5-2020

Network Analysis of Protein Structure Networks Upon Ligand Binding

David Foutch
University of Tennessee

To the Graduate Council:

I am submitting herewith a thesis written by David Foutch entitled "Network Analysis of Protein Structure Networks Upon Ligand Binding." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Life Sciences.

Albrecht von Arnim, Major Professor

We have read this thesis and recommend its acceptance:

Tongye Shen, Rachel McCord

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Network Analysis of Protein Structure Networks Upon Ligand Binding

A Thesis Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

David Alan Foutch

May 2020

To my wife and son, Sarah and Micah.

Acknowledgments

I would like to thank the members of my committee and to GST for this opportunity for which I am very grateful. I owe a debt of gratitude to Tongye Shen for facilitating one of the most satisfying endeavors of my life and for his encouragement.

I would like to acknowledge the love and support of my wife, Sarah, and my beautiful son, Micah.

I could not have completed this project without the love and support of my family: Atdaddy, Grandaddy, and Grammy–Odell Binkley, and Jeff and Ann Smith.

Abstract

Network analysis is a computational approach used to describe the structure and dynamics of complex systems. Residue-residue contacts that are made over the course of MD simulations were used to create protein structure networks (PSNs). As a case study, PSNs were generated for two protein systems: the transcription factor constitutive androstane receptor and the enzyme ribonucleotide reductase. In order to understand the changes in residue-residue contacts induced upon ligand-binding in proteins, we performed topological analyses of three CAR systems and four RNR systems under different binding conditions. Four measures of centrality were used to evaluate structural changes between ligand-free and ligand-bound systems: betweenness, closeness, degree, and eigenvector centralities. Although ligand-binding induced contact rearrangements resulting in substantial changes in centrality values for many residues, the distributions of centrality values were generally very similar for all systems. Results obtained here suggested that closeness centrality primarily identifies residues that are physically central to the three-dimensional structure of the protein. Previous reports suggested that closeness centrality identifies important residues in enzyme active sites. However, this may only be true for enzymes whose active site is centrally located. Moreover, the distributions for degree centrality are not power-law distributed, which also raises the question of whether the power-law degree distribution should be assumed for all "real-world" networks. In summary, this work demonstrated that the centrality distributions for the two representative proteins are remarkably invariant to ligand binding, despite substantial changes in centrality values for residues.

,

Table of Contents

1 Introduction **1**

1.1 Background 1

1.2 Centrality Analysis of Protein Structure Networks 3

2 Graph Theory as a Method of Analysis **6**

2.1 Elements of Graph Theory 7

2.2 Measures of centrality 10

2.3 Graph Theory That's Easier on the Eyes 12

2.4 Three Classical Random Graphs 16

3 Identifying Features in Protein Structure Networks by Measures of Centrality **22**

3.1 CAR and RNR proteins 22

3.2 Preparing the PSNs for Analysis 24

3.3 Closeness Centrality: A cautionary tale 36

3.4 Degree Centrality: "Real-worlds" and Scale-free Topology 43

3.5 Betweenness Centrality: Information Highways or the Paths Less Travelled? 46

3.6 Eigenvector Centrality: The Twilight Zone of Centralities 52

3.7 Discussion and Conclusions 57

4 Summary and Future Work **59**

4.1 Summary 59

4.2 Future Work 60

Bibliography **61**

Vita **69**

List of Tables

3.1 Mean and annotated mean contact maps 27
3.2 α-carbon table extracted from the PDB files 30
3.3 An example of the PSN edge list with xyz-coordinates for Cα atoms 32

List of Figures

1.1 Illustration of centrality distributions and set intersections for PSNs 5

2.1 An illustration of nodes related by edges. **2.1(a)** Two nodes are joined by a simple unweighted, undirected edge. **2.1(b)** Two nodes are joined by directed, weighted edges. 13

2.2 Network representation of the graph object $G(V, E)$ [68]. 14

2.3 Matrix representation of the graph object $G(V, E)$. 15

2.4 Node position and network centrality 17

3.1 Distributions of mean contacts for CAR protein systems 28

3.2 Distributions of mean contacts for RNR protein systems 29

3.3 CIT-hCAR PSN: Ribbon and Cα 31

3.4 Centrality distributions for CAR 34

3.5 Centrality distributions for RNR 35

3.6 CC values mapped to CAR molecular structure 37

3.7 CC values mapped to RNR molecular structure 38

3.8 CC distributions and scatter plots for CAR and RNR 39

3.9 Venn diagrams of top 16% of residues for mCAR CC 41

3.10 Venn diagrams of top 16% of residues for RNR CC 42

3.11 DC distributions and scatter plots for CAR and RNR 45

3.12 Venn diagrams of top 16% of residues for mCAR DC 47

3.13 Venn diagrams of top 16% of residues for RNR DC 48

3.14 BC distributions and scatter plots for CAR and RNR 49

3.16 Venn diagrams of top 16% of residues for mCAR BC 50

3.15 Venn diagrams of top 16% of residues for RNR BC 51
3.17 EC distributions and scatter plots for CAR and RNR 53
3.18 Venn diagrams of top 16% of residues for mCAR EC 55
3.19 Venn diagrams of top 16% of residues for RNR EC 56

Chapter 1

Introduction

Allostery is a process whereby ligand-binding at one site exerts a regulatory effect on the functional activity at the active site. Although over 1900 allosteric proteins have been studied over a 50 year period [53, 28], our understanding of the mechanisms underlying site-to-site communication across protein structure remains limited. We will investigate the effectiveness of four measures of network centrality to discriminate between the protein structure networks (PSNs) [42, 50, 19, 34] of three effector binding proteins under different effector binding states, and determine which if any of these measures are sensitive to contact rearrangements upon effector binding. We will use two representative protein systems: unbound and ligand-bound murine constitutive androstane receptor (mCAR); ligand-bound human constitutive androstane receptor (hCAR); and unbound and ligand-bound ribonucleotide reductase proteins (RNR), which may inform our understanding of residue-residue contact rearrangements due to effector binding. If successful, this work may be extended to the consideration of protein dynamics networks (PDNs) [41, 21, 11].

1.1 Background

Graph theory is a branch of mathematics that studies the theoretical properties of graphs. Network analysis is the application of graph theory to "real-world" systems. In the proposed research, we will investigate the relationship between distributions of network centrality and their potential sensitivity to residue-residue contact rearrangements between ligand-free and

ligand-bound states. In this work, the PSN is equivalent to the mean-contact map and the PDN is equivalent to the principal components of the co-variance matrix. The mean-contact map is a matrix in which the elements denote the pairwise mean contact between residues. A contact is a binary value defined in terms of Euclidean distance. When any two atoms between a pair of residues fall within a Euclidean distance, then a value of 1 is assigned, other wise 0. The principal components are calculated from the co-variance matrix and itself is a matrix denoting correlated motion (synchronous and asynchronous making and breaking of contacts) between residues. Both the PSN (constructed from contact maps) and the PDN (constructed from the principal components) are generated using the CAMERRA software [29]. In these networks, the residues are represented as "nodes" and pairwise contacts between residues are represented as "edges". Taken together the distribution of edges among the nodes defines the structure, or topology, of the network. The edge distribution can allow for identifying nodes that are central to the network structure. A number of centrality measures have been defined to identify potentially important nodes based on their centrality, or position in the structure of the network. For this reason, all centrality measures are positional measures. Betweenness centrality (BC) and closeness centrality (CC) are defined in terms of shortest path lengths whereas degree and eigenvector centrality are based on adjacency. The degree centrality (DC) of a node is simply the number of edges connected, or adjacent, to that node. Eigenvector centrality (EC) weights nodes that scales with the degree of adjacent nodes. A path is any sequence of nodes and edges that connects a pair of nodes. Shortest paths are simply the shortest routes connecting two nodes. It is well-established that centrality measures along with shortest paths identify essential nodes or vertices that effect the behavior of the network under perturbation [46, 36]. In the proposed research, **we investigate whether these measures of centrality are sensitive to residue-residue contact rearrangements between ligand-free and ligand-bound states.**

PSNs have gained the interest of researchers engaged in the study and prediction of allosteric mechanisms [19, 2, 50, 42]. It is generally accepted that the complementarity between protein structure and function is a result of evolutionary selection. Therefore, it stands to reason that the topology of the PSN would reflect something of these conserved relationships as well [13, 26, 49, 52, 14, 45]. CC has been used in combination with solvent

accessibility to identify catalytic sites [3]. Changes in both CC and BC were reported along with solvent accessibility under six different temperatures for β-lactamase inhibitory protein. [31]. This approach provides interesting insights on how protein conformation and the subsequent PSN are affected by the surrounding environment. A number of authors report an association between CC and active site residues [30, 55, 17]. Recently a new model of preferential attachment constructed around a weighted-betweenness coefficient was reported [62]. Unlike the BA model that prioritizes attachment to DC, the authors report that the weighted betweenness preferential attachment model reproduces features of real-world networks including the power-law degree distribution. In the search for "channels of transmission" and "energy transport pathways" in PSNs [39, 64], the DC has been associated with correlated allosteric motions both within and between clusters of residues [16]. Altogether, a strong negative correlation has been reported between BC, CC, DC and evolutionary rates of residues [22]. EV has been used to identify protein-ligand interaction effects following ligand binding at allosteric sites[39, 21]. Whereas these measures have been studied both independently and collectively; however, to our knowledge an investigation of these four centrality measures to discriminate between ligand-free and ligand-bound states has not been reported. In this research, we compare four measures of centrality from the PSNs of our seven representative protein systems. Specifically, **we will investigate how potential changes in shifts in PSN topology based on residue-residue contact rearrangements following ligand-binding correlates with these network centralities.**

1.2 Centrality Analysis of Protein Structure Networks

This works evaluates the sensitivity of four measures of network centrality to potential contact rearrangements upon ligand-binding. The fundamental question at hand is whether network analysis is sensitive enough to detect ligand-binding induced residue-residue contact rearrangements. After constructing PSNs from contact maps for our representative protein systems, we will generate betweenness, closeness, degree, and eigenvector centrality distributions for each PSN. Next, we will use the four distributions for each PSN collectively

to "profile" ligand-free and ligand-bound protein systems. In effect, we are treating the four centrality measures of ligand-free systems as a baseline signal to which the "profiles" of ligand-bound systems may be compared. Any differences between ligand-free and ligand-bound profiles is interpreted as potentially attributable to ligand-induced alterations to PSN topology. Additionally, we will investigate the intersection of the top percentage of residues for each centrality measure for ligand-free and ligand-bound systems. This will be performed without regard for the specific order of the residues. Taken together, both the centrality distributions and the intersections will be treated as measures of the degree to which contact rearrangements have resulted from ligand-binding. From these two observations–distribution shift and centrality intersections–we anticipate four possible outcomes.

First, the ligand-binding event creates no shift in the ensemble of centrality distributions and the intersection of the top residues by centrality remains approximately the same. The implication here is 1) the centrality analysis is not sensitive enough to detect the changes that are occurring, i.e., network analysis is too coarse grained, 2) there are no substantial changes in the topology of the PSN, or 3) the important rearrangements are subtle and do not significantly perturb the PSN topology 1.1a.

Second, the ligand-binding event creates no shift in the ensemble of centrality distributions, but the intersection of the top residues by centrality is reduced. In this case, the implication is that whatever contact rearrangements may have occurred the distributions do not reflect substantial changes in the topology of the PSN; however, reorganization has occurred as indicated by the depleted intersection of top residues by centrality 1.1b.

Third, the ligand-binding event shifts the ensemble of centrality distributions, but the intersection of the top residues by centrality remains approximately the same. In this case, the implication is that significant contact rearrangements have occurred as evidenced by the shift in the centrality distributions; however, the core of the PSN topology is unperturbed. This result could occur from either the loss or gain of inconsequential edges 1.1c.

Fourth, the ligand-binding event shifts the ensemble of centrality distributions, and the intersection of the top residues by centrality is reduced. In this case, the implication is that significant contact rearrangements have occurred and this has resulted in a reorganization of

the PSN topology as evidenced by the shift in the centrality distributions and the reduced intersections 1.1d.

Quantifying what constitutes a significant change in network topology is still a topic of research in network science [37, 65]. However, if this method is successful, it may potentially demonstrate the effectiveness of a novel approach using network centrality to computationally discriminate between ligand-free and ligand-bound proteins. It may even provide a for means for "profiling" other "omics" systems.

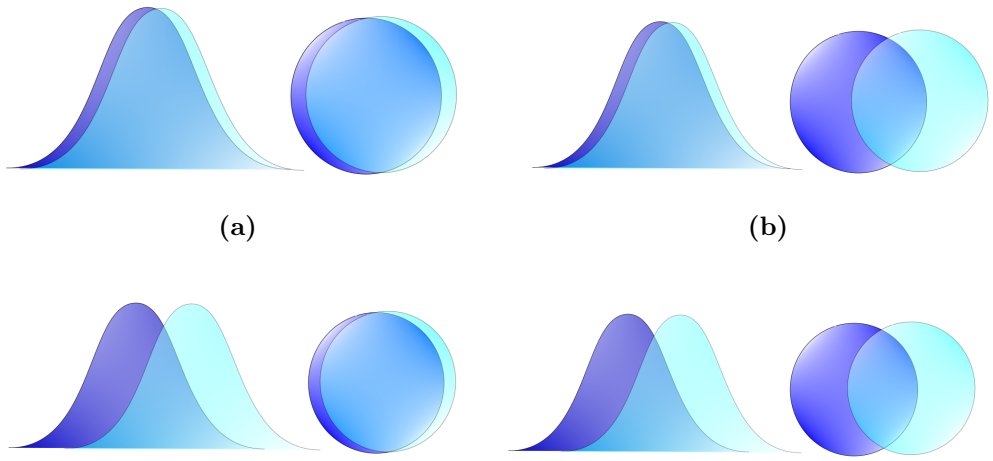


Figure 1.1: Illustration of centrality distributions and set intersections for PSNs. The figures shown above illustrate four potential outcomes that may describe the similarity between the ligand-free and ligand-bound PSNs. **(a)** No shift in the centrality distribution or reduction in the intersection, **(b)** No shift in the centrality distribution, but reduction in the intersection, **(c)** Shift in the centrality distribution, but no reduction in the intersection, and **(d)** Shift in the centrality distribution and reduction in the intersection

Chapter 2

Graph Theory as a Method of Analysis

The network paradigm is used increasingly to describe the structure and dynamics of complex systems [6, 57, 38, 7]. Researchers are discovering that many biological phenomena are not the product of single, isolated components, but derive from a complex system of interactions [60, 48]. For these reasons, a growing number of researchers are turning to the methods of systems biology to address questions that may only be investigated at the level of the network. An integral part of systems biology is the implementations of mathematical models to understand the dynamics and behaviors of complex molecular, cellular, and biological systems [9]. It approaches complex networks as a whole and investigates the control and development of emergent phenomena at the level of the network [8]. Understanding how the components of complex networks are "wired together" is critical to understanding the emergent behavior of the system. To that end, graph theory provides a useful mathematical formalism for representing and analyzing the structure and interactions of complex systems. Here we review some of the fundamental concepts of graph theory and their application to proteins as molecular networks [24, 58, 20].

