dimensional structure being quite similar.

Thus, the work presented in this thesis argues that the complex network approach to study protein three-dimensional structure can not only be an important and useful methodology to study structural attributes of a protein, but can also unravel local changes in contacts for understanding protein structure-function relationship.

Chapter 1: Introduction

Proteins

Proteins are important biochemical molecules essential for life. They are natural polymer molecules consisting of amino acid as a monomeric subunits.
Each amino acid has at least one amine and one acid functional group. The different properties result from variations in the structures of different R groups. The R group is often referred to as the amino acid **side chain**. The carbon atom next to the carboxyl group is called the α carbon.

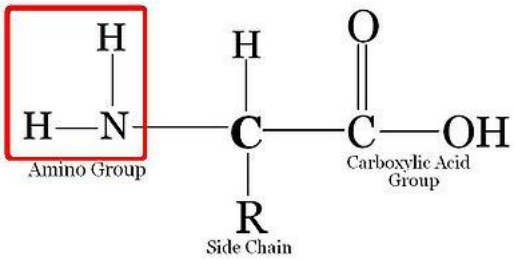


Figure 1: Structure of amino acid

Amino acids which have been incorporated into a peptide are termed as residues. Every peptide has a N-terminus and C-terminus residue on the ends. The rigid, planar structure, is a consequence of resonance interactions that give the peptide bond a 40 percent double-bond character. Due to this, the C-N bond in peptides is 0.13 $\mathring{A}$ shorter than its N-Cα single bond. C=O bond is 0.02 $\mathring{A}$ longer than that of aldehydes and ketones.

Structural Classification Of Proteins (SCOP)

SCOP database is a large manual classification of protein structural domains based on similarities of their structures and amino acid sequences. The source of protein structures is the Protein Data Bank. SCOP was created in 1994. It is maintained by Alexei G. Murzin and his colleagues at the Laboratory of Molecular Biology in Cambridge, England.
The classes are the top level, or "root" of the SCOP hierarchical classification (1).

Classes:

a. All alpha proteins: domains consisting of α-helices
b. All beta proteins: domains consisting of β-sheets
c. Alpha and beta proteins (α/β): Mainly parallel beta sheets (beta-alpha-beta units)
d. Alpha and beta proteins ($\alpha+\beta$): Mainly antiparallel beta sheets (segregated alpha and beta regions).

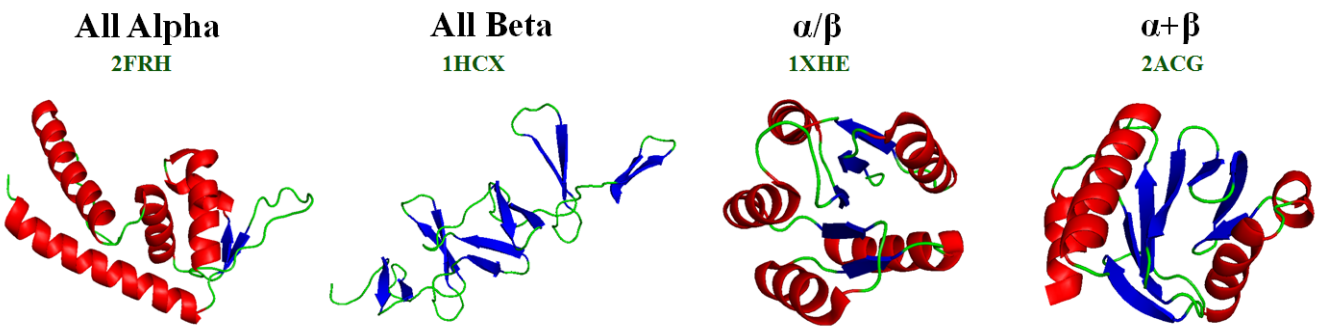


Figure 2: Different Structural Classes of Protein

Networks

A network is a graph consisting of points connected by the lines. The points are called nodes and connecting lines are called edges. The edges in the network denote the flow of information between the nodes.

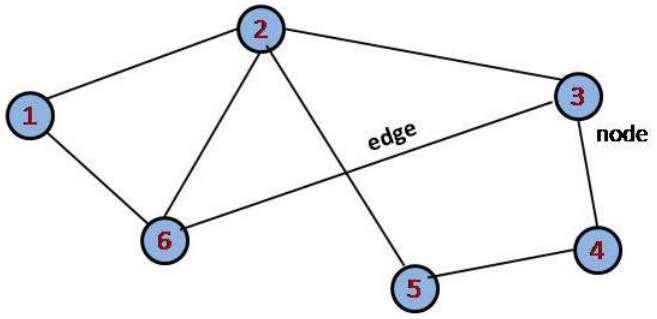


Figure 3: An undirected network consisting of six nodes.

The networks can be used to represent different kinds of system. For example, the World Wide Web is a large network where pages are the nodes and the edges are the hyperlinks between them. In a social network the nodes are the people and edges are the social relation (friends, acquaintances etc) between them e.g. Facebook.

A network is directed when the edges connecting the nodes have a particular direction. The direction of edges indicates the direction of flow of information. The example of such networks include metabolic network, World Wide Web etc. In an undirected network the edges dont have any direction and thus no particular direction of flow of information.

In an economic network nodes are companies and edges are the financial interactions between them. The nodes in such a network are different and also the edges (financial interac-

tions) are not equal. The information exchanged (financial exchange like sell and purchase of products) between two nodes is not uniform. The graphical representation is thus called a weighted network.

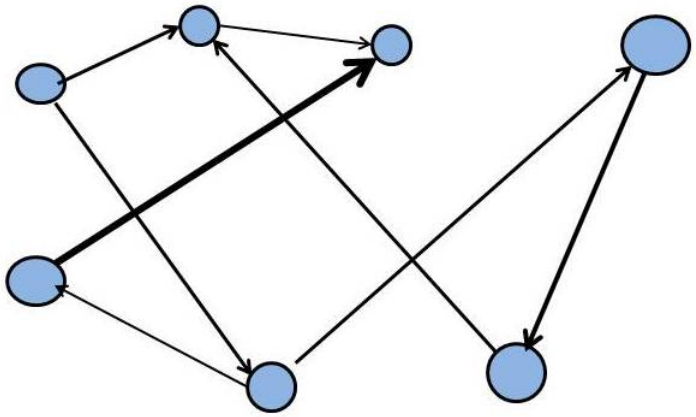


Figure 4: A weighted directed network: The thickness of edges and nodes denotes their weight.

The networks are also represented by their adjacency matrix. For an undirected and unweighted network of N nodes, an N X N adjacency matrix A can be formed. The elements of the adjacency matrix are either 0 or 1. If there is a link between two nodes then the value is 1 or else 0. The a_{ij} is the element of matrix which corresponds to the interaction between ith and jth node of the network. Thus

$$\mathbf{a}_{ij} = 0 \text{ (when there is no link between node i and j and when i=j);}$$
$$\mathbf{a}_{ij} = 1 \text{ (when there is a link between node i and j)}$$

To remove self looping, we take a_{ij} to be 0 when i=j.
The convention to write adjacency matrix for directed networks is different.
The adjacency matrix for the network in figure 1 can be written as:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Types Of Networks

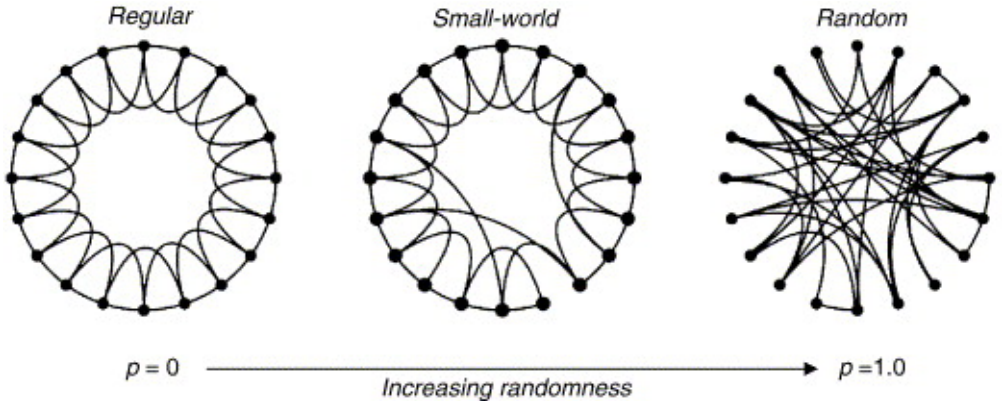


Figure 5: Different types of network topology (31).

Regular Networks

They are highly ordered as each node has exactly the same number of links.

Small-World Networks

A small-world network is a network graph in which most nodes are not neighbors of one another, but most nodes can be reached from every other node by a small number of links. A graph is considered small-world if its average clustering coefficient is significantly higher than a random graph constructed on the same vertex set, and if the graph has approximately the same mean shortest path length as its corresponding random graph (2).

Random Networks

In a random graph, unlike the regular graph architecture, the topological rule is that the node degrees may not all be equal. Instead, the degrees are distributed according to a Poisson distribution because it is assumed that any linking between nodes can happen with equal probability (2).

Scale-Free Networks

It is a highly connected network model and was studied by A L Barabasi. Compared to a random network, the scale-free network has a very different kind of connectivity because the degree distribution is defined by a power law distribution instead of the Poisson distribution associated with the random network (3,4).

Biological Networks

Any network which applies to biological system is called a biological network. Examples of such networks are - food web, neuronal networks, metabolic networks, protein-protein

interaction networks, signaling networks, gene regulatory networks, etc. These biological networks provide mathematical analysis of connections found in ecological, evolutionary and physiological studies.

Bioinformatics has shifted the focus from individual genes, protein and search algorithms to large-scale networks often denoted as -omes such as biome, interactome and proteome.

Protein Contact Network (PCN)

A protein contact network is undirected and unweighted. The structure of a protein is stabilized by a large number of interactions such as covalent bond , van der-waal forces, H-bonds, etc, the edges basically corresponds to these interactions. The nodes are the residues and edges denote the interaction between them.

The Cα atom of amino acid residues is taken as the center of mass for the residues. The distances are measured between all Cα atoms. If the distance is less than $7\mathring{A}$, the residues are taken to be interacting, else not interacting. Thus $7\mathring{A}$ makes the cutoff for building the contact network. The threshold distance can be varied from a high, fine-grained resolution to a low, coarse-grained resolution. Here the choice of threshold distance was done based on the inter-residue chemical interactions. If we increase the cutoff distance the number of contacts will increase and if we decrease the cutoff distance, the number of contacts will go down. It has been studied that beyond $7\mathring{A}$, the accuracy for the prediction of contacts decreases (5).

The pair of Cα atoms which have distance below $4\mathring{A}$ are the nearest-neighbor in the peptide chain. They have much stronger interactions than the other pairs, which can be described rather well by a $1/r^6$ dependence (6).

Other types of networks in biology

Gene Regulatory Networks

These networks analyse protein-DNA interactions. The gene expression is regulated by transcription factors. The transcription factors are proteins that bind to DNA, most of them bind at multiple sites in a genome. Cells have complex gene regulatory networks because of multiple binding sites. For example, the human genome encodes on the order of 1,400 DNA-binding transcription factors that regulate the expression of more than 20,000 human genes. Technologies to study gene regulatory networks include Chip-chip, Chip-seq, Clip-seq, and others (7).

Metabolic Networks

All the metabolic activities are result of biochemical reactions which convert one compound into another. These reactions involve enzymes as catalysts. All the compounds

(biomolecules) are a part of biochemical network of reactions, called metabolic network (8, 9).

Neuronal Networks

The network comprises of neurons that are connected and functionally associated in nervous system. These networks are often visualized as a group of neurons that perform specific physiological function in laboratory analysis.

Signaling Networks

Signaling are complex system of communication that controls basic cellular activities and co-ordinates all cell functions. Signaling networks typically integrate protein-protein interaction networks, gene regulatory networks, and metabolic networks (10).

Food Web

This network depicts feeding connection in an ecosystem. Nodes are the organisms and links map the feeding connections (who eats whom).

Quantative Description of Networks

Degree

The degree of a node is its most basic structural property, it is the number of the links of a node.

Shortest Path

Number of links that must pass by the shortest route, from one to node to another. The characteristic path lengths is the average of the shortest path lengths and is a measure of the network's overall navigability, it is calculated by

$$\boldsymbol{L} = 1/\boldsymbol{N}(\boldsymbol{N}-1)\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}-1}\sum_{\boldsymbol{j}=\boldsymbol{i}+1}^{\boldsymbol{N}}\boldsymbol{L}_{\boldsymbol{ij}}$$

where N is number of nodes in network and $L_{i,j}$ is the shortest path length between i and j.

Diameter

It is the longest among the shortest paths.

$$\mathbf{D = \max L_{ij}\ ,\ \forall\ ij\ pairs\ of\ shortest\ paths.}$$

Closeness Centrality

Closeness centrality (Ci) measure the proximity of a node i to all others within the network. The closeness centrality value C_k for residue k, is defined as

$$C_k = n-1 / \sum_{i \neq k} d(i,k)$$

where d(i,k) is the shortest path between node i and k, and n is the number of nodes in the network.
Thus, the more central a node is the lower its total distance to all other node. Closeness is a measure of how fast a information will spread from that particular node to all other nodes sequentially (11,12).

Betweenness Centrality

It is a measure of a node's centrality in a network, equal to the number of shortest path from all vertices to all other vertices that pass through that node.
Betweenness centrality is a useful measure of the load placed on the given node in the network as well as the node's importance to the network then just connectivity.
The betweenness centrality of a node v is given by the expression:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node **s** to node **t** and $\sigma_{st}(v)$ is the number of those paths that pass through v.