2.1 Elements of Graph Theory

A graph G is composed of two sets of elements, a set of vertices and a set of edges, which we denote as $G = (V, E)$. The vertex set $V(G)$ is composed of a set of elements called **vertices** or **nodes**. The edge set $E(G)$ is composed of unordered pairs of vertices. (In the case of a directed network the pairs of vertices are ordered.) Given that two vertices v_i , $v_j \in V(G)$ are connected by an edge in G, then the unordered pair $\langle v_i, v_j \rangle \in E(G)$. Vertices that are connected by an edge in a graph are said to be **adjacent** and the edge connecting two vertices is said to be **incident**. The total number of vertices adjacent to a vertex v_i is called the vertex **degree** and is denoted as δ_{v_i}. There are three coefficients that are used to describe simple random graphs: The number of vertices in the graph V, the number of edges E, and the average vertex degree denoted $\langle \delta \rangle$ also called the coordination number z.

Definition 1: For all random, undirected graphs G composed of elements $e_{ij} \in E(G)$ joining an unordered pair of elements $v_i, vj \in V(G)$, there exists an element among the set of edges such that $e_{ij} = \langle v_i, v_j \rangle$. In which case, edge e_{ij} is said to be incident on vertices v_i, v_j and vertices v_i, v_j are said to be adjacent.

For a random graph G the set V is used to denote the **order** of G and the set E is used to denote the **size** of G. A complete graph is defined as the graph that is induced by $V(G)$ such that

$$\forall\, v_i, v_j \in V(G) \,\exists\, e \in E(G)\, s.t.\, e = \langle v_i, v_j \rangle. \tag{2.1}$$

Consequently, for all pairs of vertex elements in the complete, undirected graph where self-edges are excluded it follows that the size of G is

$$\forall\, v_i \in V(G), \quad \delta_i = N - 1 \implies E(G) = \binom{N}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2}, \tag{2.2}$$

where N denotes the number of vertices. This implies that $E(G)$ corresponds to the dimensions of either the upper or lower triangle of the adjacency matrix. It follows that

all non-complete graphs are sub-graphs of complete graphs, the density of which is defined by $D = {}^{E(G)}/_{\frac{n(n-1)}{2}}$.

Definition 2: Let S be a graph of G induced by the vertex set $V^*(S) \subset V(G)$ such that $E^*(S) \subset E(G)$, and subsequently we write that the graph $S(V^*, E^*) \subset G(V, E)$. Note: $E^*(S) \subset E(G) \implies \exists\ e_{ij} \in E(G)\, s.t.\ e_{ij} \notin E^*(S)$.

The set of vertices adjacent to a given vertex v_i is also called the **neighbor set** of v_i denoted $N(v_i)$. Note that the size of the neighbor set is equivalent to the vertex degree: $N(v_i) = \delta_{v_i}$. Let $G(V, E)$ be random graph, then $N(v_i) = \{V^* : V^* \subset V(G)\}$. The corollary to $N(v_i)$ is the set of edges incident to v_i denoted $I(v_i)$ where $I(v_i) = \{E^* : E^* \subset E(G)\}$. There are three core topological coefficients associated with the neighbor set: The vertex degree, the degree of graph G and the average vertex degree for G. The degree for v_i is denoted

$$\delta_{v_i} = \text{card}\big[N(v_i)\big], \tag{2.3}$$

where $\text{card}\big[\,\big]$ denotes the count of the number of elements in set $N(v_i)$. From the vertex degree we can determine the size or degree of graph G

$$\Delta(G) = \frac{1}{2}\sum_{i=1}^{N} k_i \in V(G). \tag{2.4}$$

And from the degree of the random graph G we can determine the average vertex degree, or the **coordination number** of G

$$\langle \delta \rangle = z = \frac{\Delta(G)}{\text{card}\big[V(G)\big]}. \tag{2.5}$$

Definition 3: Let $|N|$ equal to the cardinality of the neighbor set $N(v_i)$ such that $|N| = card\big[N(v_i)\big]$ and $n = \{n \in \mathbb{N} | 1 \leq n \leq \|N|\}$. Then let V^* be a subset of $V(G)$ and E^* be a subset of $E(G)$ such that $S(V^*) \subset V(G)$ and $S(E^*) \subset E(G)$. Therefore, if we let $I(v_i) = e_{i1}, ..., e_{in}, ..., e_{i|N|} \in E^* \subset E(G) | \forall e_{ij} \in E^* \exists \langle v_i, v_j \rangle \in V^* \subset V(G) \implies S(V^*, E^*) \subset G$.

A **path** is a sequence of edges that connect two vertices as endpoints. Let $P(G)$ denote the set of all possible paths connecting all pairwise combinations of vertices in $V(G)$, and we denote the **path length** $d()$ as a distance metric defined as a count of the edges joining the vertex endpoints. Consider a set P_{ij}, where $P_{ij} \subset P(G)$ denotes all possible paths joining vertices v_i, v_j, and $p_{ij} \in P_{ij}$. Next, We let $|m| = \text{card}\big[P_{ij}\big]$ and $n = \{n \in \mathbb{N} | 1 \leq n \leq |p|\}$. Therefore, $P_{ij} = \{p_{ij}^1, ..., p_{ij}^n, ..., p_{ij}^{|m|}\}$. Given that for all p^n there exists a corresponding distance d^n, then in the case where $d^1(v_i, v_j) \leq ... \leq d^n(v_i, v_j) \leq ... \leq d^{|m|}(v_i, v_j)$ we find that there exists a **shortest path** such that

$$P_s(v_i, v_j) = \min_{i,j} p_{ij} \in P_{ij}. \tag{2.6}$$

Determining the shortest paths that connect all pairwise combinations of vertices naturally extends to the idea of the **average shortest path** as a characterizing topological coefficient for any random graph G, which we define as

$$\langle P_s \rangle = \frac{1}{w} \sum_{i=1}^{w} d^i \tag{2.7}$$

where $w = \frac{N(N-1)}{2}$ and N denotes the number of vertices in $V(G)$. This particular measure of distance allows researchers to identify vertices in large graphs that function as hubs or intermediaries between hubs. **Hubs** are high degree nodes that are strongly connected. In molecular signaling and regulatory networks, the assumption is that these vertices will be highly conserved. Short average path lengths are important for quickly routing information across a large number of targets.

2.2 Measures of centrality

Measures of centrality are more associated with network analysis than graph theory *per se*. In other words, any text book doing network analysis will include discussions on centrality along side graph theory, but in all likelihood a text book that is strictly graph theoretical will not include discussions of the measures of centrality that are considered here. In fact, the four core measures of centrality were not developed in mathematics, physics, or any of the 'hard sciences', but rather by social scientists researching social structures. In fact, Professor of sociology and behavioral science LC Freeman has famously said that "All four of the centrality measures [degree, closeness, betweenness, eigenvector centrality] from social network analysis had moved—the wrong way—into both physics and biology" which reflects an awareness that historically "physics and other, longer established fields like biology and chemistry, often fertilized the social sciences and never the other way round" [68]. This statement is underscored by an absence of rigorous treatment of centrality in graph theoretical text books. Perhaps this is because centrality measures are defined in an axiom independent manner. In other words, the definitions are independent of theoretical consideration. In a sense, they are *ad hoc*. Regardless, there is a growing body of literature on the efficiency of algorithms used for evaluating centrality for very large networks. Given that the focus of this work is the application of graph theory and network science to protein structures, this work will forego a review of the algorithms that implement centrality analyses. However, for the reasons briefly discussed above, there will also be a striking absence of graph theoretical considerations for grounding the centrality measures reviewed [40, 25].

Betweenness Centrality (BC) is a measure that utilizes the definition of shortest path to assign significance to vertices. In descriptive terms, the BC of a vertex v_i is the ratio of the number of shortest paths incident on v_i to the total number of shortest path lengths in the network. Therefore, the BC evaluated for any vertex v_i is equal to the proportion

$$B(v_i) = \frac{\sum s_i}{S(G)}, \tag{2.8}$$

where s_i denotes the set of shortest paths that include v_i and $S(G)$ denotes the complete set of shortest paths for a random graph G. Vertices with the large BC are understood to be important for maintaining the global connectivity of the graph. Vertices with larger BC also decrease the overall average path length and diameter of the graph. BC also contributes to the rate of diffusion on graphs. Vertices with large BC increase the rate of diffusion across the graph due to reduced average path lengths.

Closeness Centrality (CC) is a measure that identifies vertices that have the shortest paths to all other nodes. In other words, these are the vertices that are at the geodesic "center" of the graph. CC is defined as

$$C(v_i) = \frac{1}{\sum_j^N d(v_i, v_j)}, \tag{2.9}$$

where $d(v_i, v_j)$ is the distance between v_i and all other vertices j in G. This definition of CC implies that the largest possible CC for v_i in G is $C(v_i) = \frac{1}{N}$ and $C(v_i) \to 0$ for vertices at the graph periphery. Note: A vertex that is close to the center of a graph will have a CC approximately equal to the radius of the graph. Given the definition of CC, the question arises as to whether BC and CC are somewhat redundant. It can be the case that nodes that are 'close' are also 'between', but that is not necessarily the case. But, the principle difference between CC and BC are at the end points of the distances measured. BC evaluates the shortest path from v_h through v_i to v_j and CC begins at v_i and evaluates the distance to every other vertex v_j.

Degree Centrality (DC) is simplest of the centrality measures to understand intuitively. DC is just the count of all the nodes adjacent to v_i or alternatively a count of the edges incident on v_i. Using the definition of counting incident edges, we have

$$D(v_i) = \sum_j^N e_{ij} \tag{2.10}$$

It does not require much imagination to rationalize the logic of degree centrality: Things that are important tend to be well-connected. Whether it's social groups, interstate traffic, or biological pathways, things that are important exert a lot of influence over the system of interest. So, whereas BC and CC centrality communicate influence, vertices with high degree

centrality exert influence. Additionally, the topology of graphs with with scale-free degree distribution cohere through the existence of a small number of highly connected nodes. Targeted removal of these vertices can dramatically effect network connectivity resulting in the decomposition of large components smaller components or even bifurcation of the network.

Eigenvector Centrality (EC) builds on the concept of DC. If the vertices that are the most well-connected (highest ranked by degree centrality) exert the most influence on the network, then vertices immediately adjacent to these are likely vertices that are also "influential". EC is a concept akin to guilt-by-association. The EC of a vertex v_i is proportional to

$$E(v_i) = \frac{1}{\lambda} \sum_{j=1}^{N} A_{i,j} v_j, \tag{2.11}$$

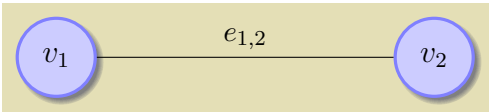
which satisfies $Av = \lambda v$. The "significance" of node v_i is defined by the eigenvector of the adjacency matrix A and scaled by the inverse of the associated eigenvalue λ. The entries of v are EC. The eigenvalue λ is the largest eigenvalue of the adjacency matrix A. One of the consequences of this centrality measure is that in scale-free networks EC 'drives' the centrality value of vertices that are not adjacent to high degree vertices to zero. In other words, EC for a vertex may be high because it has many neighbors or because it is adjacent to an influential neighbor. For this reason, in scale-free networks, EC tends to result in 'clusters' of many vertices of high EC centered on high-degree vertices with large swaths of vertices in the network characterized EC values near zero.

2.3 Graph Theory That's Easier on the Eyes

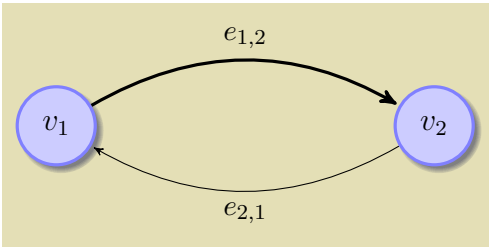
In principle, graph theory is a branch of mathematics that deals with abstract objects called *graphs*. However, in practice, the principles of graph theory when applied to networks can be understood as algorithmic operations on matrices. There is no exaggeration in saying that an adjacency matrix is the equivalent of a network. This equivalency holds because networks and matrices are merely a means of representing graph objects $G(V, E)$.

Adjacency matrices are composed of rows and columns which in practice are understood, respectively, as the sets of 'source' and 'target' nodes joined by the i, jth matrix elements. For undirected networks, an edge is incident on *unordered* pairs of vertices. That is $e_{i,j} = \langle v_i, v_j \rangle = \langle v_j, v_i \rangle = e_{j,i}$ and the corresponding adjacency matrix is symmetrical. For this reason, graph algorithms can be programmed to operate on either the upper or lower triangle. This can increase the run time efficiency on larges systems. Figure 2.1a is an example of two nodes joined by an unweighted, undirected edge. Conversely, for directed networks edges are incident on *ordered* pairs of vertices. That is $e_{i,j} = \langle v_i, v_j \rangle \neq \langle v_j, v_i \rangle = e_{j,i}$ and the corresponding adjacency matrix is asymmetrical. Figure 2.1b is an example of two nodes joined by directed, weighted edges. In adjacency matrices the presence of an edge is indicated by 1 and the absence of an edge by 0. For adjacency matrices with weighted edges, the presence of an edge will be represented by a scalar value.

To illustrate the relationship between the graph as an abstract mathematical object and its corresponding network and matrix representations consider the following example. Let $G(V, E)$ denote a graph object composed of the vertex set $V(G)$ and the edge set $E(G)$, where $V(G) = \{v_1, v_2, ..., v_{10}\}$ and $E(G) = \{e_{1,4}, e_{2,4}, e_{3,4}, e_{4,5}, e_{4,6}, e_{4,7}, e_{5,6}, e_{5,7}, e_{6,7}, e_{7,8}, e_{7,9}, e_{7,10}\}$. On inspection we find that $G(V, E)$ satisfies the formal condition that $\forall e_{i,j} \in E(G)\, \exists v_i, v_j \in V(G)\, s.t.\, e_{i,j} = \langle v_i, v_j \rangle$. Which simply means that for all edges in the set E there are



(a)



(b)

Figure 2.1: An illustration of nodes related by edges. **2.1(a)** Two nodes are joined by a simple unweighted, undirected edge. **2.1(b)** Two nodes are joined by directed, weighted edges.

corresponding vertices in V such that $e_{i,j}$ denotes the junction between two vertices v_i, v_j.

Now it is literally a game of connect-the-dots for simple graphs like $G(V, E)$. In this case, the dots are vertex elements of V and the lines that connect the dots are edge elements of E. For the graph object $G(V, E)$ a network representation is presented in 2.2. For this trivial case all the information that is present in the graph object in visually represented in the network. For more complex topology the network representation is frequently inadequate.

The 'ball-and-stick' network representation of complex data can be helpful for communicating how the structure and dynamics of a system are related, or for illustrating how centrality distributions characterize different systems. However, it is useless for calculating graph theoretical properties. Because a network represents the connectivity or interactions of a given system it is trivial to map that information to a matrix representation of $G(V, E)$. Once the matrix representation is created the entire domain of mathematical operations in linear algebra can be used to probe the system, e.g. calculating eigenvectors and eigen values. Also, other computational algorithms can be created to probe the system, e.g. calculating measures of centrality. A matrix representation of the $G(V, E)$ is shown in 2.3. Therefore,

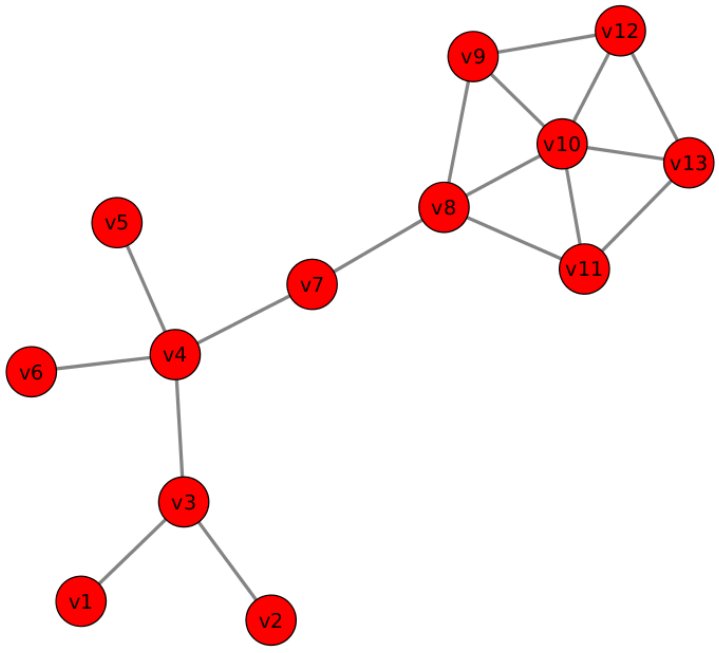


Figure 2.2: Network representation of the graph object $G(V, E)$ [68].

	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}
v_1	0	0	1	0	0	0	0	0	0	0	0	0	0
v_2	0	0	1	0	0	0	0	0	0	0	0	0	0
v_3	1	1	0	1	0	0	0	0	0	0	0	0	0
v_4	0	0	1	0	1	1	1	0	0	0	0	0	0
v_5	0	0	0	1	0	0	0	0	0	0	0	0	0
v_6	0	0	0	1	0	0	0	0	0	0	0	0	0
v_7	0	0	0	1	0	0	0	1	0	0	0	0	0
v_8	0	0	0	0	0	0	1	0	1	1	1	0	0
v_9	0	0	0	0	0	0	0	1	0	1	0	1	0
v_{10}	0	0	0	0	0	0	0	1	1	0	1	1	1
v_{11}	0	0	0	0	0	0	0	1	0	1	0	0	1
v_{12}	0	0	0	0	0	0	0	0	1	1	0	0	1
v_{13}	0	0	0	0	0	0	0	0	0	1	1	1	0

Figure 2.3: Matrix representation of the graph object $G(V, E)$.

it is trivial to note the equivalency between a network and a matrix representation of graph objects.

Because this matrix is symmetrical either the upper or lower triangle is sufficient for calculations. It should be noted that in the adjacency matrix the $e_{i,j}$ notation that has been used throughout the introduction is substituted with the scalar value 1. Formally, we write $e_{i,j} = \langle v_i, v_j \rangle = \sigma \in \mathbb{R}$. For a simple adjacency matrix, as we have in this example, all $e_{i,j} = 1$.

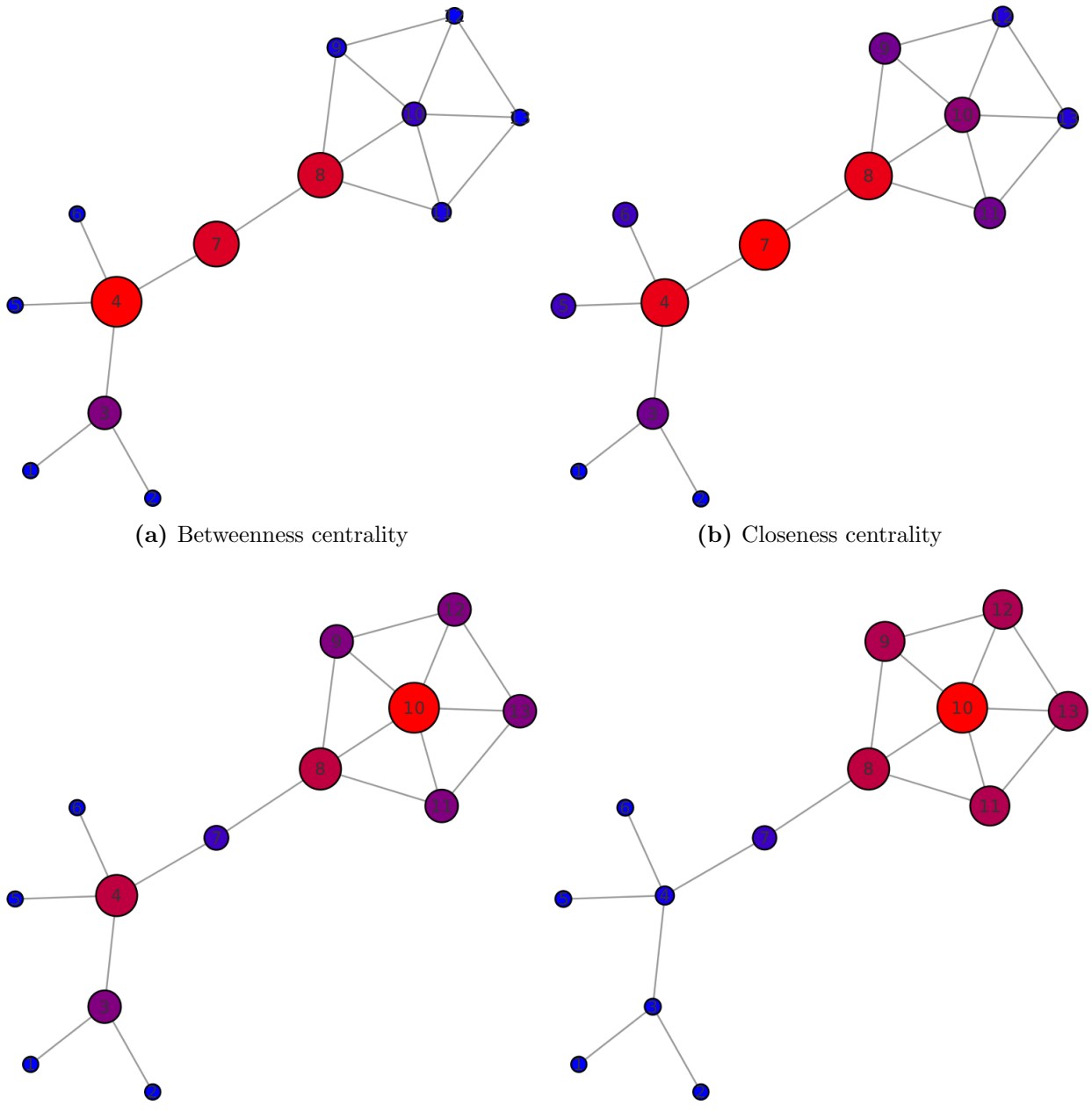
Centrality measures are *positional* measures of nodes in the network with respect to all the other nodes. Therefore, the manner in which a centrality value is assigned to nodes is determined by the algorithm used for counting edges in the network. The difference between betweenness, closeness, degree, and eigenvector centrality is in *how* the edges are counted. Figure 2.4 is a cartoon that illustrates the differences in centrality for nodes in a

simple network. In this cartoon, node size and color are used to indicate the differences in each of the four measures of centrality: Betweenness 2.4a, closeness 2.4b, degree 2.4c, and eigenvector 2.4d. A few brief observations. First, the three nodes that are highest ranked in both BC (nodes 4,7, and 8) and CC are the same. This is frequently the case as the shortest paths that connect nodes in a network will include nodes that are also geometrically central to the network. DC and EC are also very similar. This follows from the definition of EC which weights nodes by the degree of adjacent nodes. For example, consider the differences illustrated in nodes 4 and 8 shown in 2.4c and 2.4d. In figure 2.4(c), nodes 4 and 8 are of the same degree so they share the same size and color. However, figure 2.4(d) reflects the weight that is assigned to nodes based on the degree of adjacent nodes. The highest degree node adjacent to node 4 is node 3 with a degree of three, but node 8 is adjacent to node 10 which has a degree of 5.

2.4 Three Classical Random Graphs

There are many aspects that could be noted regarding the following random graphs. They have been the focus of theoretical research for the past 60 years. The intention of introducing these models is to provide background for their application to protein structure networks in chapter 4. Therefore, the following discussion will be restricted to only a few features that are characteristic of each model and relevant to making inferences about network topology from centrality distributions [40, 25].

The first random graph to have become a topic of research was introduced by **Paul Erdös** and **Alfréd Rényi** in 1959 and is recognized as the canonical example of a random graph, which bears their name as the **Erdös-Rényi random graph**, or **ER model**. In brief, the set of all ER random graphs is denoted by $ER(n, p)$ where n is the number of nodes and p is the probability that any two nodes will be joined by an edge. For an ER graph of size n and a connection probability p, the probability that the degree of a randomly selected node (v) will have exactly k neighbors out of a set of $n-1$ possible neighbors is given by [63]:



(a) Betweenness centrality

(b) Closeness centrality

(c) Degree centrality

(d) Eigenvector centrality

Figure 2.4: Node position and network centrality. Each centrality is a positional measure determined by an algorithm for counting edges. The same network is used in **(a)-(d)**. node size and color is varied to illustrate the differences in centrality. Smaller node size and blue are used to denote lower centrality values. Larger node sizes and red are used to denote higher centrality values.

$$P[\delta(v) = k] = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{2.12}$$

And, by extension, to find the average vertex degree for an ER graph is given by:

$$\begin{aligned} \sum_{k=1}^{N-1} k \cdot P[\delta(v) = k] &= \sum_{k=1}^{N-1} k \cdot \binom{n-1}{k} p^k (1-p)^{n-1-k} \\ &\vdots \\ &= p(N-1) \sum_{l=0}^{m} \binom{m}{l} p^l (1-p)^{m-l} \\ &= p(N-1) \cdot 1. \end{aligned} \tag{2.13}$$

This is an important observation for understanding clustering in ER random graphs and why ER random graphs fail to make good models for "real-world" networks. If we let z equal the average degree for an ER graph (see eq. 2.5), then from eq. 2.13 above we have $z = p(N-1)$ which implies $p = \frac{z}{N-1}$. Therefore, the implication is that clustering in ER graphs is very small. And for very large ER graphs the probability that any two vertices will be connected scales to zero, which carries the associated implication that the clustering goes to zero as well. Because the ER graph follows a binomial degree distribution with modest to zero clustering, it is frequently used as the null model for comparison to "real-world" networks which have been observed to be highly clustered even for very large networks. The significance of this is that if the ER model is a good fit for the network topology, then the topology is regarded as random.

Although the ER model has been around since 1959 it was not until 1998 that **Duncan Watts** and **Steven Strogatz** developed a method for adapting the ER model to include clustering. The algorithm is straightforward and is described in detail in any graph theory text. In short, the graph consists of n nodes arranged in a circular pattern. Nodes are joined to k neighbors, where k is an even integer value. Each node is joined to the first $\frac{k}{2}$ nodes immediately adjacent to the left and right. This is applied to all nodes. Second, each edge is rewired with a probability p and without edge duplication. The resulting graph is called a **Watts-Strogatz random graph** or **WS model** denoted by $WS(n, k, p)$. Although this procedure is somewhat straightforward the complicating aspect is that as edges are rewired

with a probability p the procedure follows the regime $n \gg k \gg 1$. This regime ensures that the resulting graph is sparse, preserves the "small-world" effect, and yields a degree of clustering. In other words, the algorithm still yields a random homogeneous network, but for large n the clustering will have a non-zero value. Without elaborating on the derivation, the clustering coefficient for WS graphs is computed to be

$$CC = \frac{3/8(k-2)}{1/2k(k-1)} = \frac{3}{4}\frac{k-2}{k-1}. \tag{2.14}$$

One of the more significant results that stems from this equation is that for any WS graph the constraint $n \gg k \gg 1$ fixes the ratio of the clustering coefficient to $\frac{3}{4}\frac{k-2}{k-1}$. From this fact several other implications follow. First, for even very large n the WS graphs will exhibit clustering, unlike ER graph. Second, the fixed ratio entails that no cluster will be complete, that is $CC \neq 1$. Third, the WS graph can be tuned by adjusting the parameter k when setting the constraint $n \gg k \gg 1$ in order to control for greater or lesser degree of clustering.

There are two key features to the third random model that make it attractive for comparison with "real-world" systems: Preferential attachment and evolutionary growth. In both the ER and WS models edges are assigned by a fixed attachment probability. In this sense, the attachment probabilities are fixed. However, the insight that is eventually formalized by **Albert-László Barabási** and **Réka Albert** (1999) is that real world networks evolve over time with the existing system at time t biasing the probability of edge assignment at time $t+1$. In other words, the incidence edge growth does not derive from a uniform probability, but exhibits *preferential attachment*. In the **Barabási and Albert random graph** or **BA model** the bias or preference is predicated by vertex degree at time t. At time $t+1$ there is a greater probability that favors vertices with greater degree. The preferential attachment implements an algorithm that biases edge assignment in favor of vertices that have more edges. For each iteration of the algorithm the probability of edge attachment for each vertex is re-evaluated. As the network evolves a *scale-free* or power-law degree distribution emerges. Networks that exhibit power-law degree distributions are dominated by many low-degree vertices that are connected by a few high-degree vertices,

also called "hubs". Several distinctive properties emerge from the preferential attachment of the BA model.

First, the BA model is tolerant, or robust, to vertex attack. Vertex attack strategies evaluate network vulnerability to loss of nodes and edges. One of the practical applications of this work is optimizing system designs that will function effectively when the structure is under attack. The scale-free property of the BA model exhibits robustness or tolerance to random vertex attack. The presence of many low-degree vertices entails an increased probability that a random attack will target a low-degree vertex. Scale-free networks can suffer the deletion of many vertices without compromising the global topology of the network structure. The probability that a random attack will hit a network hub and cripple the network is very low. It is in this sense that networks and biological systems modeled by networks are described as robust. For example, given a biological system represented by a scale-free network. The biological counterparts to network hubs are frequently presumed to have functional importance even if that function is presently unknown.