Clustering Coefficient

It is defined as the ratio of the no. of links present within the neighbors of a node to the maximum possible no. of links exist within its neighbors.
There are two subclasses of clustering coefficient: the global and the local. The global clustering coefficient gives an over all indication of the clustering in the network whereas the local clustering coefficient gives indication of the embeddedness of single node. There is a network average clustering coefficient, which is given as the average of the local cluster coefficient of all the vertices n.

$$c = (1/N) \sum_{i=1}^{N} c_i$$

C_i represents the clustering coefficient for a node n_i that ranges from 0 to 1.

Introduction to HIV-1

Human immunodeficiency virus (HIV) is a retrovirus belonging to genus lentivirus. This retrovirus is the cause behind acquired immunodeficiency syndrome (AIDS) in humans which causes failure of immune system. The two species of HIV have been characterized till now - HIV-1 and HIV-2 (13). HIV-1 is more prevalent in world and highly virulent, while HIV-2 is mostly observed in west Africa and its virulence and infectivity is lower than HIV-1. The virus infects a variety of immune cells T-cells, microglial cells and macrophages.

HIV has very high genetic variability. The high genetic variability can be attributed to its fast replication (about 1010 virions per day), high mutation rate and recombinogenic property of Reverse Transcriptase.

Viral Replication of HIV-1

The replication cycle of HIV has four main landamraks entry into host cell, reverse transcription, integration of viral DNA into host genome, expression of viral proteins, viral assembly and release.

Entry into host cell: The envelope surface glycoproteins (GP120, GP41 etc) of virus interacts with the receptors on host cell (CD4+ T cells). This is then followed by fusion of viral enevelope with cell membrane of host.

Reverse transcription: HIV has Reverse Transcriptase (RT), an enzyme which copies viral RNA and makes double stranded DNA.

Integration into host genome: The enzyme Integrase (IN) integrates the double stranded DNA formed by RT, into the host cell chromosome. This stage is called provirus. Now the viral genome is replicated along with the host cell genome when the cell divides. The integration of viral genome with host cell genome provides latency and also helps virus escape immune response.

Expression of viral proteins: Production of viral proteins and RNA takes place when the provirus is transcribed. The proteins are expressed using host cell machinery and viral regulatory protein (*tat* and *rev*).

Virus Assembly: Viral proteins are then assembled using the host cell's machinery. The virus's Protease enzyme allows the processing of newly translated polypeptides into the proteins, which are then assembled into viral particles. The virus eventually buds out of the cell.

A cell infected with a retrovirus does not necessarily lyse the cell when viral replication takes place but rather many viral particles can bud out of a cell over the course of time.

Structure of HIV-1 Reverse Transcriptase

The enzyme is a heterodimer composed of two related subunits- p66 and p51.The two subunits, p66 and p51 are 560 and 440 amino acids long, respectively. They are formed by cleavage of Gag-Pol polyprotien (it is syntheised from unspliced viral RNA) by viral Protease (PR) (14, 15). The p66 domain of heterodimer is composed of two spatially distinct active sites for RNA polymerase and RNase H activity. The p51 domain plays structural role. The polymerase domain of p66 is composed of four subdomains: fingers (residue 1-85 and 118-155), palm (residue 86-117 and 156-236), thumb (residue 237-318) and connection (residue 319-426). p51 folds into the same four subdomains as the polymerase domain of p66 (fingers, palm, thumb, and connection) but the positions of the subdomains relative to each other are different in p66 and p51 (16,17).

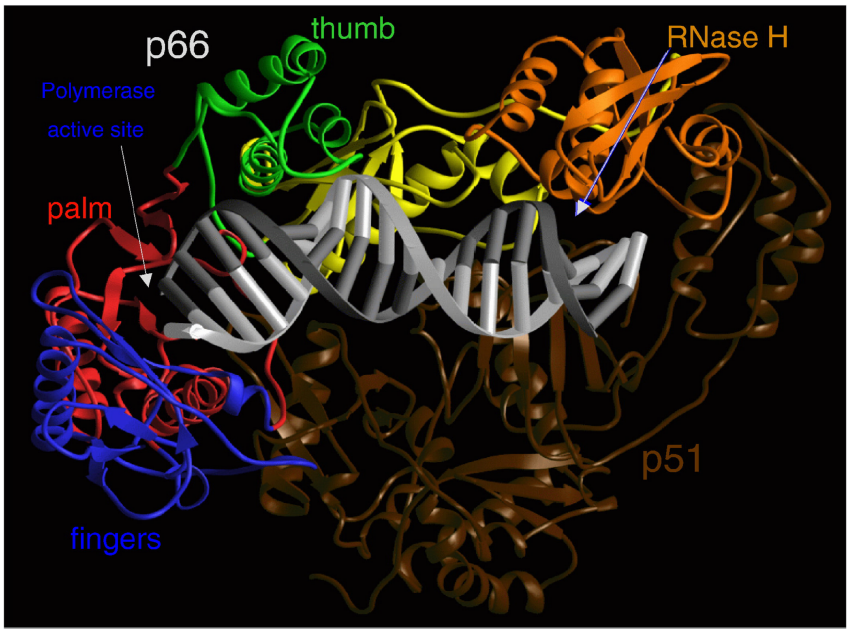


Figure 6: Ribbon representation of HIV-1 RT in a complex with nucleic acid (32).

The fingers, palm, thumb, connection, and RNase H subdomains of the p66 subunit are shown in blue, red, green, yellow, and orange, respectively. The p51 subunit is shown in dark brown. The template and primer DNA strands are shown in light and dark gray, respectively.

The cleft that binds the nucleic acid is formed by p66 fingers, palm, thumb, connection and RNase H subdomain of p66. The connection and thumb of p51 forms the floor of binding cleft. The binding cleft is formed in a way that nucleic acid contacts both polymerase and RNase H active sites. The αH and αI helices of the p66 thumb help to properly position the nucleic acid.

The DNA primer grip is a highly conserved structural motif (18) that consists of the p66 β12-β13 hairpin in HIV-1 RT (17). It helps position the 3-OH end of the primer strand at

the polymerase active site.

The polymerase active site is composed of three catalytic carboxylates in the palm subdomain of p66 (D110, D185, and D186) that bind two divalent ions (Mg2+ in *vivo*; Mn2+ in *vitro*) that are required for catalysis (19). D185 and D186 are part of the YXDD motif which is highly conserved in retroviral RTs (X is Met in HIV RT) (20). Other conserved residues that help form the dNTP binding site of RT include - R72 and K65 that are involved in the binding the β- and γ- phosphates respectively in the incoming dNTP (21); residue Y115 that contributes to the binding of the deoxyribose ring of the incoming dNTP and has been termed to be a steric gate that discriminates between deoxy and ribonucleoside triphosphates (22); and Q151, a residue that interacts directly with the 3-OH of the incoming dNTP (21).

Molecular mechanisms of HIV-1 reverse transcriptase inhibition

There are two kinds of RT inhibitors which block polymerase activity - nucleoside analogs (NRTI) and nonnucleoside analogs (NNRTI). The NRTI lacks 3-OH and when incorporated into viral DNA by RT, terminates the chain and hence stop the polymerization.

Most standard three-drug regimens involve two NRTIs combined with either a PR inhibitor or an NNRTI. Four NNRTIs have been approved for the treatment of AIDS (the first generation NNRTIs are nevirapine and delavirdine, the second generation NNRTI is efavirenz, and the third generation is etravirine).

Non-Nucleoside Revere Transcriptase Inhibitors (NNRTIs)

NNRTI-binding pocket (NNIBP) is a hydrophobic pocket adjacent to the polymerase active site in RT. The NNIBP consists of residues L100, K101, K103, V106, T107, V108, V179, Y181, Y188, V189, G190, F227, W229, L234, and Y318 of p66 and E138 of p51.

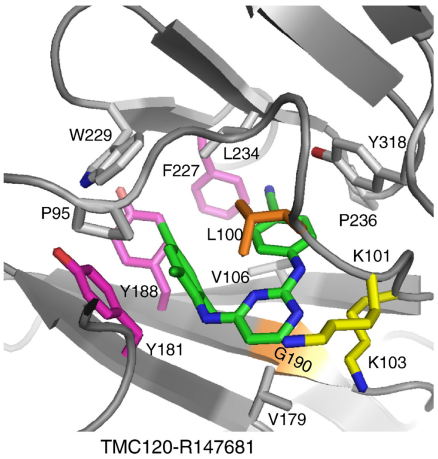


Figure 7: Ribbon representation of the NNRTI-binding pocket, showing the residues where NNRTI-resistance mutations occur (32).

NNRTIs are non-competitive inhibitors and do not directly interfere with the binding of either the dNTP or the nucleic acid substrates of RT. The molecular details of NNRTI inhibition are not clearly understood.
Structural studies of HIV-1 RT have shown that:
▷ the NNIBP primarily occurs from the p66 subunit (near the polymerase active site)
▷ the NNIBP is created by structural rearrangements, particularly an extended conformation of the primer grip and rearrangements of the aromatic-ring containing residues Y181 and Y188. This structural rearrangement locks the p66 thumb and fingers in their hyper-extended conformations (16, 23, 24);
▷ NNRTI-resistance mutations are located in and around the NNIBP (16, 23).

Chapter 2: Materials & Methods

Materials

Data Repository - The Protein Data Bank (PDB)

All protein structures are obtained from the RCSB Protein Data Bank (PDB), which is an Information portal to Biological Macromolecular Structures.
(http://www.rcsb.org/pdb/home/home.do).
It is a repository of the 3-Dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data are typically obtained by X-ray crystallography or NMR spectroscopy, and submitted by structural biologists from around the world (25). This repository was founded in 1971 by Brookhaven National Laboratory, New York, and then transferred to the Research Collaboratory for Structural Bioinformatics.
(RCSB: http://home.rcsb.org/) in 1998. The wwPDB was formed in 2003, and the PDB became an international organization.

Each structure published in PDB receives a four-character alphanumeric identifier, its PDB ID or PDB identifier. For example, 1AGS refers to a surface mutant (G82R) of a human alpha-glutathione S-transferase that shows decreased thermal stability and a new mode of molecular association in the crystal. All information related to this structure (citation, experiments, etc) can be obtained by using the PDB ID. The le format used by the PDB is known as the PDB le format. All structure data derived from X-ray diffraction and NMR studies in the PDB appear in a standard representation. Mandatory record types are present in all entries. The REMARK records present experimental details, annotations, comments, and information not included in other records.

There are REMARKS like REMARK 465, REMARK 470, etc, which are important while looking at the structure of a protein. REMARK 465 lists the residues that are present in the SEQRES records but are completely absent from the coordinates section.
REMARK 470 lists Non-hydrogen atoms of standard residues which are missing from the coordinates.

Data

The four structural classes of proteins as given in Structural Classification of Proteins (SCOP) chosen are: α, β, $\alpha + \beta$, and α/β. The α proteins are composed predominantly of α helices, and the β proteins of β sheets. The $\alpha + \beta$ proteins mainly have anti-parallel β sheets, whereas those in α/β consist of mainly parallel beta sheets. We consider 51 proteins from each of these classes whose sizes range from 106 to 787 amino acids. The structural data is obtained from the Protein Data Bank (PDB).
The PDB ID, names, size of the protein sequence, and the missing residues are listed in the following Table. The SEQRES records all residues in the sequence of the protein, but some residues may not be present in the model due to disorder, lack of electron density, etc. Those are listed as missing residues.

S. NO.	PDB ID	Name of the Protein	Size	Missing Residues
ALL ALPHA				
1	1AGS	GLUTATHIONE S-TRANSFERASE ALPHA	221	
2	1BJ5	HUMAN SERUM ALBUMIN	585	1, 2, 585
3	1BUC	BUTYRYL-COA DEHYDROGENASE	383	
4	1DSG	CYTOCHROME C PEROXIDASE	292	3
5	1E7E	SERUM ALBUMIN	585	1, 2, 585
6	1E7G	SERUM ALBUMIN	585	1, 2, 585
7	1G5Z	OUTER SURFACE PROTEIN C	164	
8	1GLN	GLUTAMYL-TRNA SYNTHETASE	468	
9	1GNE	GLUTATHIONE S-TRANSFERASE	232	
10	1HA2	SERUM ALBUMIN	585	1, 2
11	1HLB	HEMOGLOBIN (DEOXY)	158	
12	1KNY	KANAMYCIN NUCLEOTIDYLTRANSFERAS	253	
13	1MBS	MYOGLOBIN	153	
14	1N29	PHOSPHOLIPASE A2, MEMBRANE	124	
15	1NXG	CITRATE SYNTHASE	427	1, 1000
16	1O2E	PHOSPHOLIPASE A2	123	
17	1ONP	1-DEOXY-D-XYLULOSE-5-PHOSPHATE	398	398(A,B)
18	1PMB	MYOGLOBIN	153	
19	1PMT	GLUTATHIONE TRANSFERASE	203	202, 203
20	1QM6	PHOSPHOLIPASE C	370	
21	1QUU	SKELETAL MUSCLE ALPHA-ACTININ 2	250	2, 49, 250
22	1R5A	GLUTATHIONE TRANSFERASE	218	1, 216-218
23	1SAV	ANNEXIN V	320	2, 319, 320
24	1SCH	PEANUT PEROXIDASE	294	
25	1SIR	GLUTARYL-CoA DEHYDROGENASE	394	1, 2, 393, 394
26	1SRY	SERYL-tRNA SYNTHETASE	421	
27	1TD7	PHOSPHOLIPASE A2, ISOFORM 3	119	

28	1UA5	GLUTATHIONE S-TRANSFERASE	218	0, 216, 217
29	1ULY	HYPOTHETICAL PROTEIN PH1932	192	1, 192
30	1UX5	BNI1 PROTEIN	411	
31	1VAR	MANGANESE SUPEROXIDE DISMUTASE	198	
32	1XDP	POLYPHOSPHATE KINASE	687	
33	1XJL	ANNEXIN A2	319	
34	1XWM	PHOSPHATE UPDATE REGULATOR	217	216, 217
35	2ALX	RIBONUCLEOTIDE-DIPHOSPHATE REDUCTASE	340	
36	2BXI	SERUM ALBUMIN	585	1, 2, 585
37	2BXM	SERUM ALBUMIN	585	1, 2, 585
38	2EA2	METHIONINE AMINOPEPTIDASE 2	369	
39	2FNP	STAPHYLOCOCCAL ACCESSORY REGULATOR A	124	101, 102(A/B)
40	2FRH	STAPHYLOCOCCAL ACCESSORY REGULATOR A	127	
41	2G0G	PEROXISOME PROLIFERATOR-ACTIVATED RESP	271	
42	2GTU	GLUTATHIONE S-TRANSFERASE	217	
43	2HSG	GLUCOSE-RESISTANCE AMYLASE	332	1, 59, 69, 70
44	2J08	DEOXYRIBODIPYRIMIDINE PHOTO-LYASE	420	
45	2OAD	GLUTATHIONE S-TRANSFERASE P1	209	
46	2OSN	PHOSPHOLIPASE A2 ISOFORM3	118	
47	2VDB	HUMAN SERUM ALBUMIN	579	
48	2ZBH	PHOSPHOLIPASE A2	121	
49	3B9L	SERUM ALBUMIN	585	1, 2, 585
50	3BQD	GLUCOCORTICOID RECEPTOR	255	
51	3C3S	SUPEROXIDE DISMUTASE	198	197, 198 (A/B)
ALL BETA				
1	1A72	HORSE LIVER ALCOHOL DEHYDROGENASE	374	
2	1ADW	PSEUDOAZURIN	123	
3	1ASP	ASCORBATE OXIDASE	552	
4	1B0O	BETA-LACTOGLOBULIN	162	
5	1B88	T CELL RECEPTOR V-ALPHA DOMAIN	114	
6	1B90	PROTEIN (BETA-AMYLASE)	516	
7	1BAG	ALPHA-1,4-GLUCAN-4-GLUCANOHYDROLASE	425	
8	1BNM	CARBONIC ANHYDRASE	259	2, 3
9	1BVZ	PROTEIN (ALPHA-AMYLASE II	585	
10	1CD8	T CELL CORECEPTOR CD8	114	
11	1CDJ	T-CELL SURFACE GLYCOPROTEIN CD4	178	
12	1CK1	PROTEIN (ENTEROTOXIN TYPE C-3)	239	
13	1DAB	P.69 PERTACTIN	539	
14	1DU5	ZEAMATIN	206	
15	1DXM	H PROTEIN	131	
16	1E3L	ALCOHOL DEHYDROGENASE, CLASS II	376	1, 2, 3
17	1ECZ	ECOTIN	142	

18	1EFT	ELONGATION FACTOR TU	405	
19	1F7R	POL POLYPROTEIN	136	134-136
20	1FLG	QUINOPROTEIN ETHANOL DEHYDROGENASE	582	
21	1G8G	SULFATE ADENYLYLTRANSFERASE	511	1
22	1HCX	MAJOR AUTOLYSIN	127	
23	1HDY	ALCOHOL DEHYDROGENASE	374	
24	1HDZ	ALCOHOL DEHYDROGENASE	374	
25	1HNF	CD2	182	1, 2, 3
26	1HSO	CLASS I ALCOHOL DEHYDROGENASE 1	374	
27	1I9E	CYTOTOXIC TCELL VALPHA DOMAIN	115	
28	1IIU	CYTOTOXIC TCELL VALPHA DOMAIN	174	
29	1JE6	MHC CLASS I CHAIN PROTEIN	275	
30	1JSG	ONCOGENE PRODUCT P14TCL1	114	1, 2, 3
31	1KWQ	CARBONIC ANHYDRASE II	260	1, 2
32	1LIL	LAMBDA III BENCE JONES PROTEIN CLE	212	
33	1LU1	LECTIN	253	
34	1LWH	4-ALPHA-GLUCANOTRANSFEARSE	441	
35	1LWJ	4-ALPHA-GLUCANOTRANSFEARSE	441	
36	1LXA	UDP N-ACETYLGLUCOSAMINE O-ACYLTRANSFERASE	262	
37	1MC5	ALCOHOL DEHYDROGENASE CLASS III	374	
38	1MOE	ANTI-CEA mAb T84.66	240	
39	1QD0	VHH-R2 ANTI-RR6 ANTIBODY	128	
40	1QGL	PROTEIN (SUCCINYLATED CONCANAVALIN A)	237	
41	1QM6	PHOSPHOLIPASE C	370	
42	1THU	THAUMATIN ISOFORM B	207	
43	1ULC	GALECTIN-2	150	
44	1XYH	CYCLOPHILIN-LIKE PROTEIN PPIL3B	161	161
45	1Z24	INSECTICYANIN A FORM	189	
46	1ZAP	SECRETED ASPARTIC PROTEINASE	342	251
47	1ZSA	CARBONIC ANHYDRASE II	259	2, 3
48	2ANS	ADIPOCYTE LIPID-BINDING PROTEIN	131	
49	2BML	AUTOLYSIN	126	
50	3BLG	BETA-LACTOGLOBULIN	162	
51	3DPA	CHAPERONE PROTEIN PAPD	218	
$\alpha + \beta$				
1	1AN9	D-AMINO ACID OXIDASE	340	
2	1B02	PROTEIN (THYMIDYLATE SYNTHASE)	279	
3	1B8U	PROTEIN (MALATE DEHYDROGENASE)	329	1, 2
4	1BQ1	THYMIDYLATE SYNTHASE	264	
5	1BQI	PAPAIN	212	
6	1BSP	THYMIDYLATE SYNTHASE A	278	23

7	1C68	PROTEIN (LYSOZYME)	164	
8	1CF5	PROTEIN (BETA-MOMORCHARIN)	249	
9	1CJ3	P-HYDROXYBENZOATE HYDROXYLASE	392	
10	1DC4	GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE	330	
11	1DD5	RIBOSOME RECYCLING FACTOR	185	1
12	1DTP	DIPHTHERIA TOXIN	190	
13	1E5D	RUBREDOXIN OXIDOREDUCTASE	402	1
14	1E7M	GLUCOSE 6-PHOSPHATE 1-DEHYDROGENASE	485	
15	1EU4	SUPERANTIGEN SPE-H	204	
16	1EVI	D-AMINO ACID OXIDASE	340	
17	1FRX	FERREDOXIN	106	
18	1G62	RIBOSOME ANTI-ASSOCIATION FACTOR	224	
19	1H94	GLUCOSE 6-PHOSPHATE 1-DEHYDROGENASE	485	
20	1IUV	P-HYDROXYBENZOATE HYDROXYLAS	394	
21	1JMF	THYMIDYLATE SYNTHASE	316	
22	1JTQ	THYMIDYLATE SYNTHASE	264	
23	1KNY	KANAMYCIN NUCLEOTIDYLTRANSFERASE	253	
24	1LCA	THYMIDYLATE SYNTHASE	316	
25	1NJA	THYMIDYLATE SYNTHASE	316	
26	1NJC	THYMIDYLATE SYNTHASE	316	
27	1PAF	POKEWEED ANTIVIRAL PROTEIN	262	
28	1PBB	P-HYDROXYBENZOATE HYDROXYLASE	394	392-394
29	1PRQ	PROFILIN IA	125	
30	1QS5	LYSOZYME	162	
31	1QTC	LYSOZYME	162	
32	1RBB	RIBONUCLEASE B	124	
33	1RAA	ASPARTATE CARBAMOYLTRANSFERASE	310	
34	1RTB	RIBONUCLEASE A	124	
35	1SRY	SERYL-tRNA SYNTHETASE	421	
36	1TSW	THYMIDYLATE SYNTHASE	316	
37	1TSX	THYMIDYLATE SYNTHASE	316	
38	1VAR	MANGANESE SUPEROXIDE DISMUTASE	198	
39	1VZA	THYMIDYLATE SYNTHASE	316	
40	1YFP	YELLOW FLUORESCENT PROTEIN	227	
41	1ZPR	THYMIDYLATE SYNTHASE	264	
42	2ACG	PROFILIN II	125	
43	2DPG	GLUCOSE 6-PHOSPHATE DEHYDROGENASE	485	
44	2EQL	HORSE MILK LYSOZYME	129	
45	4MDH	CYTOPLASMIC MALATE DEHYDROGENASE	334	
46	7LYZ	HEN EGG WHITE LYSOZYME	129	
47	8LYZ	HEN EGG WHITE LYSOZYME	129	
48	171L	T4 LYSOZYME	164	1, 63, 164

49	177L	T4 LYSOZYME	164	1, 63, 164
50	189L	T4 LYSOZYME	164	
51	231L	T4 LYSOZYME	164	
α/β				
1	1GZD	GLUCOSE-6-PHOSPHATE ISOMERASE	557	555-557
2	1HJQ	BETA-1,4-GALACTANASE	332	
3	1I4N	INDOLE-3-GLYCEROL PHOSPHATE SYNTHASE	251	
4	1IEI	ALDOSE REDUCTASE	316	
5	1IPE	TROPINONE REDUCTASE-II	259	
6	1IPF	TROPINONE REDUCTASE-II	259	
7	1J9A	OLIGORIBONUCLEASE	184	
8	1J42	RNA-BINDING PROTEIN	189	1, 2, 189
9	1KAK	PROTEIN-TYROSINE PHOSPHATASE	298	1
10	1KF0	PHOSPHOGLYCERATE KINASE	416	
11	1L9G	CONSERVED HYPOTHETICAL PROTEIN	192	192
12	1LKZ	RIBOSE-5-PHOSPHATE ISOMERASE	219	
13	1LWJ	4-ALPHA-GLUCANOTRANSFERASE	441	
14	1N8T	GLUCOSE-6-PHOSPHATE ISOMERASE	557	
15	1NUH	GLUCOSE PHOSPHATE ISOMERASE	558	556, 557
16	1NVG	NAD-DEPENDENT ALCOHOL DEHYDROGENASE	347	
17	1ONP	REDUCTOISOMERASE	398	398
18	1P5R	FORMYL-COENZYME A TRANSFERASE	428	1
19	1PP4	RHAMNOGALACTURONAN ACETYLESTERASE	233	
20	1PWZ	HALOHYDRIN DEHALOGENASE	254	1, 254
21	1Q7C	3-OXACYL-[ACYL-CARRIER-PROTEIN] REDUCTASE	244	
22	1R5A	GLUTATHIONE TRANSFERASE	218	
23	1R8W	GLYCEROL DEHYDRATASE	787	1
24	1RHD	RHODANESE	293	
25	1SBT	SUBTILISIN BPN	275	
26	1T3B	THIOL:DISULFIDE INTERCHANGE PROTEIN DSBC	211	211
27	1TIM	TRIOSEPHOSPHATE ISOMERASE	247	
28	1TJD	THIOL:DISULFIDE INTERCHANGE PROTEIN DSBC	216	
29	1TO6	GLYCERATE KINASE	371	
30	1TUU	ACETATE KINASE	399	399
31	1UA5	GLUTATHIONE S-TRANSFERASE	218	216, 217
32	1UD6	AMYLASE	480	
33	1UXP	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE	501	1, 2
34	1VFF	BETA-GLUCOSIDASE	423	
35	1W93	ACETYL-COENZYME A CARBOXYLASE	553	535-538
36	1XHE	arcA	123	1, 123
37	1XLC	D-XYLOSE ISOMERASE	394	1

S. NO.	PDB ID	Name of the Protein	Size	Missing Res.
38	1Y8C	METHYLTRANSFERASE	246	
39	1YS9	PROTEIN SPy1043	254	1
40	1ZMG	MALTOSE-BINDING PERIPLASMIC -PROTEIN	370	
41	2AXP	HYPOTHETICAL PROTEIN BSU20280	173	1
42	2BFR	HYPOTHETICAL PROTEIN AF1521	192	
43	2BKW	ALANINE-GLYOXYLATE AMINOTRANSFERASE 1	385	
44	2BVM	TOXIN B	542	
45	2DSV	CHITINASE-3-LIKE PROTEIN 1	361	
46	2J0W	LYSINE-SENSITIVE ASPARTOKINASE 3	449	1, 2
47	2NTJ	ENOYL-[ACYL-CARRIER-PROTEIN] REDUCTASE	268	
48	2OV4	TRYPTOPANYL-tRNA SYNTHETASE	328	
49	2PKX	TRANSCRIPTIONAL REGULATORY PROTEIN	121	120-121
50	2PR2	ENOYL-ACP REDUCTASE	269	1, 2
51	2QMY	DIHYDRONICOTINAMIDE DEHYDROGENASE	230	

Data for HIV-1 RT

A dataset of inhibitor bound HIV-1 Reverse Transcriptase with and without resistance mutation is selected from the Protein Data Bank, taking efavirenz (EFZ) as nonnucleoside reverse transcriptase inhibitor. For the Apo form Reverse transcriptase we are using 1RTJ.
While selecting the pdb file of data the following points were kept in mind:
▷ The experimental condition for crystallization were similar for the dataset.