Second, Hubs in the BA model result in shorter average path lengths. In ER and WS models the uniform edge attachment probability yields homogeneous network structure. In contrast, the BA model employs an algorithm that biases the probability of edge assignment over time. After multiple iterations of the algorithm the emerging network structure creates an increasingly nonuniform probability for edge attachment yielding a heterogeneous or scale-free network structure. Underlying the heterogeneity is a number of high-degree vertices, or hubs. The system of hubs decreases the overall diameter of the network and decreases the number of edges that must be traversed to connect any two pairs of vertices. Because the BA model is characterized by shorter average path lengths, the speed with which a perturbation propagates across the system faster than in the ER and WS models.

Third, hubs associated with shorter average path lengths yield the emergence of node betweenness. It is possible to discuss BC in the context of arbitrary degree networks; however, BC takes on a significance in BA models that it does not have in the ER and WS models. In the basic versions of the ER and WS models there is likely to be a large overlap between vertex centrality and vertex BC simply because of the homogeneity of the network. However, in the BA model the scale-free degree distribution creates a heterogeneity that

reduces that average shortest paths. Among the set of all shortest paths between vertices, a set of vertices emerges for which many of these shortest paths pass through. These vertices will have a much higher BC coefficient. These will contribute to the robustness of the network to vertex attack as well as the speed with which information and perturbation propagates across the network. Because vertices with high BC coefficients emerge in scale-free networks there is a positive correlation between vertex degree and vertex betweenness.

The brief review of graph theory and canonical models provides the context for understanding how these concepts can be used to analyze network representations of biological systems. It provides a language for describing global network characteristics, like homogeneity and heterogeneity, as well as a means for quantifying network properties in order to make inferences and predictions about biological systems more rigorous.

Chapter 3

Identifying Features in Protein Structure Networks by Measures of Centrality

Beyond its formalism as a sub-branch of discrete mathematics, graph theory as an applied science provides a unique set of tools for evaluating complex systems. When dynamic and qualitative attributes are mapped to nodes and edges, network representations help researchers gain insight into how systems function and which components and interactions may be the most important. Proteins are molecular machines [1, 23] where structure and function are related. Therefore, it seems natural to approach the study of protein structure and dynamics with the tools available in network analysis with the goal of understanding protein structure and dynamics. This chapter will discuss (1) how protein structure networks (PSN) are constructed from protein simulation data, (2) how network analysis of PSN is used to characterize the underlying network topology, and (3) what that topology says about underlying biology.

3.1 CAR and RNR proteins

To illustrate the application of these methods, analysis of four measures of network centrality was performed on the PSNs of seven systems under different effector binding states. An

unbound and bound murine constitutive androstane receptor (mCAR) along with a ligand bound human CAR (hCAR), an ortholog of mCAR. Additionally, four human ribonucleotide reductase (RNR) systems were also studied under different effector binding states, including a ligand-free and three ligand-bound systems.

The proteins mCAR and hCAR are members of a broader receptor class called nuclear hormone receptors (NR). NRs are able to directly bind to DNA and regulate the expression of target genes. Generally, NRs must first undergo agonist dependent activation prior to functioning as a transcription regulator. CAR proteins appear to be an exception. CAR gene regulation occurs constitutively in the absence of an agonist. All NRs contain three conserved domains: An N-terminal domain, a DNA-binding domain (DBD), and a ligand-binding domain (LBD) [54, 43]. Binding an effector molecule at the LBD results in NR mediated gene expression. It has been noted that perhaps the most important consequence of effector binding at the LBD is the modified interaction between LBD and co-activators or co-repressors which consequently modulates the recruitment of RNA polymerase [54].

The computational design focused on the structural-mechanical effects of effector binding in the CAR monomers. This research reports on the analysis of two of these: 1,4-Bis[2-(3,5-dichloropyridyloxy)]benzene (TCPOBOP) and 6-(4-chlorophenyl)imidazo(2,1-b)(1,3)thiazole-5-carbaldehyde O-(3,4-dichlorobenzyl) oxime (CITCO). The ligand TCPOBOP acts as a strong agonist in mCAR. In this work, TCPOBOP bound at the LBD of mCAR is abbreviated as TCP-mCAR. The ligand CITCO acts as a strong agonist for hCAR [43]. In this work, CITCO bound at the LBD of hCAR is abbreviated as CIT-hCAR. This work explores how contact rearrangements from agonist binding events can be explained in graph theoretical terms specifically with the use of fundamental measures of centrality.

Ribonucleotide reductases are a highly conserved family of proteins that maintains the pool of deoxyribonucleotides used in the synthesis of DNA through the reduction of ribonucleotides [33, 27]. RNR oligomerization is essential to the function of RNR, and interface formation during oligomerization is modulated by effector binding. The activity of human RNR is regulated by two allosteric sites: An activity site (a-site)–not to be confused with the enzyme active site–and a specificity site (s-site) [61]. The a-site regulates overall enzymatic activity and the s-site determines the specific type of reduction reaction. As

mentioned previously this work investigates the PSN of RNR monomers under four different binding states. The abbreviations for these binding state follow an ASE-RNR format: "A" denotes the activity site, "S" the specificity site, and "E" the enzymatic site. The prefixes indicate whether an effector is bound at these sites. So, 000-RNR indicates the apo form of the human RNR protein, 0T0-RNR indicates that dTTP is bound at the s-site and the a-site is unliganded. AT0-RNR indicates that ATP is bound at the a-site and dTTP is bound at the s-site. dAT0-RNR indicates that dATP is bound at the a-site and dTTP is bound at the s-site [44]. Again, the simulation of protein dynamics focused on the human RNR protein. Analysis of both the CAR and RNR PSNs will investigate how centrality distributions and set intersections correspond to rearrangements in residue-residue contacts relative to the unbound or ligand-free systems.

3.2 Preparing the PSNs for Analysis

The PSN were produced from mean contact maps generated from protein simulation described in [29, 44]. In brief, a mean contact map represents the average "contact" between two residues over the time course of the simulation. A "contact" between two residues is said to occur whenever any atom between two residues is within a distance cutoff of $4.2\mathring{A}$. Over the course of the protein dynamics simulation residues will make and break contacts: Move inside and outside the Euclidean cutoff.

For example, consider a protein having n-residues that undergoes a dynamics simulation for a duration of time T during which S number of snapshots are taken and the xyz-coordinates for n-residues may be saved as vectors eq.3.1 such that

$$\begin{aligned} \vec{r}_1 &= \big[\,(x(t_i), y(t_i), z(t_i)) : 1 \leq i \leq S\,\big] \\ \vec{r}_2 &= \big[\,(x(t_i), y(t_i), z(t_i)) : 1 \leq i \leq S\,\big] \\ &\vdots \\ \vec{r}_n &= \big[\,(x(t_i), y(t_i), z(t_i)) : 1 \leq i \leq S\,\big] \end{aligned} \tag{3.1}$$

Here $\vec{r}_i$ is an oversimplification of position for residue i for demonstration purpose. If we let $I = 4.2\mathring{A}$ threshold and $|\vec{r}_i - \vec{r}_j|$ define the Euclidean distance between between two residues, then the contact u_{ij} is defined as

$$u_{ij} = \Theta(I - |\vec{r}_i - \vec{r}_j|) \tag{3.2}$$

where Θ is the Heaviside function eq.3.3 which, for a given snapshot, assigns 1 when a contact is observed and 0 when a contact is not observed

$$\Theta(x) = \begin{cases} 1 & \text{if } x > 0 \\ \\ 0 & \text{if } x \leq 0 \end{cases}. \tag{3.3}$$

And so we define the average or mean contact is defined as the ratio of actual contacts to total possible contacts

$$U_{ij} = \langle u_{ij} \rangle = \frac{1}{S} \sum_{S} u_{ij}. \tag{3.4}$$

This entails that the range of mean contact values is $1 \geq U_{ij} \geq 0$. Given the geometry of the protein some residues will never make contact and other residues will always be in contact. Their mean values will be 0 denoting no contact and 1 denoting constant contact. Therefore, we can define a contact matrix U such that

$$U = \begin{bmatrix} U_{11} & U_{12} & \cdots & U_{1n} \\ U_{21} & U_{22} & \cdots & U_{2n} \\ \vdots & & \ddots & \vdots \\ U_{n1} & U_{n2} & \cdots & U_{nn} \end{bmatrix}. \tag{3.5}$$

Because U is a square, symmetric matrix the computational complexity can be reduced by considering only the upper triangle as well as neglecting the self-interactions along the

diagonal. It is from this remaining set of $\frac{n(n-1)}{2}$ mean contacts that the PSN is derived. However, the upper-triangle of the matrix is easily converted to an 2xm edge list, where m is the selected contacts. In table 3.1a a portion of the mean contact map for CIT-hCAR is shown. The ordinal values in the first two columns of the mean contact map indicate the order of residues as they appear in the output of the contact map, e.g. first residue is denoted "1", the second residue denoted "2", etc. In the annotated mean contact map of table 3.1b the absolute index given in the PDB file for hCAR is used. The absolute index for hCAR is the ordinal index of the contact map + 106. As explained in chapter 2 networks and matrices are equivalent mechanisms for representing data. For the reasons mentioned previously, it is trivial to show that the mean contact map is merely the $\frac{n(n-1)}{2}$ equivalent of the upper triangle of the adjacency matrix. Therefore, the contact map constitutes the "edge list" of the PSN. Table 3.1b shows the annotated mean contact map for CIT-hCAR. Node attributes including a three-letter amino acid code and the internal index, are assigned to the contact map. This makes it convenient to identify residues and map residue positions to secondary protein structure.

From the foregoing, it is shown that a PSN is a network representation of the mean contact map, and the mean contact map is simply an edge list. The nodes in the PSN are residues and the edges are denoted by mean contacts. However, in the context of analysing protein dynamics, static edges and close to static edges are not of interest. Because static elements of the network do not contribute to understanding protein dynamics; therefore, they are trimmed from the PSN. Although interest is shown in determining cutoffs for contact invariant topology [47], our principle motivation in trimming the PSNs is driven by our long-term interest in building a protein dynamics network.

It is of interest to note that most edges aggregate at the min/max values of the mean contact distribution. Figure 3.1a shows a histogram of the distribution of mean contacts for CIT-hCAR. There are 661 edges aggregated at the min/max or static values of the distribution. When contacts $u_{ij} \leqslant 0.02$ and $u_{ij} \geqslant 0.98$ are cut this results in a PSN network with 550 edges, which is just less than half of the total edges reported for the CIT-hCAR PSN. This method of visually identifying static edges was used to trim the PSN for TCP-mCAR, 000-RNR, 0T0-RNR, AT0-RNR, and dAT0-RNR proteins (figure 3.1b).

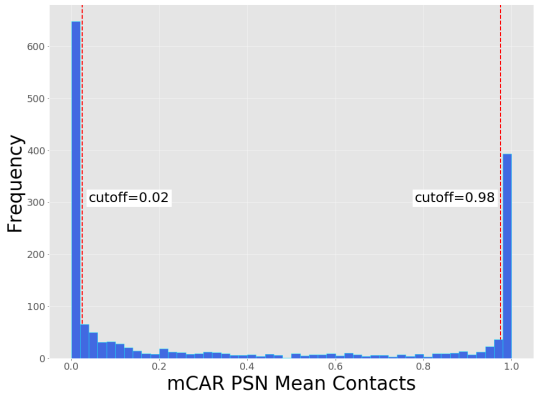
Table 3.1: Mean and annotated mean contact maps illustrated using CIT-hCAR

(a) Mean contact map

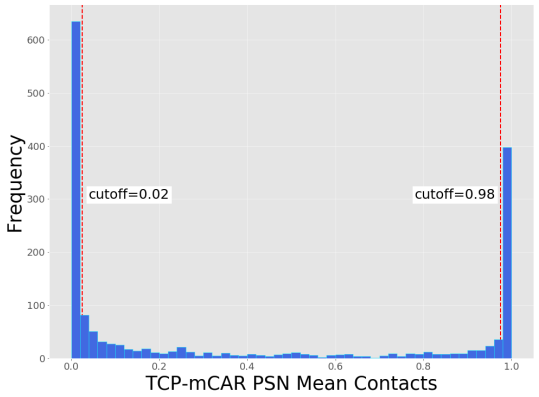
Res	Res	Mean contact
1	3	0.64
1	4	0.19
1	5	0.04
1	6	0.03
1	7	0.01
2	4	0.85
2	5	0.09
2	6	0.13
2	7	0.31
2	9	0.00
2	10	0.03
3	6	0.96
3	7	0.92
3	8	0.04
3	9	0.10
3	181	0.01
3	188	0.00

(b) Annotated mean contact map

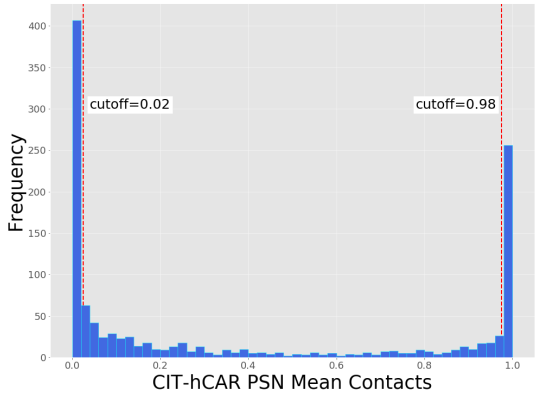
Source	Target	Mean contact
SER_107	GLU_109	0.64
SER_107	GLN_110	0.19
SER_107	GLU_111	0.04
SER_107	GLU_112	0.03
SER_107	LEU_113	0.01
LYS_108	GLN_110	0.85
LYS_108	GLU_111	0.09
LYS_108	GLU_112	0.13
LYS_108	LEU_113	0.31
LYS_108	ARG_115	0.00
LYS_108	THR_116	0.03
GLU_109	GLU_112	0.96
GLU_109	LEU_113	0.92
GLU_109	ILE_114	0.04
GLU_109	ARG_115	0.10
GLU_109	GLU_287	0.01
GLU_109	SER_294	0.00



(a)

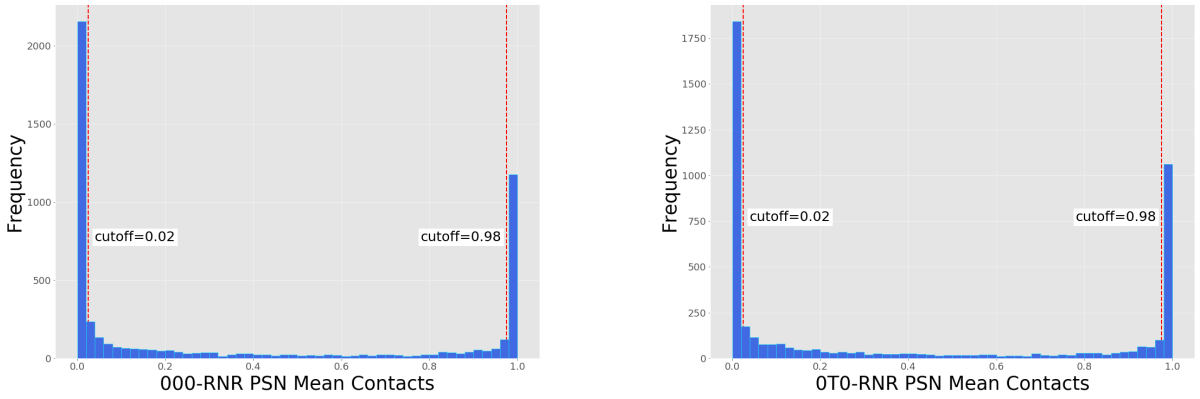


(b)



(c)

Figure 3.1: Distributions of mean contacts for CAR protein systems. Shown here are mean contact distributions for **(a)** the mCAR apo protein, **(b)** TCP-mCAR holo protein, and **(c)** CIT-hCAR holo protein. The mean contact values are indicated on the horizontal axis and frequency of mean contacts is indicated on the vertical axis.