▷ The PDB ID's with less number of missing residues were selected.

▷ The dataset with resolution less than $3\mathring{A}$ was selected.

PDB Files	Inhibitor Bounded	Resistance Mutation	Resolution
1RTJ	_	_	2.35
1IKW	EFZ	_	3
1FKO	EFZ	K103N	2.9

Methods

Softwares & Packages

Introduction to Matlab

MATLAB is a software package developed to perform numerical calculations on vectors and matrices. It was developed by MathWorks. The software has many features :
▷ It can do sophisticated graphics in two and three dimensions.
▷ It allows matrix manipulations and plotting of functions and data.
▷ It contains a high-level programming language (a baby C) which makes it quite easy to code complicated algorithms involving vectors and matrices.
▷ It can numerically solve nonlinear initial-value ordinary differential equations.
▷ It can numerically solve nonlinear boundary-value ordinary differential equations.
▷ It contains a wide variety of toolboxes which allow it to perform a wide range of applications from science and engineering.
▷ It can interface with programs written in other languages, which includes C, C++, Java and Fortran.

The one crucial feature of MATLAB is that it can group large amounts of data in arrays and perform mathematical operations on this data as individual arrays rather than as groups of data. This makes it very easy to apply complicated operations to the data and also the chances of errors are less (26).

Introduction to R

R is a package of software facilities for data manipulation, calculations and graphical representation. It was initially written and released as an open source software by Ross Ihaka and Robert Gentleman at University of Auckland during 90s.

Some salient features of R :
▷ An effective storage facility and data handling.
▷ It has various operators for calculations on arrays & matrices.
▷ Large, coherent and integrated collection of functions.
▷ Graphical facilities for data analysis.
▷ It has a well developed, simple and effective programming language (called S) which includes conditionals, loops, user defined recursive functions and input and output facilities.

All things in R like functions, datasets, results, etc. are called OBJECTS. Graphs are not stored as objects. Script is written so as to create an object and required graphs (27).

Introduction to igraph

igraph is a free software package for creating and manipulating undirected and directed graphs. It includes implementations for classic graph theory problems like minimum spanning trees and network flow, and also implements algorithms for some recent network analysis methods, like community structure search.
Some salient features of igraph:
▷ It has functions for generating regular and random graphs,
▷ There are functions for manipulating graphs.
▷ Numeric or textual attribute can be assigned to the vertices or edges of the graph, like edge weights or textual vertex ids.
▷ It can calculate various structural properties, graph isomorphism.
▷ It supports many file formats, e.g., GraphML, GML or Pajek.
▷ The R package and python can visualize graphs in many ways - 2D and 3D, interactively or non-interactively.
▷ igraph runs on most operating systems -MS Windows, Mac OSX and various Linux versions.
▷ The software for igraph is different for C library, the R package and python extension (27).

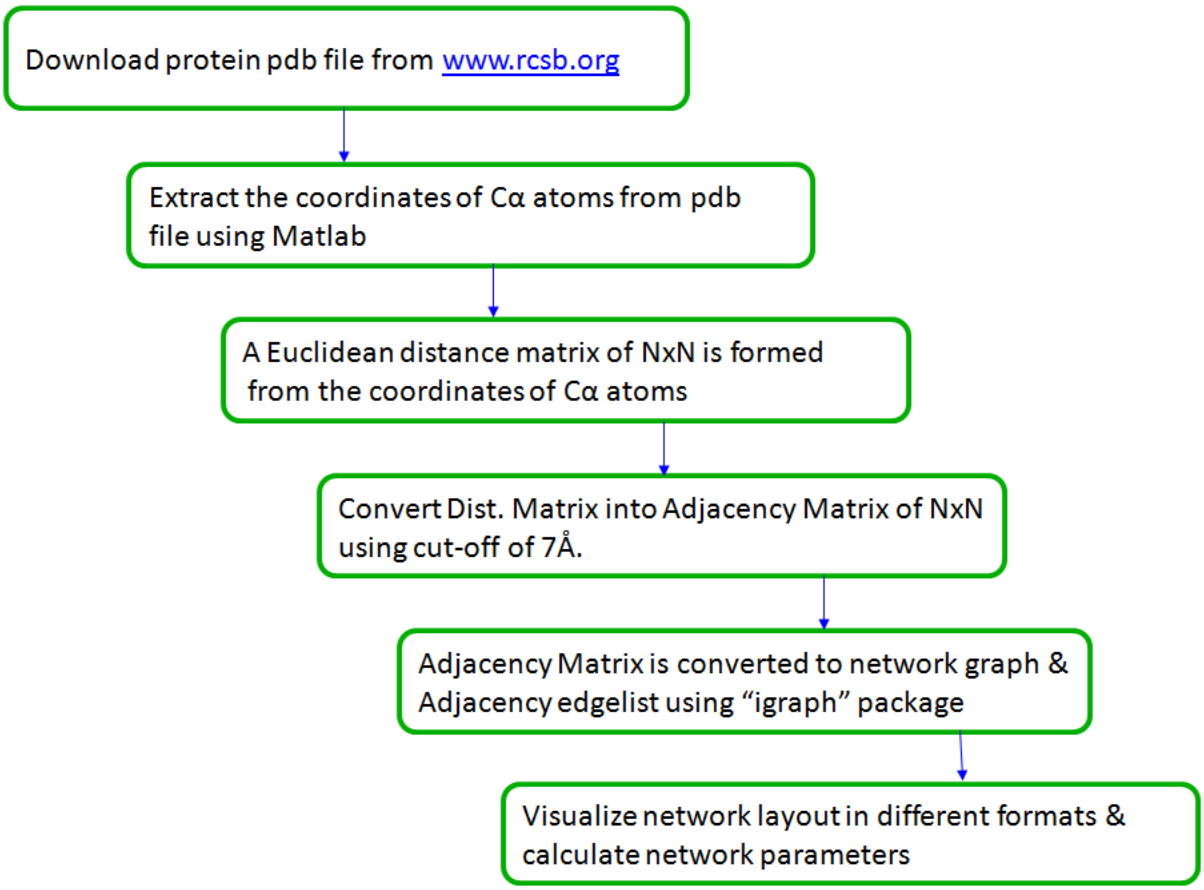
Introduction to Pymol

PyMol is an open source molecular graphics software. It was created by Warren Lyford DeLano and commercialized by DeLano Scientific LLC, which is a private software company. The software is based on python language.
The software helps in visualizing and creating images of complex macromolecules and is widely used in structural studies (28).

Introduction to Pajek

The word 'Pajek' is from Solvenian language and it means spider. The pajek is a program for analysis and visualization of large networks having some thousands or even millions of vertices.
The motivation behind development of Pajek was the observation that there exist several sources of large networks. Pajek provide tools for analysis and visualization of such large networks like: collaboration networks, organic molecule in chemistry, protein-receptor interaction networks, Internet networks, citation networks, data-mining (2-mode networks), etc.
The main goals in the design of Pajek are:
▷ to provide the user with some powerful visualization tools;
▷ to support abstraction by (recursive) decomposition of a large network into several smaller networks that can be treated further using more sophisticated methods;
▷ to implement a selection of efficient (subquadratic) algorithms for analysis of large networks (29).

Constructing Protein Contact Network



Extracting Cα Coordinates

▷ We are using Matlab for extracting Cα coordinates from MODEL present in the PDB file format. The program is specific for ATOM and chain A residues of the Cα atoms.
▷ A Euclidean distance matrix is formed by the coordinates using, giving the NxN matrix representing the spacing of a set of N points in Euclidean space, where N represents Cα atom.
▷ We convert the distance matrix into a Adjacency Matrix of (0,1) data point, all the distances above 7$\mathring{A}$ are given value **0** and less than 7$\mathring{A}$ are **1**. The adjacency matrix gives us the linkage or interaction of a Cα atom with each other. Creating a **Network Map** with Cα atoms as **Nodes** and interaction of different Cα atom as their **Links**.

Constructing ligand binding residue contact network

When ligand is interacting with a protein, we should know the key amino acid residues which forms contact with the ligand and helps in understanding the ligand binding pocket inside the protein, the below flowchart is a basic algorithm for creating a ligand binding contact network.

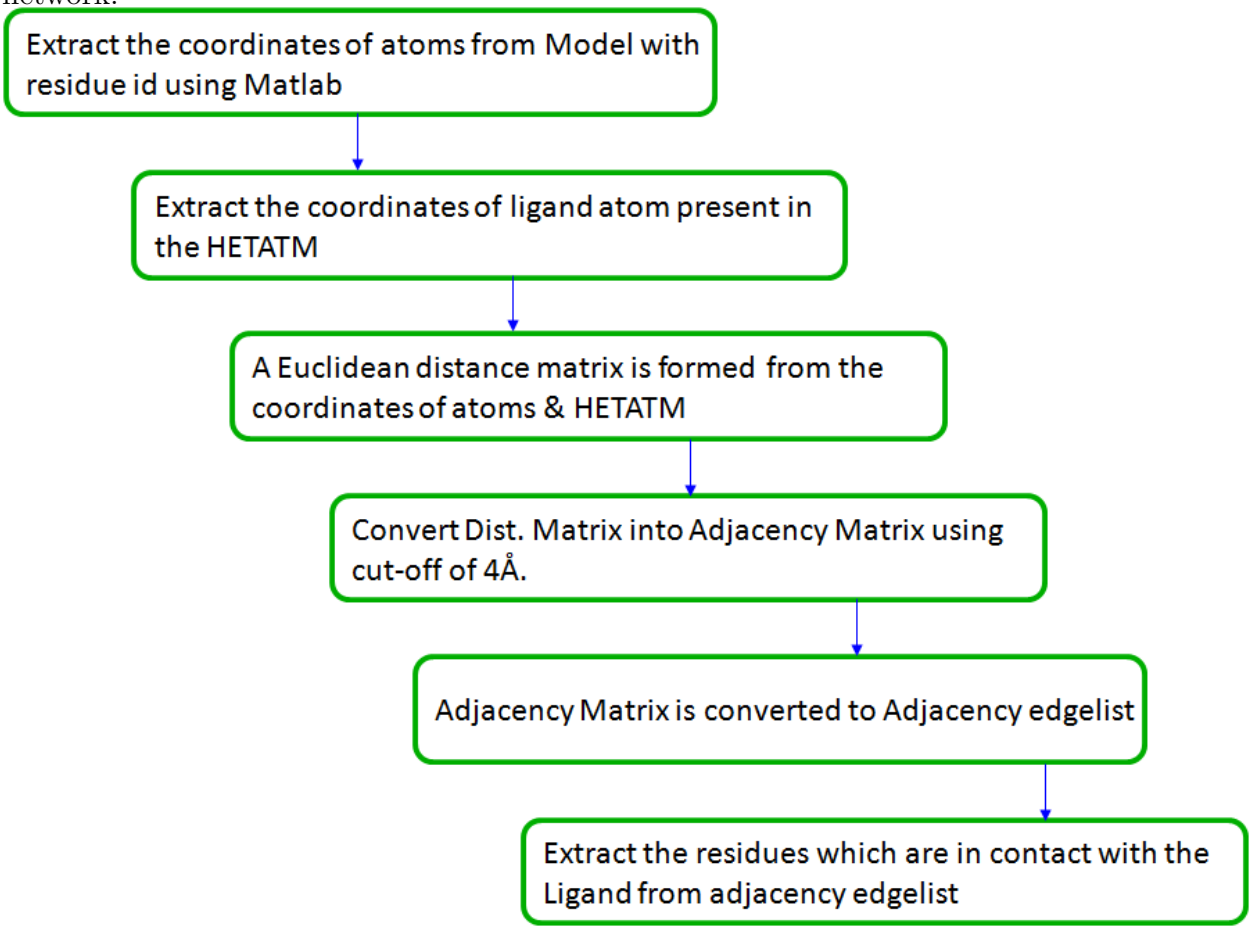


Figure 8: Flowchart for constructing a contact network of ligand binding residue.

Network Visualization

Contact Matrix

The two dimensional contact plot gives us the interaction between Cα atoms of all the residues in the protein. In the alpha helix, there is hydrogen bond interaction between oxygen atom of i residue and NH4+ of the (i+4) residues. The beta sheets have hydrogen bonding between the atoms of adjacent strands which can run either parallel or antiparallel.

The alpha helices are depicted as the interaction of a Cα with its four immediate neighbours and is shown on the diagonal of the graph.

The β-sheets are shown by interaction between residues of adjacent strands in the beta sheets. The parallel β-sheets are depicted as a set of interactions lying parallel to the diagonal, whereas antiparallel sheets are perpendicular to the diagonal.