(a) **(b)**

(c) **(d)**

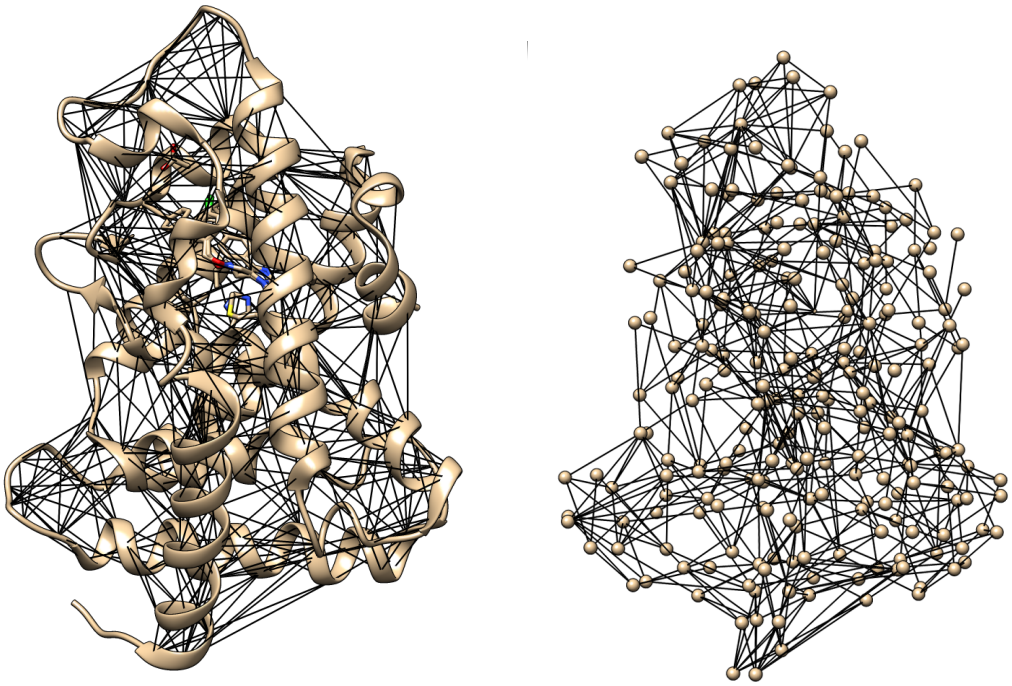
Figure 3.2: Distributions of mean contacts for RNR protein systems. Shown above are mean contact distributions for **(a)** the 000-RNR apo protein, **(b)** 0T0-RNR holo protein, **(c)** AT0-RNR holo protein and **(d)** dAT0-RNR holo protein. The mean contact values are indicated on the horizontal axis and frequency of mean contacts is indicated on the vertical axis.

Table 3.2: Example of the α-carbon table extracted from the hCAR PDB file

Atom number	Cα	Residue	Index	x-coord	y-coord	z-coord
5709	CA	SER	107	52.78	52.89	4.29
5715	CA	LYS	108	51.22	49.62	3.15
5724	CA	GLU	109	51.07	48.75	6.86
5733	CA	GLN	110	49.09	51.73	8.15
5742	CA	GLU	111	46.93	51.29	5.03
5751	CA	GLU	112	46.41	47.56	5.54
5760	CA	LEU	113	45.84	48.27	9.27
5768	CA	ILE	114	43.13	50.93	9.20
5776	CA	ARG	115	41.36	48.73	6.59
5787	CA	THR	116	41.26	45.84	9.08
5794	CA	LEU	117	40.22	48.01	12.05
5802	CA	LEU	118	37.33	49.65	10.19
5810	CA	GLY	119	36.65	46.30	8.52
5814	CA	ALA	120	35.65	44.42	11.68
5819	CA	HIS	121	34.46	47.63	13.34

In order to visualize how the results from the network analysis map back to the protein structure, it is necessary to construct 3D networks as either a stand-alone visualization or to be superimposed on the molecular simulation of protein structures (see figure 3.3). To simplify this complicated process residues in the network were connected by edges that are joined at the ends by residue Cα atoms. First, Cα data was extracted from the PDB file for each respective protein and saved to a separate file (see table 3.2). Next, a second file was created containing the PSN edge list (denoted "Source" and "Target") with their respective xyz-coordinates in adjacent columns (see table 3.3). From this table two types of networks were generated. One 3D network was used for investigating how the network structure is related to secondary protein structure (figure 3.3a). The other 3D network was generated for the purpose of investigating how values of network centrality were distributed in the network apart from the secondary structure (figure 3.3b). Both of these tasks were facilitated using the pseudobond utilities provided by UCSF-Chimera. The first and second columns of table 3.3 compose the edge list that is used to generate node IDs for edges in both the PSNs and the PDNs.

As discussed previously, centrality distributions are a means for evaluating the underlying network topology. In chapter 2, the ER, WS, and BA random graphs each exemplify specific



(a) CIT-hCAR PSN projected onto ribbon structure

(b) CIT-hCAR PSN

Figure 3.3: Two protein structure networks. Each figure presents the PSN for CIT-hCAR. **(a)** The PSN for CIT-hCAR projected onto the ribbon structure. **(b)** The same PSN presented in the standard nodes-and-edges structure. The nodes are the positioned according to the Cα atomic coordinates from the PDB file. The ligand CITCO acts as a strong agonist for hCAR. CITCO-bound at the ligand-binding domain of hCAR is abbreviated as CIT-hCAR

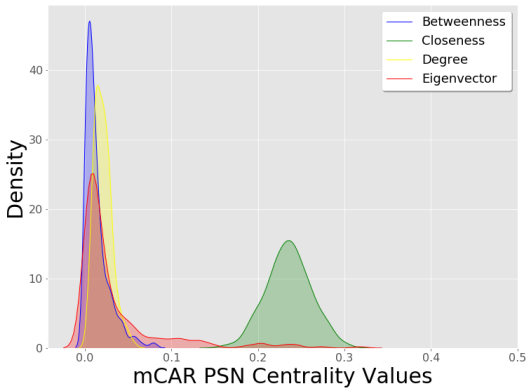
Table 3.3: PSN edge list with Cα coordinates

Source	Target	x_s coord	y_s coord	z_s coord	x_t coord	y_t coord	z_t coord
SER_107	GLU_109	52.78	52.89	4.29	51.07	48.75	6.86
SER_107	GLN_110	52.78	52.89	4.29	49.09	51.73	8.15
SER_107	GLU_111	52.78	52.89	4.29	46.93	51.29	5.03
SER_107	GLU_112	52.78	52.89	4.29	46.41	47.56	5.54
SER_107	LEU_113	52.78	52.89	4.29	45.84	48.27	9.27
LYS_108	LEU_113	51.22	49.62	3.15	45.84	48.27	9.27
LYS_108	ARG_115	51.22	49.62	3.15	41.36	48.73	6.59
LYS_108	THR_116	51.22	49.62	3.15	41.26	45.84	9.08
LYS_108	PRO_180	51.22	49.62	3.15	44.31	38.36	13.77
LYS_108	VAL_181	51.22	49.62	3.15	46.64	40.24	16.10
LYS_108	SER_184	51.22	49.62	3.15	47.20	35.43	18.23
GLU_109	GLU_287	51.07	48.75	6.86	52.19	50.77	16.50
GLU_109	SER_294	51.07	48.75	6.86	49.87	59.71	10.83
GLN_110	ARG_115	49.09	51.73	8.15	41.36	48.73	6.59
GLN_110	GLU_286	49.09	51.73	8.15	52.95	52.13	19.97
GLN_110	GLU_287	49.09	51.73	8.15	52.19	50.77	16.50
GLN_110	MET_288	49.09	51.73	8.15	48.69	52.15	16.87

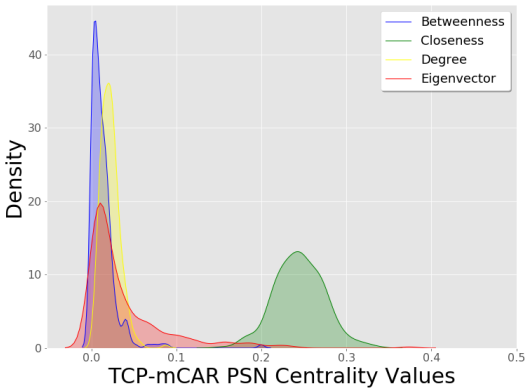
features that have been used to characterize differences between random and "real-world" networks. For example, a strong association between the power-law degree distribution is widely regarded as a non-random feature of "real-world" networks [5, 10]. In this work, we consider four measures of network centrality to determine which, if any, may be indicative of residue rearrangements that occur upon ligand-binding. Additionally, we are preparing the way to ask questions regarding the profile created by taking these four measures of centrality as a whole. Are they best described as random, or non-random features? Determining this similarity may help inform our expectations for future network analyses.

Moving forward we will consider the four centrality distributions for each protein: CAR and RNR. However, we will focus the analysis to mCAR/TCP-mCAR and 000-RNR/dAT0-RNR, that is a ligand-free system with a ligand-bound system for each protein. We will investigate these four protein systems as test cases to determine what each of the four centrality distributions may indicate about the underlying topologies of these two proteins and their various bound and unbound states. Centrality distributions for each of the seven protein systems are shown in figures 3.4 and 3.5. On inspection, we observe a high degree of similarity across all centrality distributions for both CAR and RNR systems. The hCAR system is similar to the the mCAR systems; however, there is much less variability in the hCAR EC distribution compared to the mCAR systems. From these data it is evident that ligand-binding does not perturb the shape of the centrality distributions.

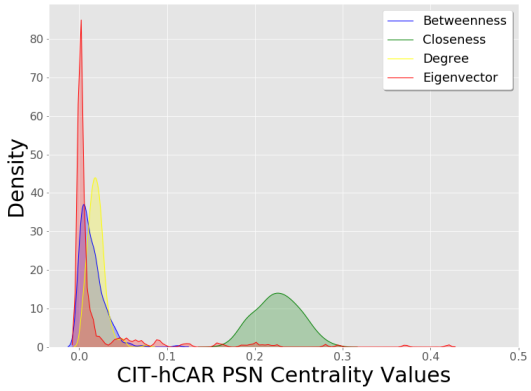
As we proceed there are four potential outcomes from the centrality analyses that we anticipate by comparing and contrasting the ligand-free from the ligand-bound systems. First, the ligand-binding event creates no shift in the ensemble of centrality distributions and the intersection of the top residues by centrality remains approximately the same. Second, the ligand-binding event creates no shift in the ensemble of centrality distributions, but the intersection of the top residues by centrality is reduced. Third, the ligand-binding event shifts the ensemble of centrality distributions, but the intersection of the top residues by centrality remains approximately the same. And fourth, the ligand-binding event shifts the ensemble of centrality distributions, and the intersection of the top residues by centrality is reduced. The consequence is the determination of which if any of the centrality measures are indicative of the residue-residue rearrangements that are occurring upon ligand-binding.



(a)



(b)



(c)

Figure 3.4: Centrality distributions for CAR. Shown above are centrality distributions for **(a)** the mCAR apo protein, **(b)** TCP-mCAR holo protein, and **(c)** CIT-hCAR holo protein. The value of centrality coefficients is indicated on the horizontal axis and frequency of coefficients is indicated on the vertical axis.

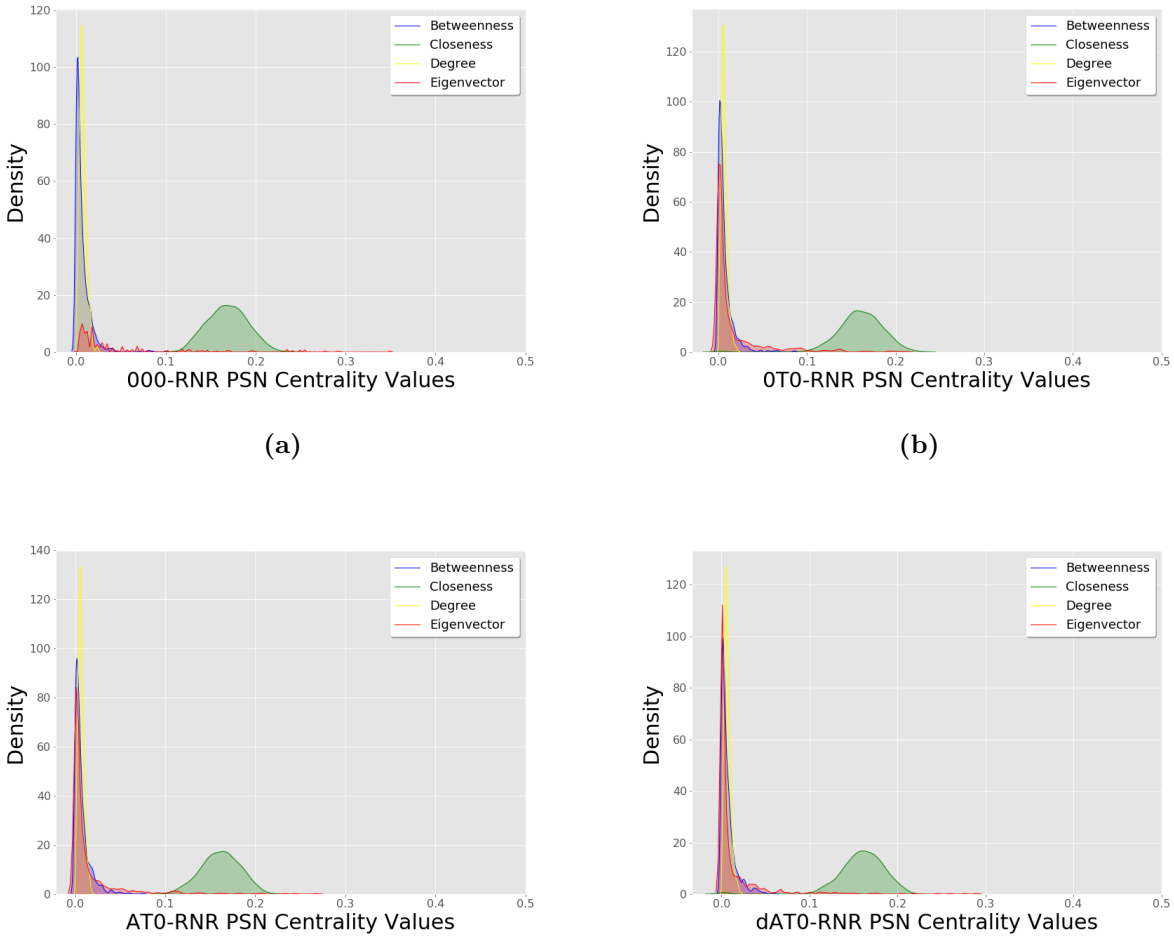


Figure 3.5: Centrality distributions for RNR. Shown above are centrality distributions for **(a)** the 000-RNR apo protein, **(b)** 0T0-RNR holo protein, **(c)** AT0-RNR holo protein and **(d)** dAT0-RNR holo protein. The value of centrality coefficients is indicated on the horizontal axis and frequency of coefficients is indicated on the vertical axis.

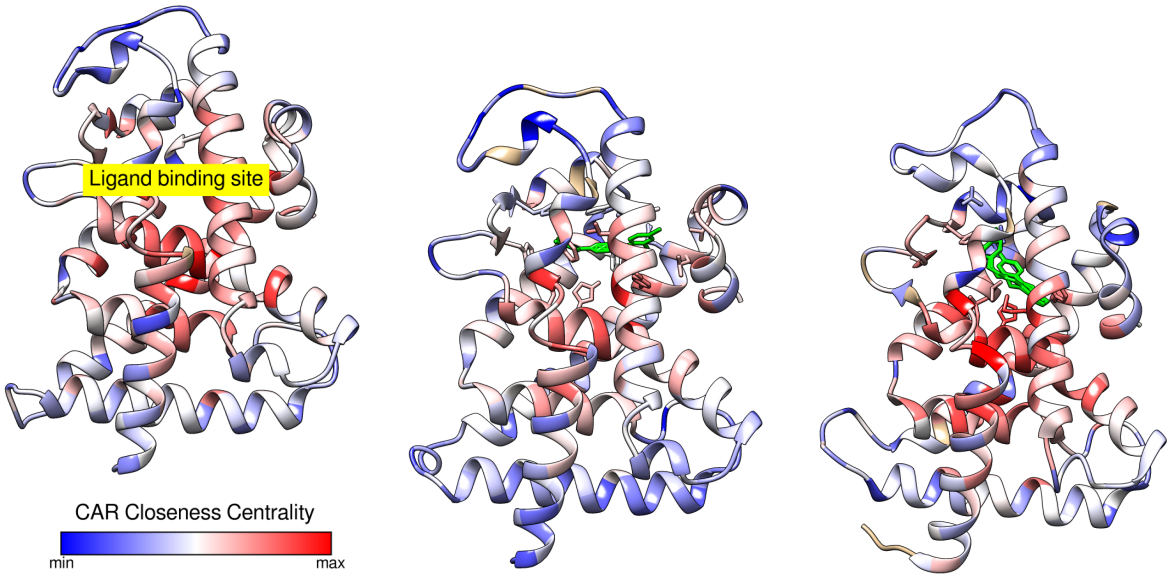
3.3 Closeness Centrality: A cautionary tale

CC has frequently been reported to identify important residues associated with protein function. Amitai *et al* [3] reported that of the 178 representative enzymes they investigated, their method of analysis accurately identified ≈ 350 or 70% of the ≈ 500 known active-site residues. The authors generated networks they call residue interaction graphs (RIGs) from a suite of computational tools integrated into one program as described in [56]. The results of these collective analyses determined the residue-residue contacts in the RIGs. Although the authors report identifying 70% of known active-site residues, they also only report a sensitivity of 47% and a specificity of 9.4%. Others have also reported the effectiveness of CC in identifying residues that are important to protein activity. After evaluating 21 different methods for constructing residue-residue contacts from 128 representative proteins, Cusack *et al* [15] report on a method that yields 70% sensitivity and 70% specificity in predicting residues involved in protein activity. They also report that 85% of those residues are located in the protein core. Additionally, Li *et al* [35] report on a novel implementation for building networks of residue-residue contacts. The authors report the successful identification of residues involved in protein function by CC in a representative set of 285 proteins. This method results in a 91% sensitivity and 88.8% specificity. However, the position of the identified residues with respect to the protein core is not mentioned. Unlike Amitai *et al* the methods employed by Cusack *et al* and Li *et al* each enlist a different Euclidean cutoff in PDB structure: $5\mathring{A}$ and $2\mathring{A}$ respectively. This raises the question of the effectiveness of using an indirect means for constructing residue-residue contacts over reliance on physical proximity [67, 51]. It is reported that residues with high CC are frequently highly conserved [17, 18, 22], found in networks exhibiting small-world topology [4], and contribute to protein rigidity [22]. Taken together, there is a consensus in the literature as to the importance of CC in identifying residues that are important to protein structure and function [12].

In figures 3.6 and 3.7 the CC values for the seven systems are mapped to their respective CAR and RNR ribbon structures. For all CAR protein systems, the highest CC values fall along helix 5 (H5) which sits 'beneath' the ligand binding site. It also appears that the lowest CC values are found along H9 and the N-terminal domain as well as the loop between

H2" and H3. The RNR systems are much larger protein systems than CAR; however, the results exhibit an analogous pattern. Upon inspection, it is observed that residues having the largest CC centrality are found at the physical center of the protein and the residues with the lowest CC values are at the activity and selectivity sites. It should also be noted that this is true regardless of the binding state. Taken together these few examples support the conclusion that in PSNs the highest CC values identify residues that are physically central to the protein structure. Perhaps this may provide a common-sense understanding of why a straightforward use of CC may successfully identify residues involved in the active site: If the active site is physically central to the protein structure or contiguous with the physical center, then the CC will provide a straightforward method for at least identifying the region of active site.

To investigate whether an analysis of CC is a useful indicator of contact rearrangements following ligand-binding, we began by plotting the CC distributions for all seven systems. Upon visual inspection, we find that the CC distributions are very similar for all systems (figures 3.8a and 3.8b). That the distributions show little difference between unbound and



(a) **(b)** **(c)**

Figure 3.6: Three molecular structures for CAR proteins with CC values mapped to each residue. From top left **(a)** the mCAR apo protein, **(b)** TCP-mCAR holo protein, and **(c)** CIT-CAR holo protein. Ligands are denoted by the color green. Lower values are indicated by blue and higher values are indicated by red.

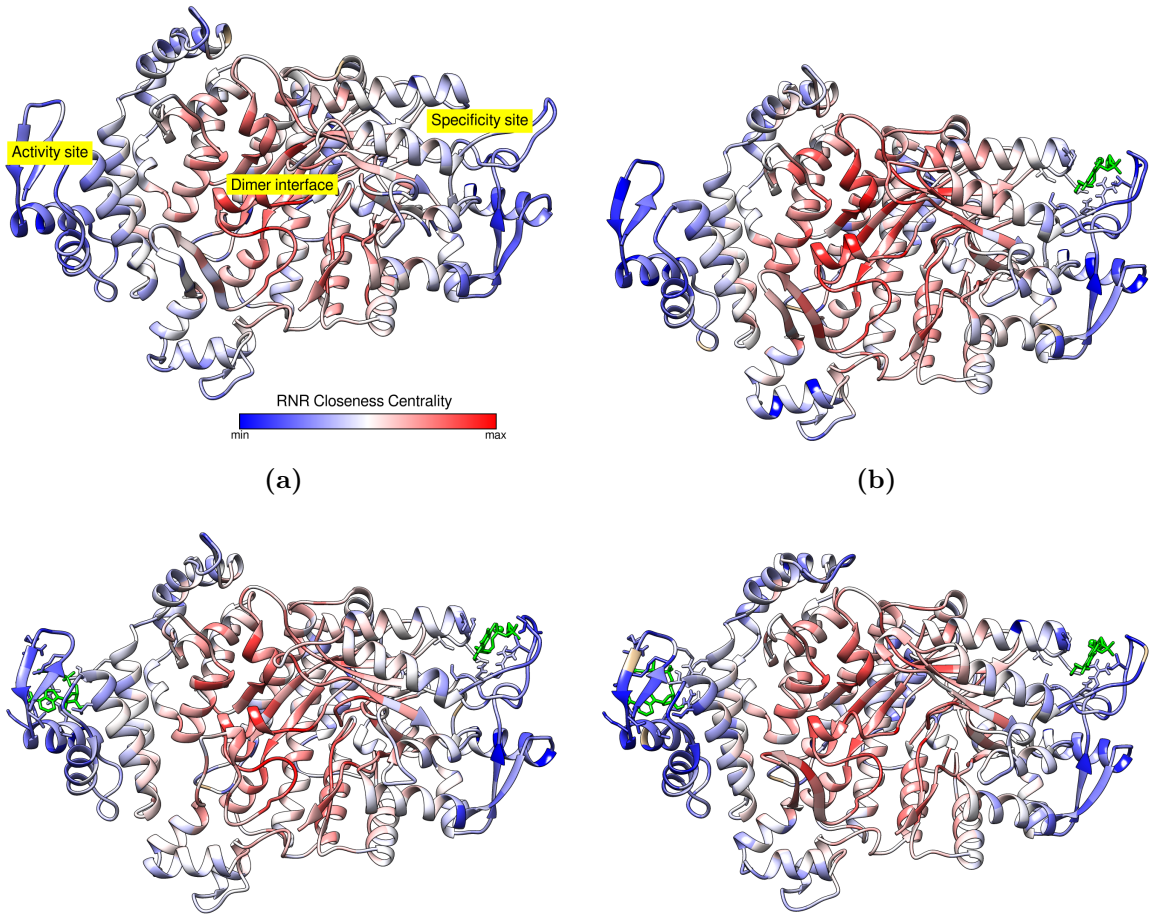


Figure 3.7: Four molecular structures for RNR proteins with CC values mapped to each residue. Displayed are four molecular structures of the RNR protein with CC values mapped to each residue. From top left **(a)** the 000-RNR apo protein, **(b)** 0T0-RNR holo protein, **(c)** AT0-RNR holo protein and **(d)** dAT0-RNR holo protein. Ligands are denoted by the color green. The range of CC values are denoted by blue-red color heatmap. Lower values are indicated by blue and higher values are indicated by red.

bound states suggesting that the rearrangements that are occurring are either dramatic, but follow the same distribution pattern or are insufficient to perturb the CC distribution. Comparing the scatter plots in figures 3.8c and 3.8d we observe that upon ligand-binding substantial contact rearrangements occurred substantially shifting the CC values; however, the overall distributions remained similar.

In figures 3.9 and 3.10 we generate Venn diagrams to examine the intersection of the set of residues with the highest CC between the ligand-free and ligand-bound systems and then map those values to residues in their respective protein structures. We define the top residues to be the 16% of residues with the highest CC values. Although a frequent cutoff for centrality values for top nodes is 5% [22, 47, 59], a cutoff as high as 20% [32] has been reported.

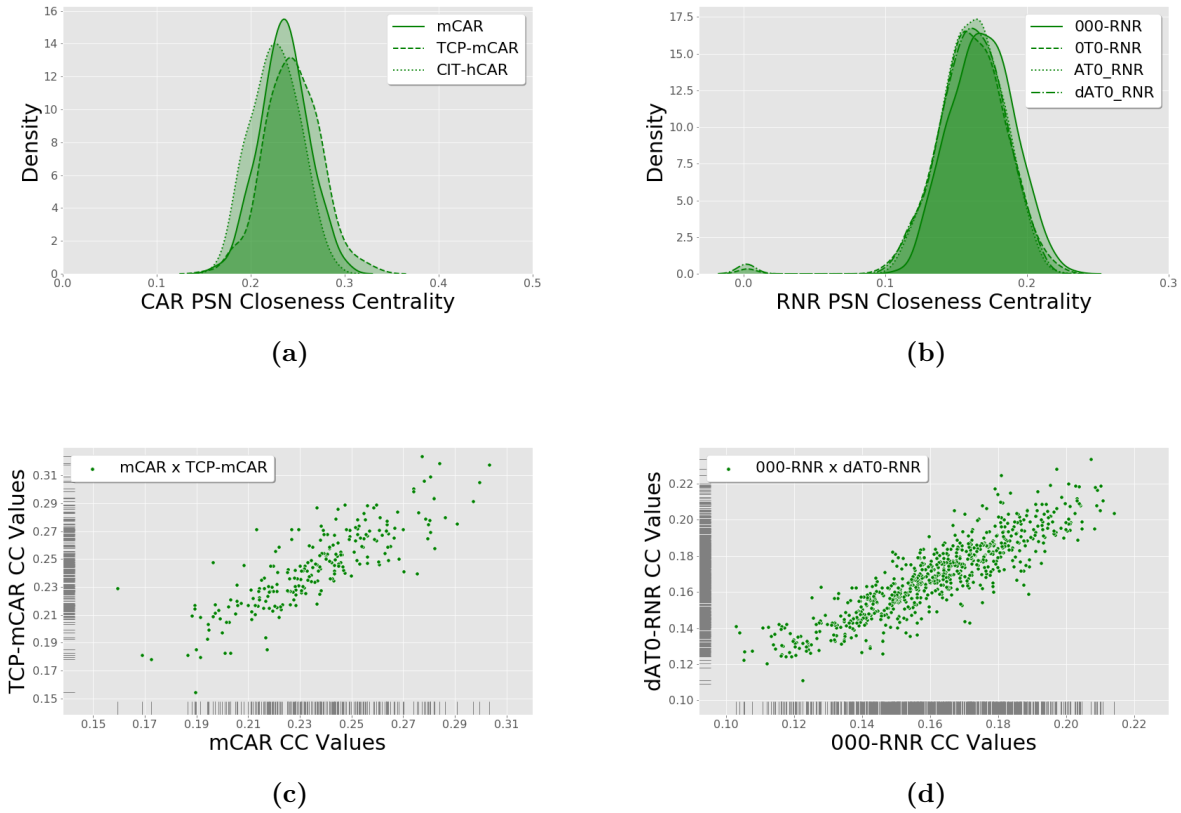
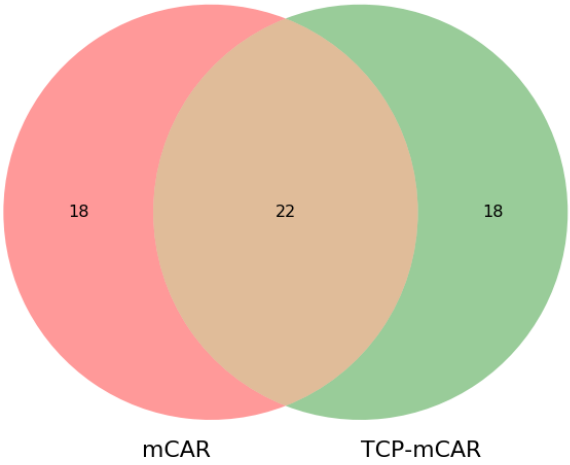


Figure 3.8: CC distributions and scatter plots for CAR and RNR. CC distributions for **(a)** CAR and **(b)** RNR. The scatter plots for **(c)** mCAR and TCP-mCAR and **(d)** 000-RNR and dAT0-RNR.