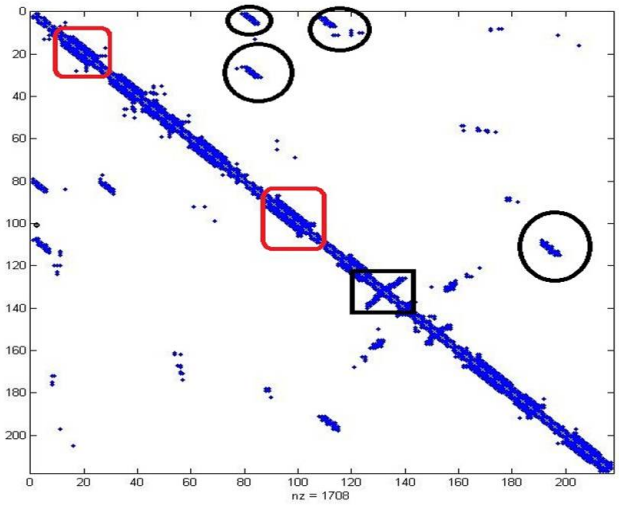


Figure 9: Contact Plot

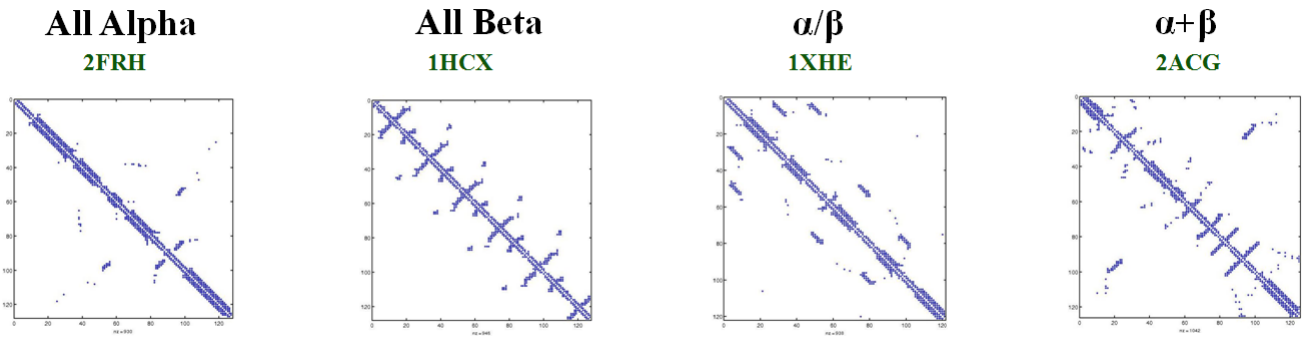


Figure 10: Different contact plots for different Structural Classes of Protein

2-Dimensional network layout

The Kamada-Kawai algorithm is a force directed layout algorithm and it achieves faster convergence and is used to layout networks of all sizes. The force directed layout algorithm considers a force between any two nodes. In this algorithm, the nodes are represented by steel rings and the edges are springs between them.

The basic idea is to minimize the energy of the system by moving the nodes and changing the forces between them. To generate the layout, the global energy of all the nodes and

their derivatives are calculated. At each iteration, the node with the highest energy gradient is displaced to make it zero (30).

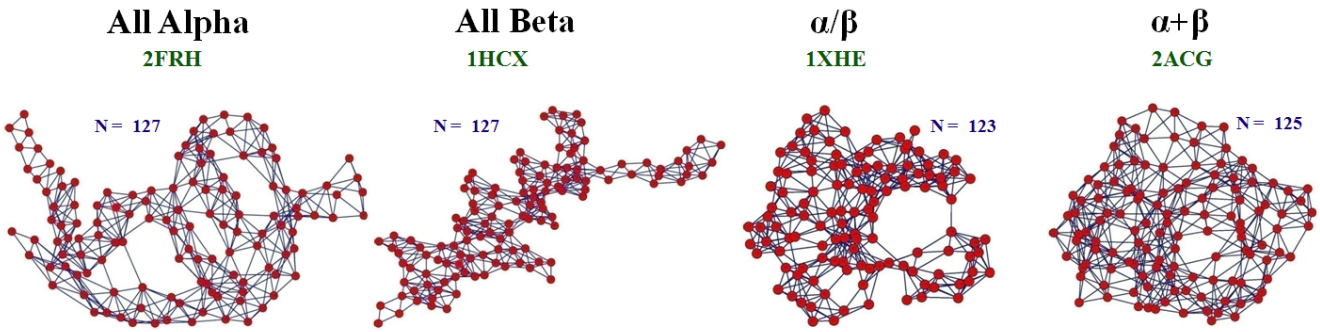


Figure 11: 2-Dimensional network layout

Ring Graph Representation

In this network layout the backbone the of the protein are aligned in a circle. Ring graph helps in visualizing the long range contacts. The long range contacts gives the tertiary structure of proteins.

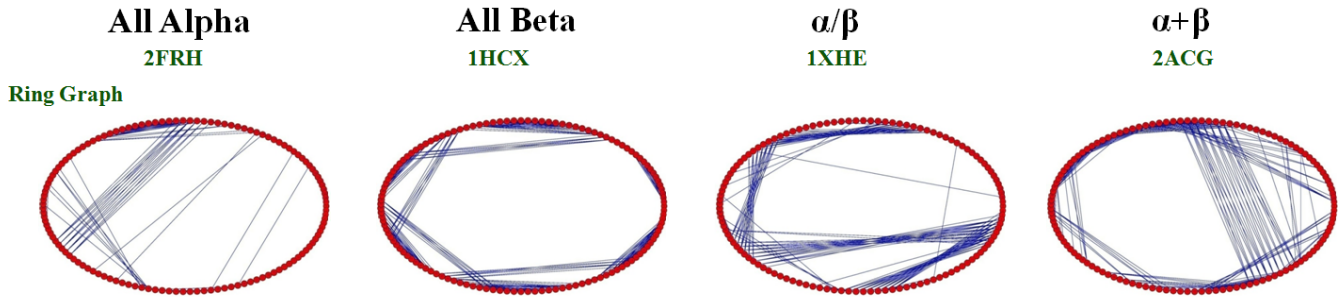


Figure 12: Ring Graph Representation of Proteins

All of the four structural classes of proteins were visualized in different network layout for better understanding of the structural features among the four classes. The red dots refers to the Cα atoms of the corresponding amino acid residue in the above proteins and the connecting line between two red dots (node) are the links which shows the interaction between two residues.

We can clearly see the difference among 2-Dimensional network plots and observe the secondary structural information from it. Ring graph representation shows the long range contacts of the four different structure. The long range contacts differs among the classes as the tertiary structure of proteins are different.

Calculating Network Parameters

We use igraph package in R for calculating network parameters. First the adjacency matrix file is read in R. Library igraph is then uploaded to create graph from adjacency matrix for further analysis. A function tkplot() plots the network graph of the graph adjacency matrix. Now the network parameters are extracted from the graph adjacency matrix.

Statistical tools

Kolmogrov-Smirnov test: It is a non parametric statistical test. KS-test tries to determine if two datasets differ significantly. The KS-test has the advantage of making no assumption about the distribution of data.

Chapter 3: Results

The results are organised in 2 parts:
▷ Part I shows the study of network parameters of the PCNs of all four structural classes of proteins.
▷ Part II shows the development of the PCNs of HIV-1 RT for the apo and holo forms, and illustrate the structural changes during drug binding and resistance mutation to understand their structure-function relationship.

Part I: Major structural classes of proteins

Four datasets, comprising of the three-dimensional coordinates of 51 proteins from each structural class were taken from the Protein Data Bank. PCNs were generated by taking a coarse-grained approach of representing the Cα atoms of the amino acids as nodes, and the interaction distance of 7$\mathring{A}$ between two amino acids as a link between any two Cα atoms.

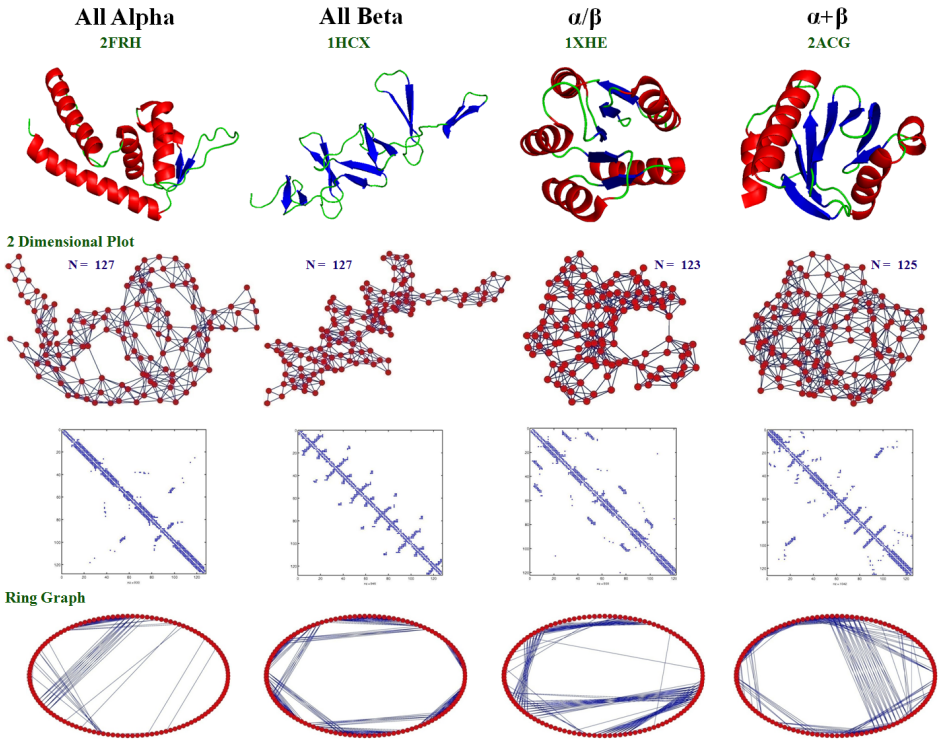


Figure 13: Shows the 3-dimensional structure, 2-dimensional structure PCNs, plots of the adjacency matrix and the ring graph representation of a representative protein from each class.

Each of these network representations highlight different aspects of the protein structure.

Network Parameter Analysis

The network parameters (Degree, Shortest path, Clustering coefficient) were calculated for each node for all 204 proteins. The average shortest path, global clustering coefficient and the degree distribution were found for each PCN of 204 proteins.

Average Shortest Path & Global Clustering Coefficient

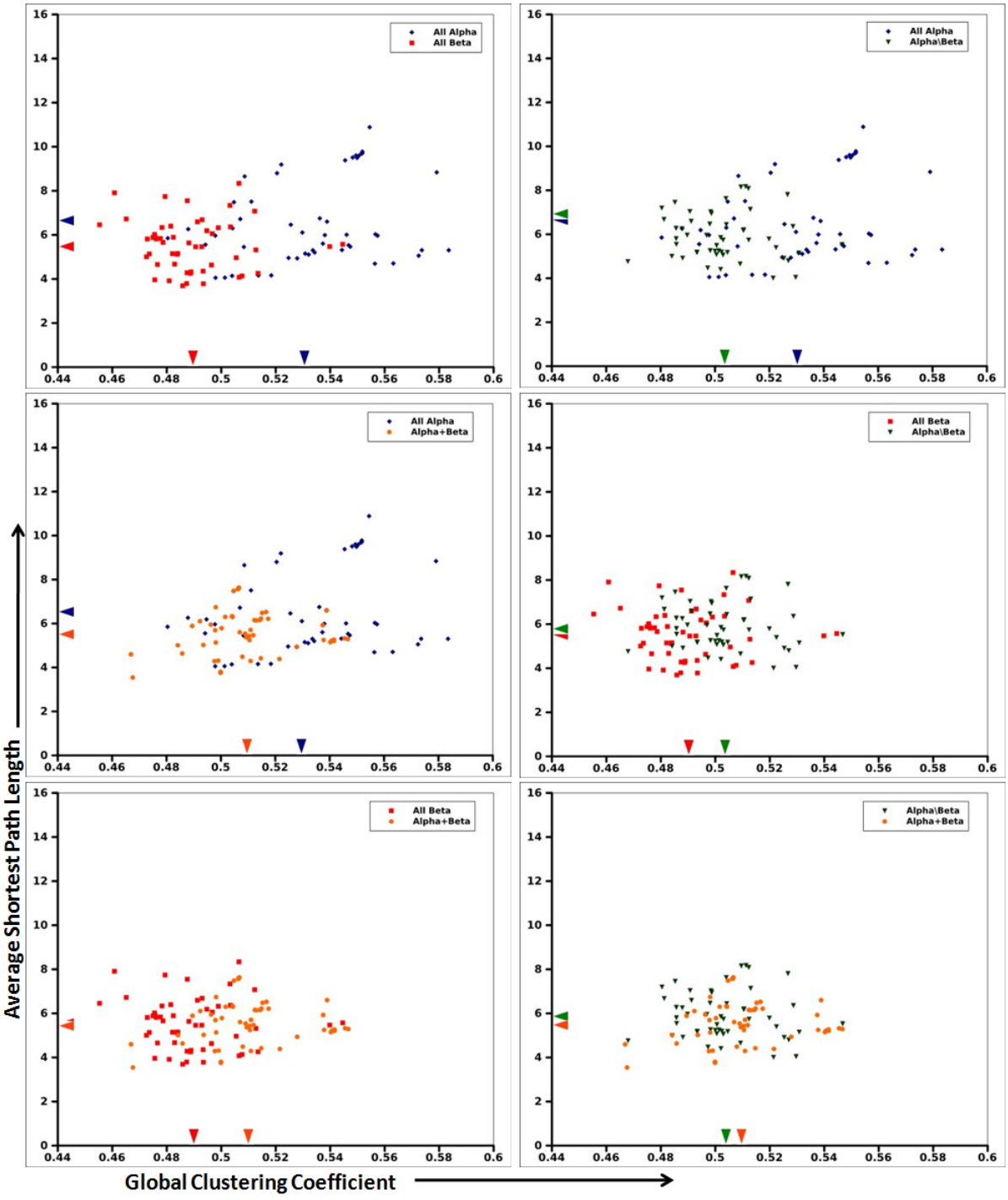


Figure 14: Shows the comparative values (pairwise) of average shortest path and clustering cofficient of the PCNs of four structural classes of proteins.