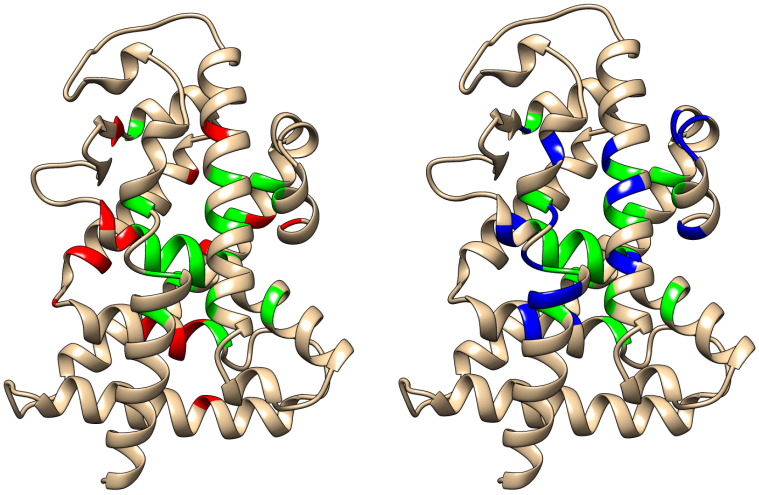
We *ad hoc* use a slightly less liberal cutoff of 16%. This constitutes a set of 40 residues for CAR and a set of 120 residues for RNR. We observe that in figures 3.9a and 3.10a that 55% of the CAR residues are found in the intersection while 72% of the RNR residues are in the intersection. This indicates that contact rearrangements following ligand-binding creates a greater rearrangement in topology in CAR than in RNR. In figure 3.9 the residues of the Venn diagrams are mapped to their respective protein structures. Green indicates the residues that are found in the intersection of the Venn diagrams. Red indicates residues in the top 16% by CC for the ligand-free system that are outside the intersection. Conversely, blue indicates residues in the top 16% by CC for the ligand-bound system that are outside the intersection. In mCAR, the residues outside the intersection are found throughout the protein structure: H5, H8, H9, and H11. A number of residues outside the intersection are found in H1, H2' and H2"-H3 loop. However, it is noteworthy that in TCP-mCAR, residues that are not in the intersection of the top 16% by CC in mCAR are found in activation function 2 (AF2) located in the C-terminal domain (H11,HX,H12). The positioning of H12 is crucial for receptor activation [66]. Binding the agonist ligand strengthens the AF2 conformation and promotes CAR transactivation [43]. In the case of RNR, the 000-RNR residues outside the CC intersection appear biased toward the specificity site and are focused in a localized region. However, for dAT0-RNR CC seems more widely dispersed along the dimer interface contiguous with residues in the RNR intersection.

The similarity of the CC distributions across all seven systems suggests that ligand-binding does not result in major changes in the distribution of CC values. Also, the reduction in the intersection of the top 16% of residues combined with the variance in the scatter plots suggests substantial contact rearrangements following ligand-binding. This is also reflected in how the CC values are mapped to the protein structure. Moreover, in TCP-mCAR, the evidence perhaps suggests that the shift in CC may have functional significance. Taken together, these results support the first alternative outlined in 1.1b.

Taken as a whole, the evidence gathered is consistent with the observation that CC identifies residues that are central to the physical structure of the protein. The implication is that if the active site is physically central to the protein structure, then CC may identify residues that contribute to protein activity. However, if the active site is not physically

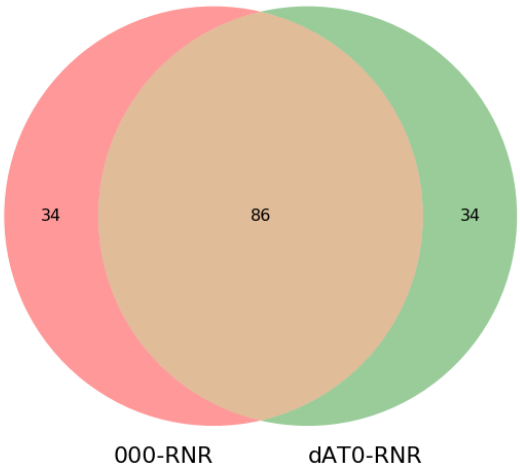


(a)

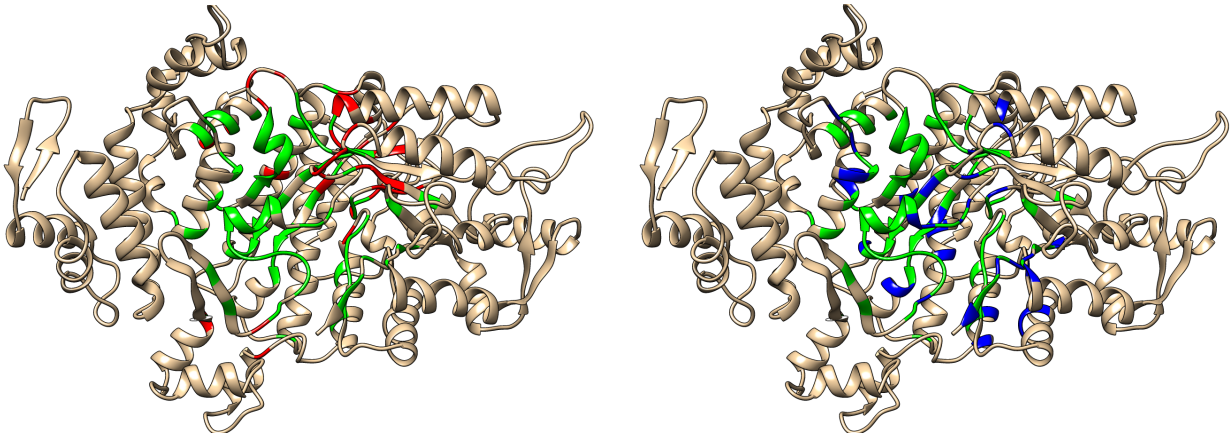


(b) **(c)**

Figure 3.9: Venn diagrams of top 16% of residues by CC for mCAR. **(a)** The intersection of the top 16% of residues between mCAR and TCP-mCAR. The intersections of the two sets are shown in brown. Red indicates the set of residues from the mCAR system outside the intersection. Green indicates the set of residues from the TCP-mCAR system outside the intersection. **(b,c)** The intersection of the top 16% of residues between mCAR and TCP-mCAR mapped onto the protein structure. The intersecting set of common residues between mCAR and TCP-mCAR are indicated by green. **(b)** The set of residues for mCAR that are outside the intersection are indicated by red. **(c)** The set of residues for TCP-mCAR that are outside the intersection are indicated by blue.



(a)



(b) **(c)**

Figure 3.10: Venn diagrams of top 16% of residues by CC for RNR. (a) The intersection of the top 16% of residues between 000-RNR and dAT0-RNR. The intersections of the two sets are shown in brown. Red indicates the set of residues from the 000-RNR system outside the intersection. Green indicates the set of residues from the dAT0-RNR system outside the intersection. **(b,c)** The intersection of the top 16% of residues between 000-RNR and dAT0-RNR mapped onto the protein structure. The intersecting set of common residues between 000-RNR and dAT0-RNR are indicated by green. **(b)** The set of residues for 000-RNR that are outside the intersection are indicated by red. **(c)** The set of residues for dAT0-RNR that are outside the intersection are indicated by blue.

central, then caution CC is not a good indicator of the active site residues. To make the argument that there is a specific relationship between CC and the active site, research would need to focus on testing proteins where the active site is physically 'off-center' to the protein.

3.4 Degree Centrality: "Real-worlds" and Scale-free Topology

One of the principal features that characterize network topology is the degree distribution. The reason for this is that other network features are strongly influenced by DC such as clustering and the emergence of hubs. The ER, WS, and BA models [25, 40] each characterize different network topologies and–with some caveats–an associated degree distribution. Algorithms for building ER and WS graphs generate topologies that follow binomial and Poisson degree distributions respectively. The ER procedure for connecting N vertices with m edges follows a connection probability of p with a corresponding 'non-connection' probability of $1 - p$; therefore, the resulting binomial degree distribution for ER graphs yields a uniform or homogeneous topology with low clustering. The significance of this is that these networks fail to produce important features of "real-world" networks. As such the ER graph are often used as the null-graph. The principal contribution of the WS model is the inclusion of a tuneable clustering coefficient that generates graphs with greater or lesser degree of clustering. Additionally, the WS graphs exhibit the well-known "small-world" effect where the average shortest path length scales logarithmically with graph size N. The advantage to this procedure is that random graphs can be produced with clustering that is analogous to that observed in "real-world" networks. The contribution of BA graphs is the development of "preferential attachment". In short, the procedure begins with an equal connection probability for N vertices which changes over the course of the procedure to favor vertices with more edges. In other words, the connection probability becomes biased in favor of vertices with larger degree. To begin, nodes have equal probability of having an edge assigned. For each iteration of the algorithm the degree of each vertex is evaluated and the probability of edge assignment becomes biased towards vertices with the highest

degrees. This method of generating a random graph results in the canonical scale-free or power-law degree distribution. BA graphs are characterized by a small number of highly connected vertices and a large number of single-degree vertices. The BA graphs exhibit the "small-world" effect, increased clustering, and scale-free topology all of which correspond to features that have come to be identified with "real-world" networks. For these reasons the BA or scale-free graph has become a reference model in the study of social and biological networks. Specifically, the concept of preferential attachment with its corresponding network hubs has spawned a considerable amount of interest and research because of the assumption that hubs in "real-world" networks are formed by nonrandom, preferential mechanisms and, consequently, are important. This assumption has not been without merit.

The ER, WS, and BA models can provide researchers with a quick-and-dirty method of evaluating network structure. Consider the degree distributions for the seven protein systems in figures 3.11a and 3.11b. Upon visual inspection it is seen that the PSNs for CAR and RNR do not follow a power-law degree distribution. Again, the feature that makes power-law degree distributions of interest is the assumption hubs emerge non-randomly and are consequently highly significant. Therefore, in the absence of a power-law degree distribution and the corresponding assumptions, it isn't clear what high-degree nodes in the PSN signify. In the BA model high-degree vertices are a non-random feature. In contrast, in the WS model high-degree vertices are a random feature. Therefore, one may ask whether high-degree residues in PSNs are a random or non-random feature. Do PSNs follow a network connectivity consistent with the ER or the WS models? At this time that remains an open question that awaits further investigation. As mentioned previously, this will require the generation of an ensemble of models and determining which ensemble offers the closest description. At the very least it can be said that even though PSNs are not "scale-free", they are nonetheless 'things' we find in the "real-world", which perhaps suggests that we reconsider what assumptions we should make regarding biological networks.

Upon visual inspection it is clear that the DC distributions for CAR are highly similar as are the DC distributions for RNR (figures 3.11a and 3.11b) indicating that residue-residue contact rearrangements following ligand binding do not perturb the PSN DC distribution. Additionally, the distribution of DC values for CAR systems are significantly

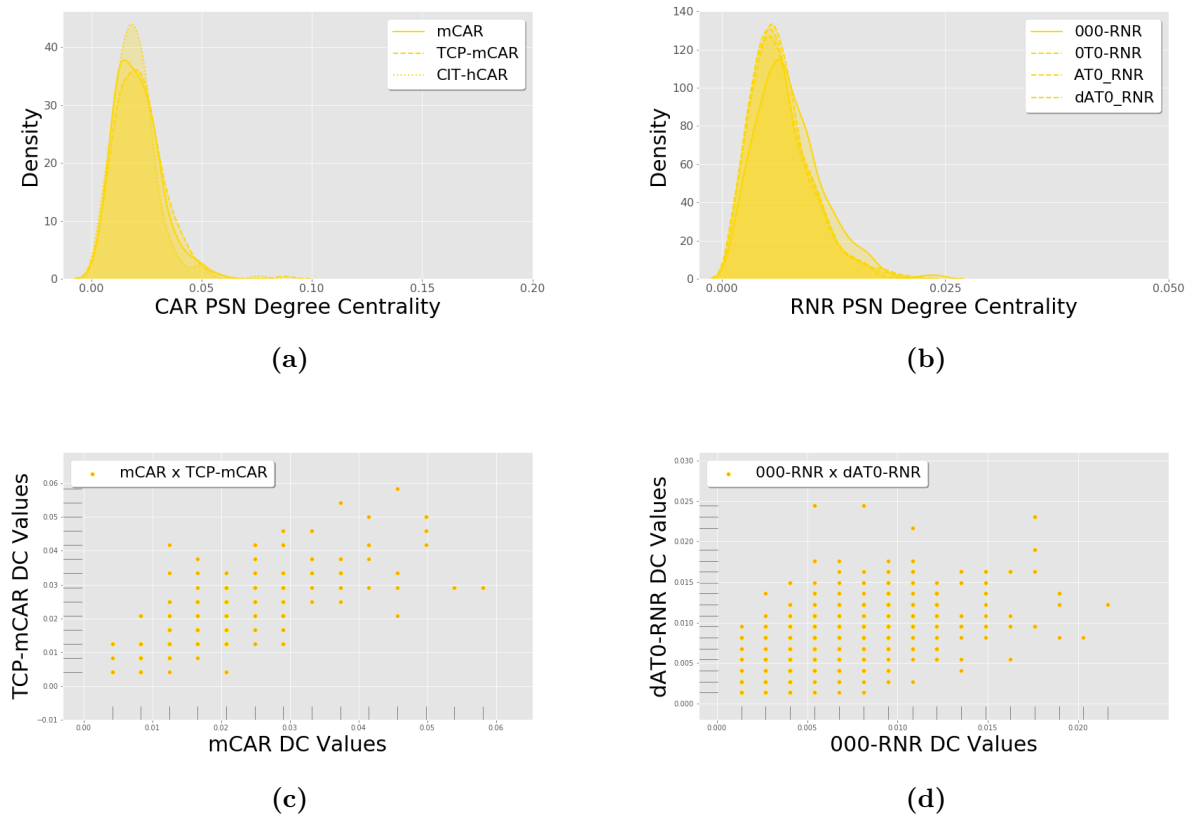


Figure 3.11: DC distributions and scatter plots for CAR and RNR. DC distributions for **(a)** CAR and **(b)** RNR. The scatter plots for **(c)** mCAR and TCP-mCAR and **(d)** 000-RNR and dAT0-RNR.