The arrows in the axes of the plots represent the mean of each structural classes.

The following table gives the means and the standard deviations of the average shortest path and clustering coefficient of different structural classes.

Structural Classes	Average Shortest Path Length (L)	Global Clustering Coefficient (C)
All Alpha (α)	6.559 ± 1.900	0.532 ± 0.025
All Beta (β)	5.552 ± 1.158	0.489 ± 0.017
α/β	5.897 ± 1.072	0.503 ± 0.016
α+β	5.556 ± 0.976	0.510 ± 0.018

To find stastical difference among the different classes, Kolmogrov Simirnov test is done and the results are shown in the following two tables.

Clustering Coefficient Dataset 1	Clustering Coefficient Dataset 2	P valve	Difference b\w Datasets
All Alpha (α)	All Beta (β)	1.88e-11	Significant
All Alpha (α)	α/β	1.75e-07	Significant
All Alpha (α)	α+β	4.73e-06	Significant
All Beta (β)	α/β	1.31e-05	Significant
All Beta (β)	α+β	4.48e-09	Significant
α/β	α+β	0.034	Not Significant

There is consistent difference in the C(Global Clustering Coefficient) of α and β proteins. Except for $\alpha + \beta$ and α/β, all other structural classes show difference in global clustering cofficient at $p < 0.01$, this can be due to the helical structure of the α proteins, where the amino acids are densely packed compared to that of the flat β sheets. This may contribute to the small increase in the Average Clustering Coefficient of the α proteins. Since $\alpha + \beta$ and α/β have a mixed composition of α helices and β sheets the KS-test shows no statistical difference, in this network parameter.

Avg. Shortest Path Length Dataset 1	Avg. Shortest Path Length Dataset 2	P valve	Difference b\w Datasets
All Alpha (α)	All Beta (β)	0.0602	Not Significant
All Alpha (α)	α/β	0.0602	Not Significant
All Alpha (α)	α+β	0.0602	Not Significant
All Beta (β)	α/β	0.2513	Not Significant
All Beta (β)	α+β	0.6902	Not Significant
α/β	α+β	0.5210	Not Significant

It is clear from both figure 14 and the table that the average shortest path do not differ much among the four structural classes. KS test also shows we can not differentiate among classes since the parameter do not differ significantly.

Degree Distribution

The degree of each node in 204 proteins were calculated and their distribution plotted in figure 15. The shape of degree distributions of α, β, $\alpha + \beta$ and α/β protein networks are close to bell-shape, following normal distribution. The number of nodes with very high degree falls off rapidly in all PCNs.

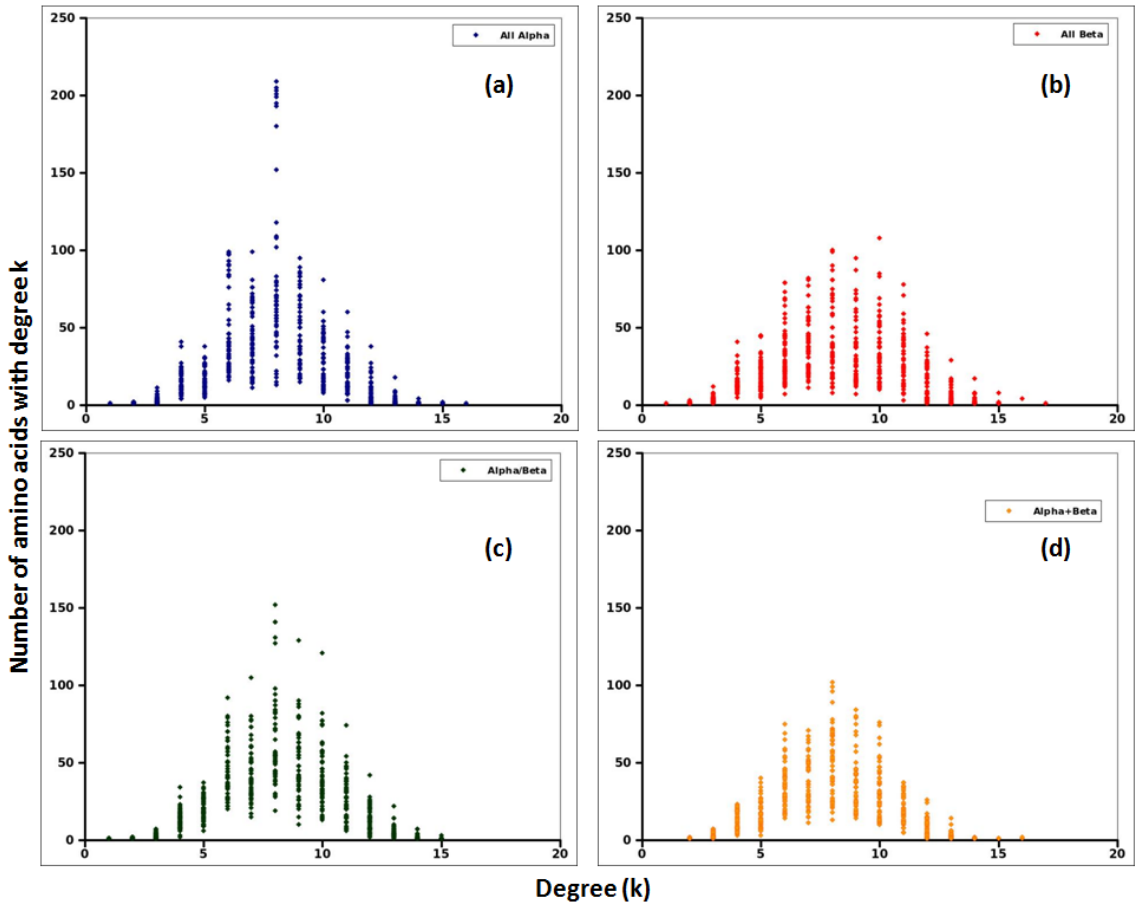


Figure 15: Degree distributions for (a) α, (b) β, (c) α/β and (d) $\alpha + \beta$ proteins, 51 of each class.

This is because there is a physical limit on the number of amino acids that can occupy the space within a certain distance around another amino acid.

The plots clearly show that all four degree distribution show a maximum around 8, but the α protein PCNs seem to have larger number of nodes having degree 8. In figure 16 the number of nodes (amino acids) having degree 8 is plotted for the four structural classes.

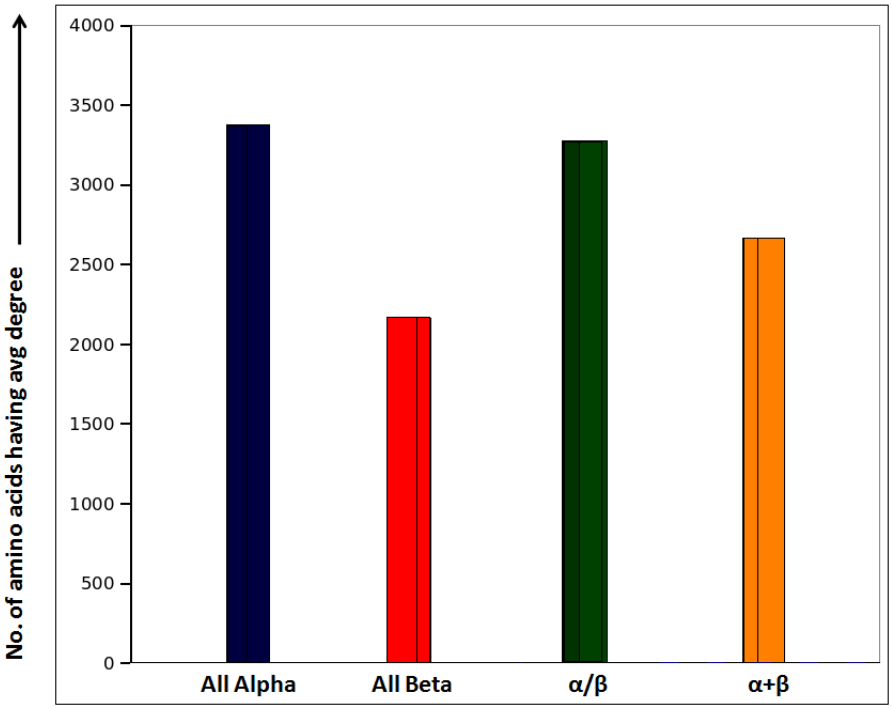


Figure 16: Quantifying number of nodes with degree 8 for α, β, α/β and $\alpha + \beta$ proteins

The above figure quantifies the number of amino acids having degree that is equal to 8. The occurrence of the most common degree differs in different structural classes of protein. In all alpha, 3365 amino acids has 8 degree, while for β, α/β and $\alpha + \beta$ proteins the occurrence of the most common degree are 2138, 3294 and 2652 respectively. This parameter seems to differ between α and β classes.

Protein networks are Small World Networks

We calculated the L (Average shortest path length) and C (Global clustering coefficient) for each protein. As control PCNs, we calculate the L and C of random graphs and regular graphs wih the same no. of nodes (N) and edges (K) (table below).

Network Types	Average Shortest Path Length (L)	Global Clustering Coefficient (C)
Regular Network	34.95 ± 3.216	0.667
Protein Contact Network	5.892 ± 1.381	0.509 ± 0.025
Random Network	3.489 ± 0.551	0.306 ± 0.062

The Kolmogorov-Smirnov test shows that the differences between L and C of the proteins and random or regular lattices are statistically significant.
Figure 17 shows the random control and PCNs of proteins from four structural classes in the L-C plot.

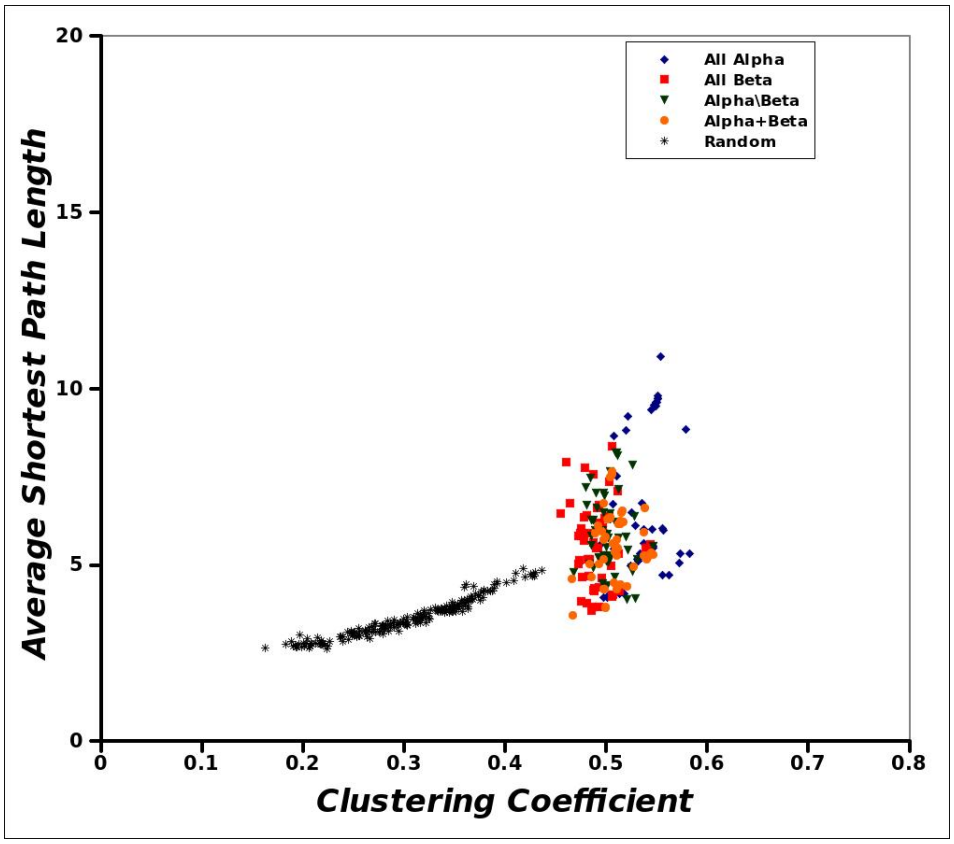


Figure 17: The L-C plot of proteins from four structural classes.

Protein networks have significantly high clustering coefficient than their random counterparts and the L and C-values fall between the random and regular networks in the LC plot. This property of high C and medium L, indicate that protein networks are **Small World Networks**.

Our preliminary study of few network parameters indicate that some parameters differ among structural classes, which can have some functional and structural relevance.

Part II: Relationship between structure and function in proteins

Proteins perform an array of functions in the cell. They perform these specific functions by virtue of their precise structure and chemistry. Structures are a critical determinant of their functions. Hence the study of structure-function relationship, prediction of structure given the sequence etc., are important areas of research.
In this section we study **The Structure-Function Relationship** through network analysis using HIV-1 Reverse Transcriptase as our model protein, we study the functional aspects related to conformational changes in a protein when ligand interacts with the protein. In case of HIV-1 Reverse transcriptase our ligand is efavirenz (EFZ) which is a non-nucleoside revere transcriptase inhibitors (NNRTIs).

Efavirenz binds to NNRTIs binding pocket (NNIBP) which leads to inhibition of HIV-1 RT protein, our objective is to study the protein contact map and to see how contacts of the nodes differ in presence (holo form) and in absence of ligand (apo-form) and also in the protein having resistance mutation.

It is difficult to visualize HIV-1 reverse transcriptase in a network contact map because it's a big protein of about 1000 amino acid residues. To get information regarding structural dissimilarities is tough from the network layout. The 2-dimensional layout for HIV-1 RT is formed with the help of Cytoscape's organic layout. The three structures whose PCNs are shown in figure 18 are - 1RTJ is HIV-1 RT without drug bound to it (apo form), 1IKW and 1FKO are HIV-1 RT with EFZ as a non-nucleoside revere transcriptase inhibitors (NNRTIs) bound to it, and 1FKO is a resistance mutant at K103N.

By looking at the different network visualization techniques (figure 19) one can observe differences among the three proteins.
The contact matrix for the three HIV-1 RT proteins are almost similar with some differences in contacts. The ring graph representation of the proteins show many differences, in long distance contacts.

The conversion from apo to holo form changes the function of the protein (inhibition of RT activity), our aim is to find out the structural aspect in terms of contact pattern changes which leads to the functional change.

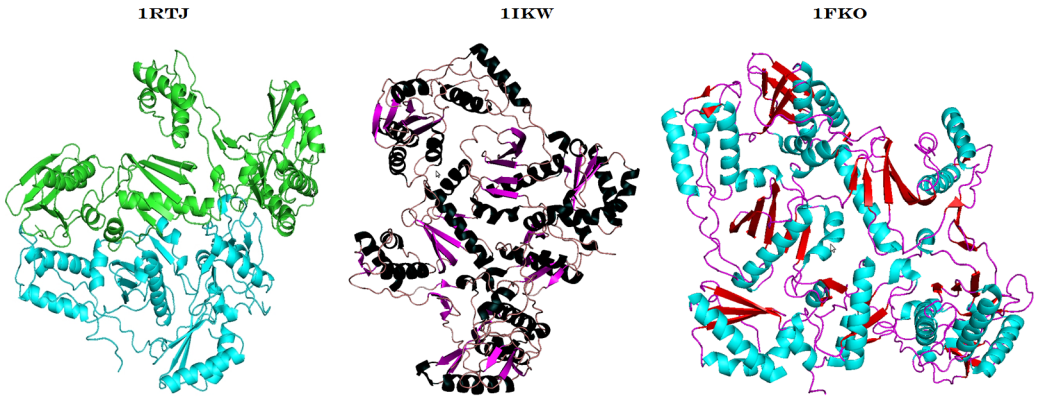


Figure 18: Pymol images for 1RTJ, 1IKW & 1FKO

For this we have looked at the ligand binding pocket of the protein and study the possible loss and gain of contact after drug binds to the apo form.

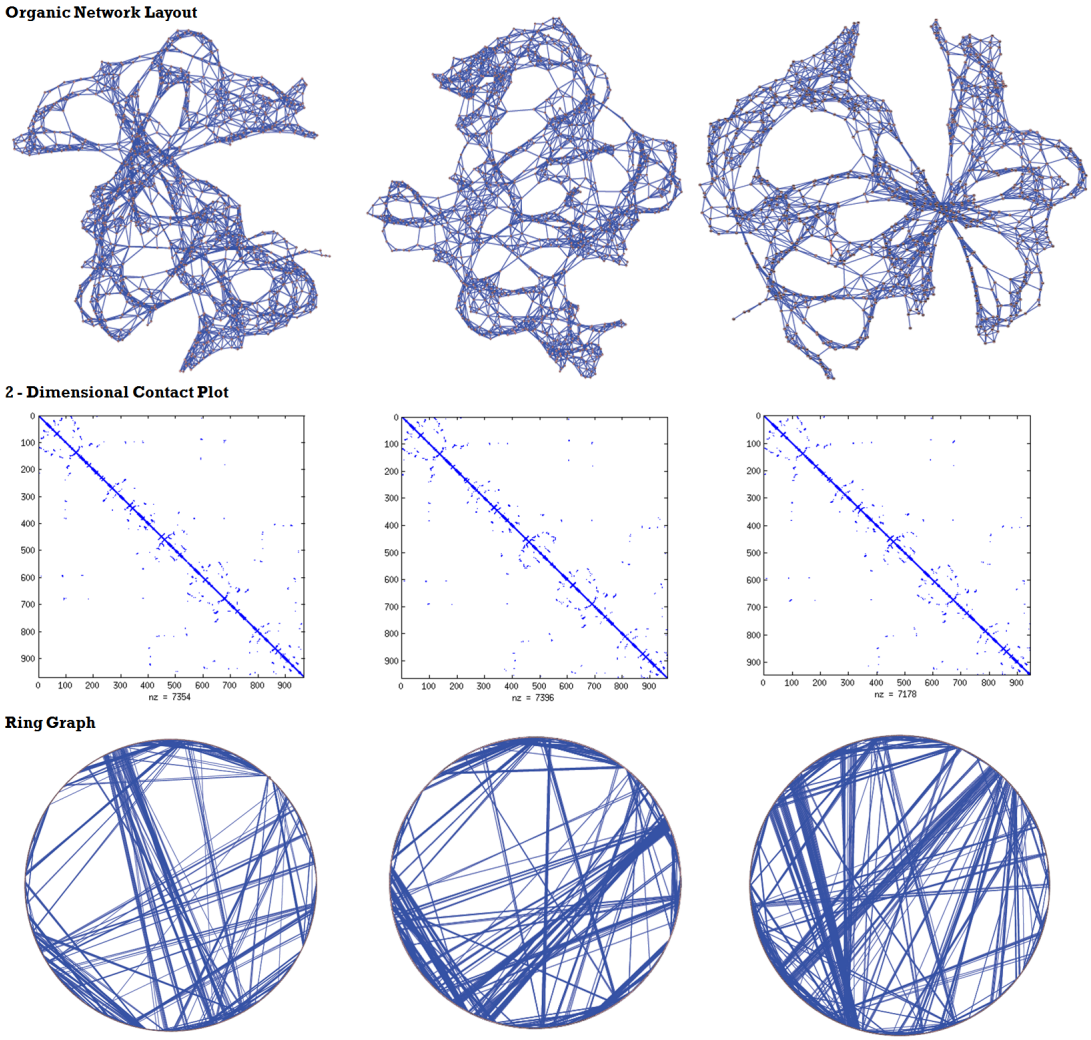


Figure 19: Protein contact network for 1RTJ, 1IKW & 1FKO

Root Mean Square Deviation (RMSD)

The three dimensional structure of the proteins are similar as accounted by average RMSD values.
Figure 20 and the table show the overlap of the three structures and their cross-structural RMSD values. It is clear that the three proteins have quite similar overall structure.
For finer analysis we have done protein contact plot and network parameters analysis to find out the how the proteins with so much of structural similarity differs in functions.

RMSD	1RTJ	1IKW	1FKO
1RTJ	0	1.742	1.015
1IKW	1.742	0	1.393
1FKO	1.015	1.393	0

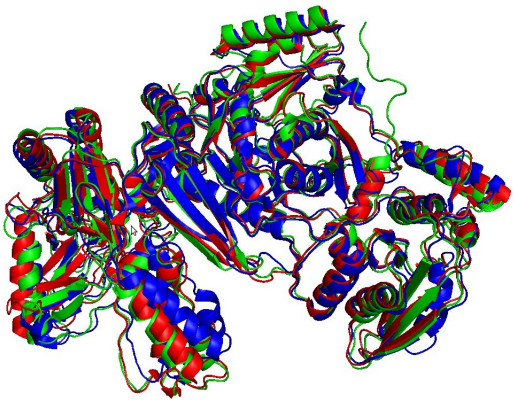


Figure 20: Overlapping Structure - Red : 1FKO; Green : 1RTJ; Blue : 1IKW

Network Parameter Analysis

The network parameters are extracted from the graph adjacency matrix with graphical output for each parameter using igraph package in R language.

Network Parameters Analysis		
PDB Files	Average Shortest Path Length	Global Cluster Coefficient
1RTJ	10.744	0.502
1IKW	10.586	0.499
1FKO	10.865	0.503

The differences of global network parameters among the proteins are not significant. This is indicative of there structural resemblance.

Contact Pattern Analysis

Transition from Unbound State to Bound State

1RTJ is unbound state (apo), while 1IKW and 1FKO are bound state with efavirenz (efz). The following table lists the total number of contacts changes.

▷ **Loss of Contacts** : The contacts between residues which are present in 1st state (unbound

state) but after transition got lost because of the conformation change in the structure.
▷ **Gain of Contacts** : The new contacts formed after the transition to the 2nd state (bound state).

Loss Cont.\Gain Cont.	1RTJ	1IKW	1FKO
1RTJ	0	-	-
1IKW	283\304	O	-
1FKO	147\235	311\202	O

It is clear from the above table that we can not study the transition by taking the whole network contact graphs because too many contacts are changing. Therefore we considered only that residues which are interacting with ligand (Ligand Binding Pocket) in the protein.

Contact changes in ligand binding pocket

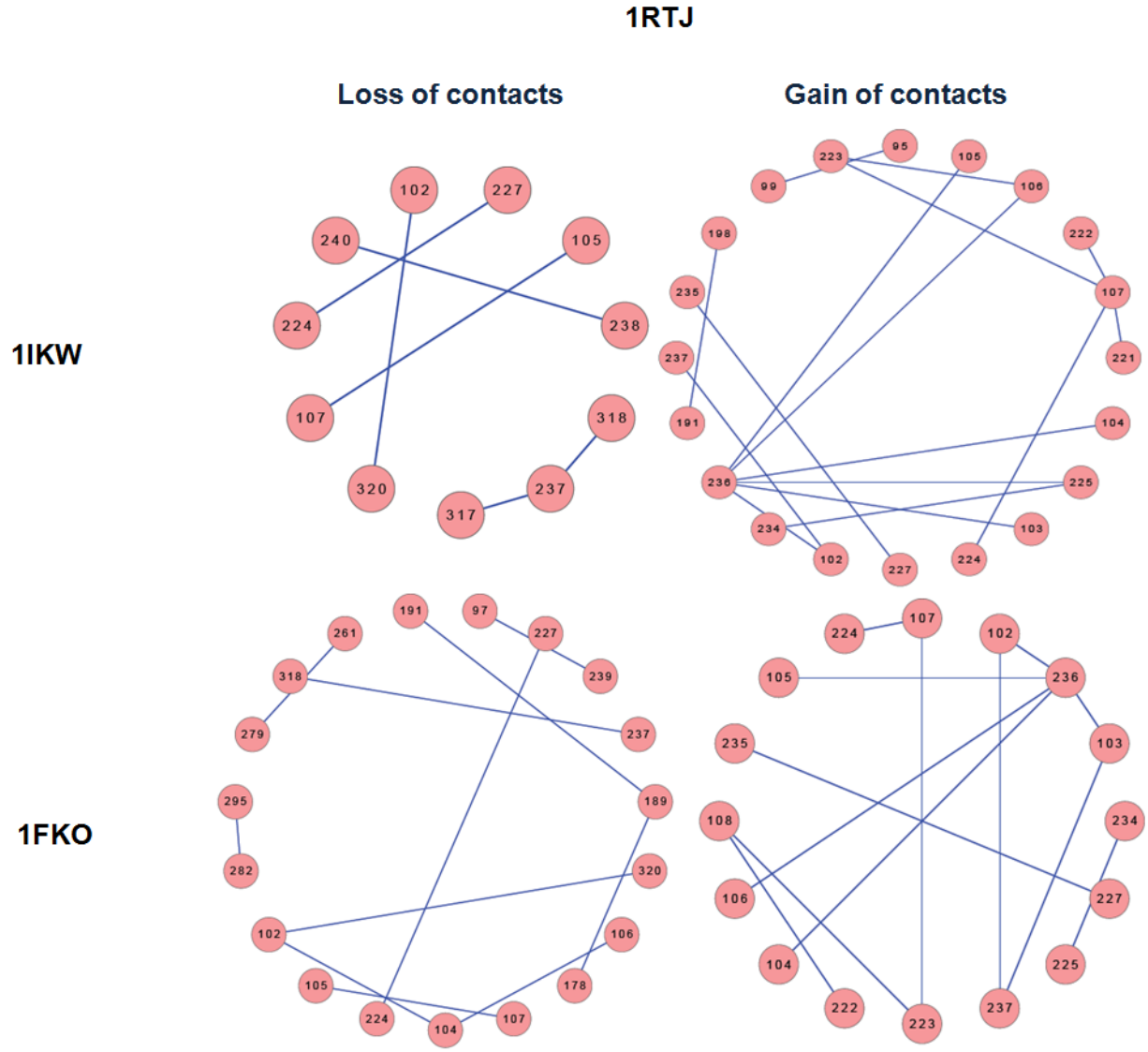


Figure 21: Network representation of contacts which are lost and gained during conformational change

Figure 21 shows all the ligand binding residue contacts gained and lost. Consider 1RTJ first, here the loss of contact means the contacts were present in 1RTJ but got lost when ligand binds, which is 1IKW. The numbers written inside the nodes corresponds to the amino acid residue. The new contacts arising due to ligand interaction are the gain of contacts. Similarly, 1FKO also undergoes loss and gain of contacts as it is a bound state conformation with efz.

Contacts of residues such as 105-107, 102-320, 224-227, 320-102 are observed to be lost in both 1IKW and 1FKO but are present in 1RTJ (apo form).

The common gain of contacts are 102-237, 235-227, 107-223, 105-236, 104-236, 106-236, 107-224. None of these residues were present in the apo form (1RTJ), but due to ligand binding they occurred in both 1FKO and 1IKW. The conformational change because of the ligand binding induced these interactions.

The loss and gain of contacts from a node of network provides information regarding the structural change taking place in the protein.

Understanding p66 (chain A) and p51 (chain B) interface

The HIV-1 Reverse Transcritptase is a heterodimer with two subunits. Our objective is to study the interaction between the larger subunit p66 with the smaller subunit p51, when there is a conformational change from apo to holo form in the protein. The inhibitor binding site is present in chain A (p66).

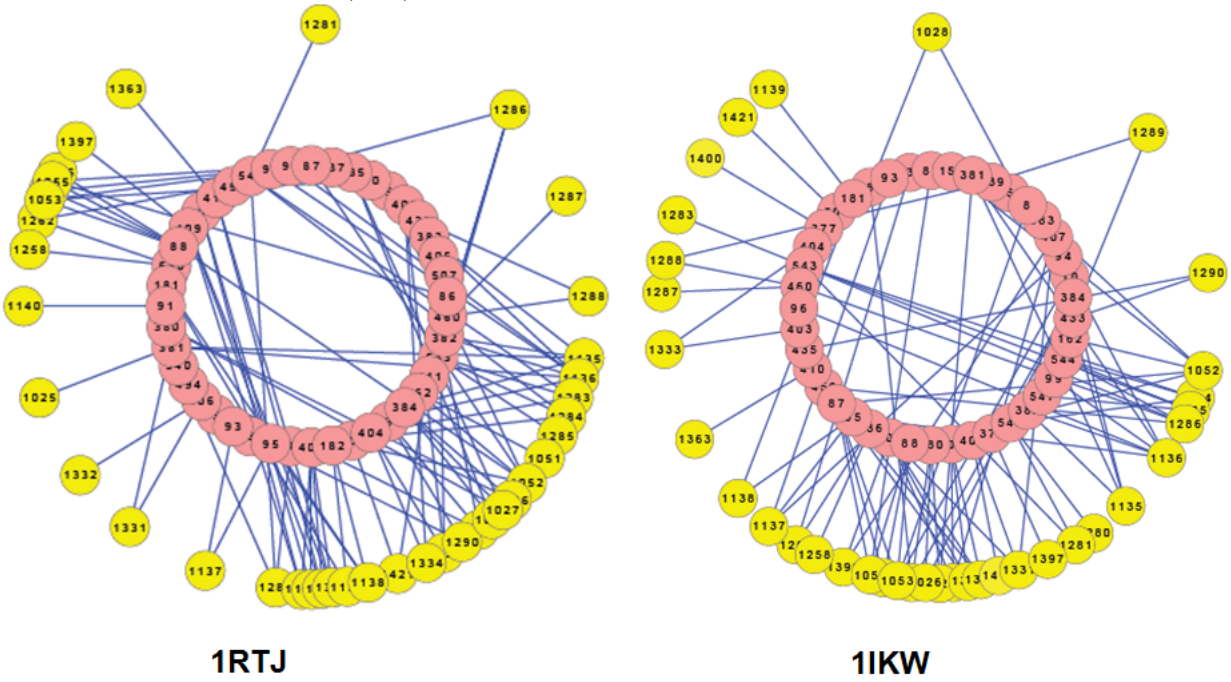


Figure 22: Interaction between residues of p66 (pink) and p51 (yellow) subunits of RT in radial network layout

Figure 22 plots the residues in the two subunits which interact in the apo and holo forms. Many of the contacts are common, so we remove the common interface contacts present in both 1RTJ (apo) and 1IKW (holo), leaving only these residue contacts which are only present in the individual conformations (figure 23).

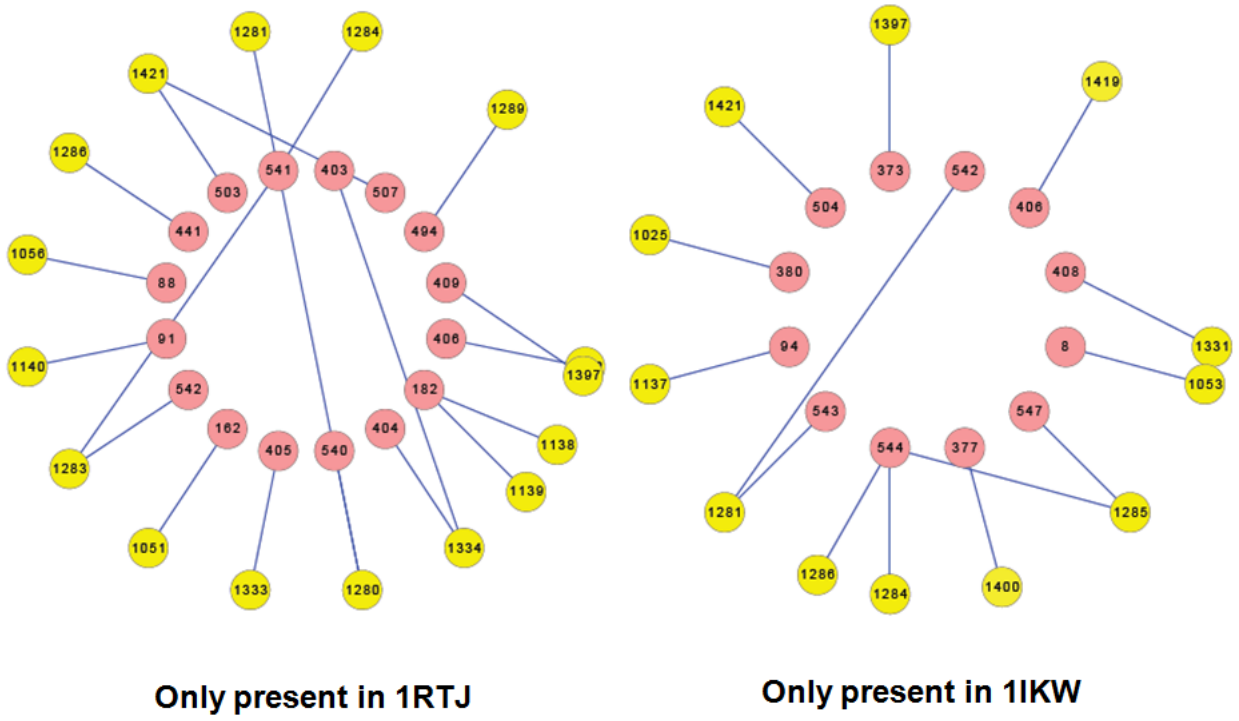


Figure 23: Radial network layout for 1RTJ (apo) and 1IKW (holo) showing interface contacts exclusive to each conformation

The pink nodes depict residues of chain A (p66 subunit) and the yellow nodes are chain B (p51 subunit) residues. We can observe the change in contact area of two subunits because of the binding of ligand. This clearly shows that ligand binding brings about conformational change in the interaction of two subunits p66 and p51.

Chapter 4: Discussion

Discussion

This thesis work is devoted to understanding protein structure and structure-function relationships using the protein contact networks. First we analysed 51 proteins from each of the four different structural classes all α , all β, α+β and α/β. This study views each protein structure as a network of convalent and noncovalent connections between amino acids. Each amino acid in a protein structure is a node, and the strength of the interactions which is reflected in the distance between two amino acids is evaluated for edge determination. The protein contact network for 204 proteins were constructed as a function of the cutoff distance. It is clearly observed that the contact maps are different for the different structural classes. The network parameters and their combinations can be used to describe a structure and for delineating the differences among different protein classes. The Kolmogorov-Smirnov test showed that some network parameters of the α and β classes are significantly different thereby establishing the same. All of the four structural classes of protein were visualized in different network layout (2-D contact map, kamada kawai layout and ring graph) for better understanding of the structural features among the four classes. We also showed that, irrespective of structural classes (all α, all β, α+β and α/β), all proteins show **small world** behavior. This implies that the proteins are neither randomly structured, nor have fully regular topology.

After understanding the PCN for these structures, we chose HIV-1 Reverse Transcriptase as our model protein to observe the protein-ligand interaction using the approach of contact map. We did the analysis with the apo and holo form of the protein in order to observe structure-function relationship and to show the conformation changes between the two forms using PCN approach.

The two functionally important subunits of the HIV-1 RT p66 and p51 were also studied using contact map. The interactions between the residues of these two subunits were found to be changing upon ligand binding (Holo form). We clearly observe the change in contact area of two subunits because of the binding of ligand, which implies that ligand binding brings about conformational change in the interaction of two subunits p66 and p51. The contacts between amino acid residues are lost and new contacts are formed. Even though many contacts change in the whole network, but here we focused only on residues in the ligand binding pocket. We observed many common contacts in 1IKW and 1FKO which were lost when protein was in holo form (ligand bound). Similarly, in both of these protein models (1IKW and 1FKO) some new contacts were observed which were not present in the apo form (1RTJ).
The PCN approach allows easy visualization of these changes in large proteins like the HIV-1 RT. Given that the cross structural RMSD for these proteins are low, the PCN approach offers an alternative way of studying the confromational changes. The results can be supplemented and validated with molecular dynamics simulation.

Bibliography

1. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures J. Mol. Biol. (1995).

2. D. J. Watts, Small Worlds The Dynamics of Networks Between Order and Randomness, Princeton University Press (1999).

3. A.-L. Barab si, R. Albert, Emergence of Scaling in Random Networks Science (1999).

4. R. Albert, A.-L. Barabasi, Statistical Mechanics of complex networks Rev. Mod. Phys. (2002).

5. G. Pollastri and P Baldi Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation form all four cardinal corners, bioinformatics (2002).

6. Konrad Hinsen, Elastic Network Models for proteins,
http://dirac.cnrs-orleans.fr/plone/Members/hinsen/elastic-network-models-for-proteins.

7. R. Albert, H. G. Othmer, The topology of the regulatory interactions predicts the expression pattern of the Drosophila segment polarity genes. J. Theor. Biol. (2003).

8. E. Ravsaz, A. L. Somera, D. A. Mongru, Z. N. Oltvai,A.-L. Barabsi, Hierarchical organization of modularity in metabolic networks Science (2002).

9. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A.-L.Barabasi. Lethality and centrality in protein networks, Nature (2000).

10. Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros and Ruth Nussinov, Residues crucial for maintaining short paths in network communication mediate signaling in proteins, Molecular Systems Biology (2006).

11. A. Aszodi, W. R. Taylor, Connection topology of proteins, CABIOS (1993).

12. L. H. Greene, V. A. Higman, Uncovering network systems within protein structures J. Mol. Biol. (2003).

13. Rambaut, A; Posada, D; Crandall, KA; Holmes. The causes and consequences of HIV evolution. Nature Reviews Genetics (2004).

14. Di Marzo Veronese F, Copeland TD, DeVico AL, Rahman R, Oroszlan S, Gallo RC, Sarngadharan MG. Characterization of highly immunogenic p66/p51 as the reverse transcriptase of HTLV-III/LAV. Science (1986).

15. Lowe DM, Aitken A, Bradley C, Darby GK, Larder BA, Powell KL, Purifoy DJ, Tisdale M, Stammers DK. HIV-1 reverse transcriptase: crystallization and analysis of domain structure by limited proteolysis. Biochemistry (1988).

16. Kohlstaedt LA, Wang J, Friedman JM, Rice PA, Steitz TA. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science (1992).

17. Jacobo-Molina A, Ding J, Nanni RG, Clark AD Jr, Lu X, Tantillo C, Williams RL, Kamer G, Ferris AL, Clark P, et al. Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double stranded DNA at 3.0 A resolution shows bent DNA. Proc Natl Acad Sci USA (1993).

18. Xiong Y, Eickbush TH. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J (1990).

19. Larder BA, Purifoy DJ, Powell KL, Darby G. Site-specific mutagenesis of AIDS virus reverse transcriptase. Nature (1987).

20. Delarue M, Poch O, Tordo N, Moras D, Argos P. An attempt to unify the structure of polymerases. Protein Eng (1990).

21. Huang H, Chopra R, Verdine GL, Harrison SC. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. Science (1998).

22. Martin-Hernandez AM, Domingo E, Menendez-Arias L. Human immunodeficiency virus type 1 reverse transcriptase: role of Tyr115 in deoxynucleotide binding and misinsertion-delity of DNA synthesis. Embo J (1996).

23. Ding J, Das K, Hsiou Y, Sarafianos SG, Clark AD Jr, Jacobo-Molina A, Tantillo C, Hughes SH, Arnold E. Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 A resolution. J Mol Biol (1998).

24. Ren J, Esnouf R, Garman E, Somers D, Ross C, Kirby I, Keeling J, Darby G, Jones Y, Stuart D, et al. High resolution structures of HIV-1 RT from four RT-inhibitor complexes. Nat Struct Biol (1995).

25. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourneand, W. R. Taylor, The Protein Data Bank Nucleic Acids Res. (2000).

26. MATLAB Manual and Introductory tutorials, Ivan Graham (2005).

27. An Introduction to R :Notes on R: A Programming Environment for Data Analysis and Graphics Version 3.0.0 (2013-04-03).

28. PyMOL User's Guide, Warren L. DeLano and Sarina Bromberg (2004).

29. Pajek Program for analysis and visualization of large networks, Reference Manual version2.05, Vladmir Batagelj and Andrej Mvar.

30. NWB Team. (2006). Network Workbench Tool. Indiana University, Northeastern University, and University of Michigan,
https://nwb.slis.indiana.edu/community/?n=VisualizeData.ForceDirected

31. Duncan J. Watts and Steven H Strogatz, Collective dynamics of small world networks, Nature 393, 440-442 (1998).

32. Stefan G. Sarafianos, Bruno Marchand, Kalyan Das, Daniel Himmel, Michael A. Parniak, Stephen H. Hughes4, and Eddy Arnold Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. J Mol Biol. (2009).