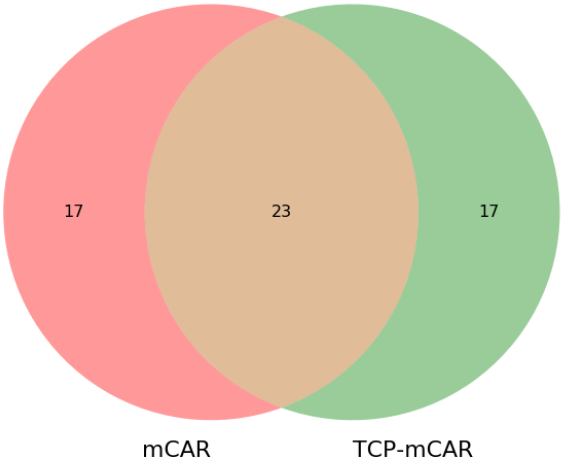
larger suggesting that the structure for the CAR systems is more compact than those of RNR. The scatter plots for DC may appear peculiar (figures 3.11c and 3.11d). This results from the normalization of discrete DC values. Upon visual inspection a positive linear relationship with a slight positive, right skew in both the ligand-free and ligand-bound systems for CAR and RNR. This is indicates that the mass or density of the DC values is concentrated at lower values. Figures 3.12a and 3.13a show the Venn diagrams for the CAR and RNR proteins. The intersection of the top 16% of residues by DC for CAR is 58% which is similar to the CAR intersection for CC. However, the intersection of the top 16% of residues by DC for RNR is at 47%. The implication is that even though distributions of DC values do not exhibit noticeable differences following ligand-binding, the intersection indicates that a significant shift in connectivity in fact occurred. This is supported by the figures 3.13b and

3.13c that show residues outside the DC intersection are interspersed throughout the protein structures of 000-RNR or dAT0-RNR. In figures 3.12b and 3.12c it is observed that a number of residues that fall inside the intersection for DC are at the N-terminus and at H9, which is physically contiguous to the N-terminus domain. Again, residues outside the intersection for TCP-mCAR are at the AF2 domain.

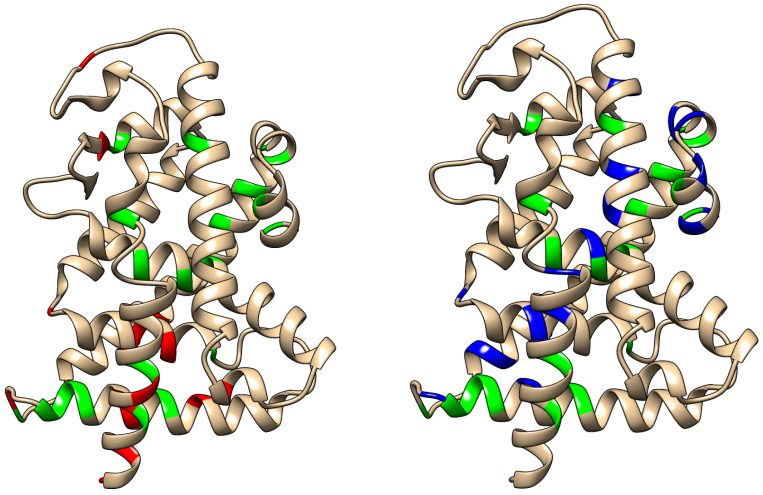
The similarity of the DC distributions across all seven systems suggests that ligand-binding does not result in major changes in the distribution of DC values. Also, the reduction in the intersection of the top 16% of residues combined with the variance in the scatter plots suggests substantial contact rearrangements following ligand-binding. This is also reflected in how the DC values are mapped to the protein structure. Moreover, the range of DC values and the height of the distributions between CAR and RNR systems suggests that CAR has a more compact tertiary structure than RNR. Additionally, in TCP-mCAR, the evidence perhaps suggests that the shift in DC may have functional significance with respect to the AF2 domain. Taken together, these results support the first alternative outlined in 1.1b.

3.5 Betweenness Centrality: Information Highways or the Paths Less Travelled?

In scale-free systems, BC is associated with high-degree nodes (network hubs). This follows from the observation that the distance between any two nodes in the network is reduced by the presence of a few, highly interconnected hubs. Because shortest paths decrease the distance across even very large networks they are often referred to as 'information highways'. As described previously, BC is a measure of the total number of shortest paths that 'pass-through' a given vertex. In scale-free networks–like the BA model–BC and DC are related by the high-degree hubs. However, the degree distributions for the PSNs for CAR and RNR are not scale-free. Therefore, this raises the question to what extent residues with high BC are influenced by high-degree residues (Perhaps a simple scatter plot of the DC and BC values would be sufficient to answer the question). To put it in other words, are residues

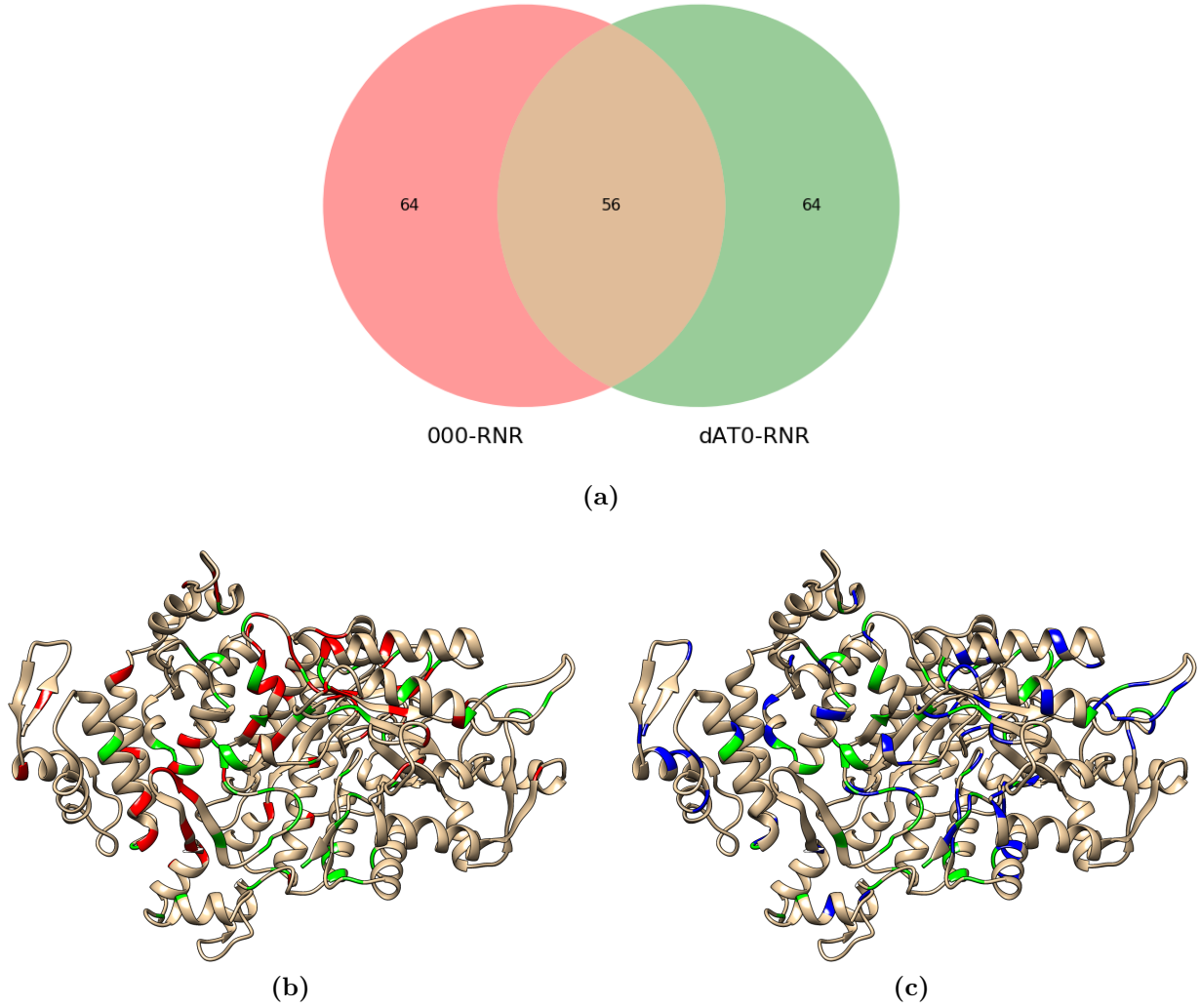


(a)



(b) **(c)**

Figure 3.12: Venn diagrams of top 16% of residues by DC for mCAR. **(a)** The intersection of the top 16% of residues between mCAR and TCP-mCAR. The intersections of the two sets are shown in brown. Red indicates the set of residues from the mCAR system outside the intersection. Green indicates the set of residues from the TCP-mCAR system outside the intersection. **(b,c)** The intersection of the top 16% of residues between mCAR and TCP-mCAR mapped onto the protein structure. The intersecting set of common residues between mCAR and TCP-mCAR are indicated by green. **(b)** The set of residues for mCAR that are outside the intersection are indicated by red. **(c)** The set of residues for TCP-mCAR that are outside the intersection are indicated by blue.



(a)

(b)

(c)

Figure 3.13: Venn diagrams of top 16% of residues by DC for RNR. **(a)** The intersection of the top 16% of residues between 000-RNR and dAT0-RNR. The intersections of the two sets are shown in brown. Red indicates the set of residues from the 000-RNR system outside the intersection. Green indicates the set of residues from the dAT0-RNR system outside the intersection. **(b,c)** The intersection of the top 16% of residues between 000-RNR and dAT0-RNR mapped onto the protein structure. The intersecting set of common residues between 000-RNR and dAT0-RNR are indicated by green. **(b)** The set of residues for 000-RNR that are outside the intersection are indicated by red. **(c)** The set of residues for dAT0-RNR that are outside the intersection are indicated by blue.

with high BC in PSNs heavily-trafficked pathways communicating important motion and "information" throughout the network, or are they less functional...like a side road?

Upon visual inspection it is clear that the BC distributions for CAR are highly similar as are the BC distributions for RNR (figures 3.14a and 3.14b). The BC distributions for the RNR systems are hardly distinguishable. The scatter plots for BC (figures 3.14c and 3.14d) exhibit a positive, right skew in both distributions. As mentioned before, this suggests that the mass of the BC coefficients are concentrated at lower values. Given that the BC distributions mirror those of DC, perhaps this suggests a a relationship between DC and BC in PSNs. Figures 3.16a and 3.15a display the Venn diagrams for the CAR and RNR proteins. The intersection of the top 16% of residues by BC for CAR is 58% which is similar to both the CC and DC. However, the intersection of the top 16% of residues by BC for RNR is at 48%. This is similar to the intersection observed in the RNR DC intersection. This is

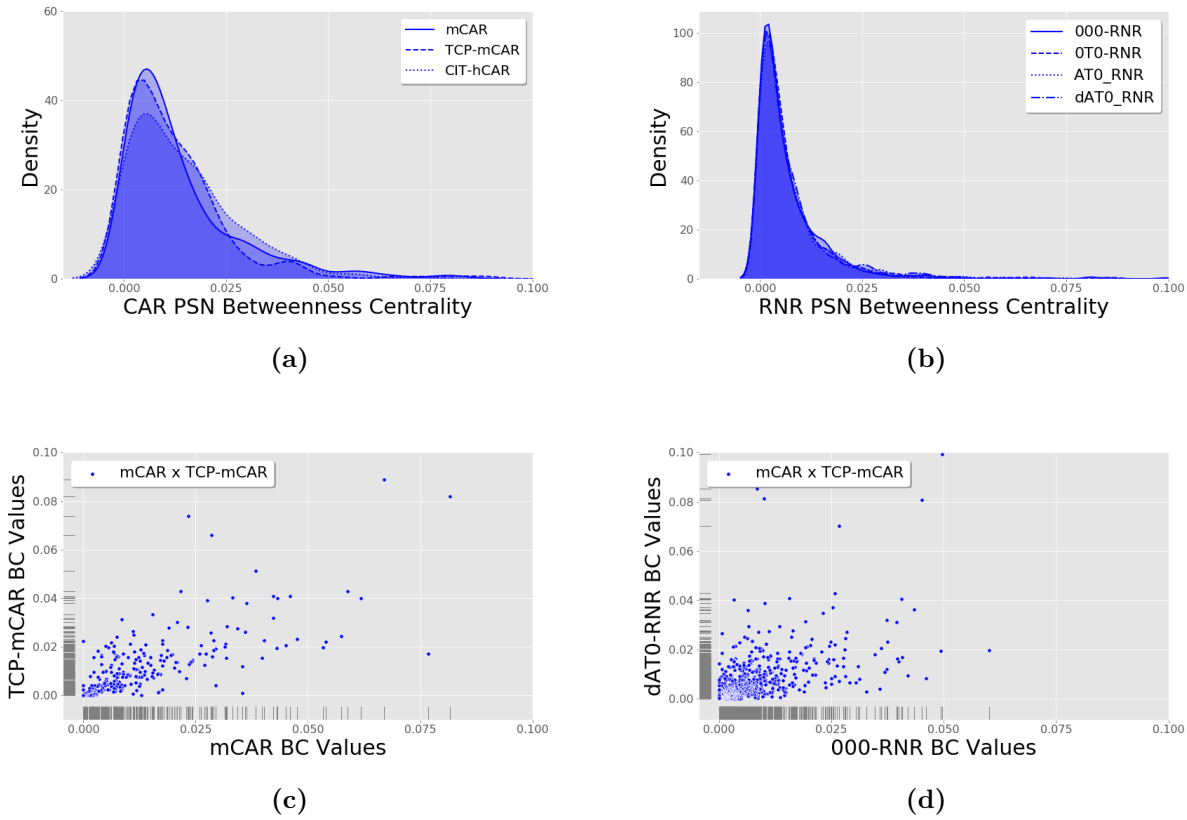
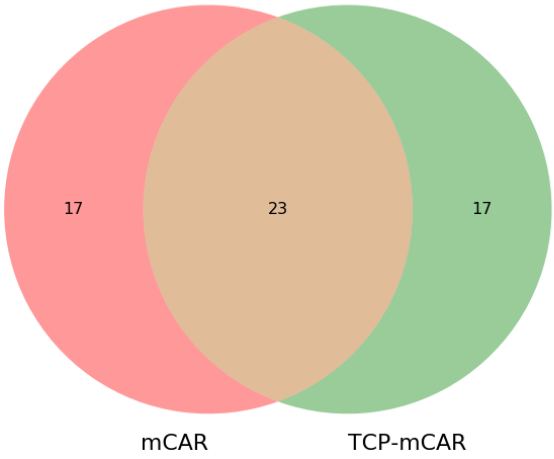
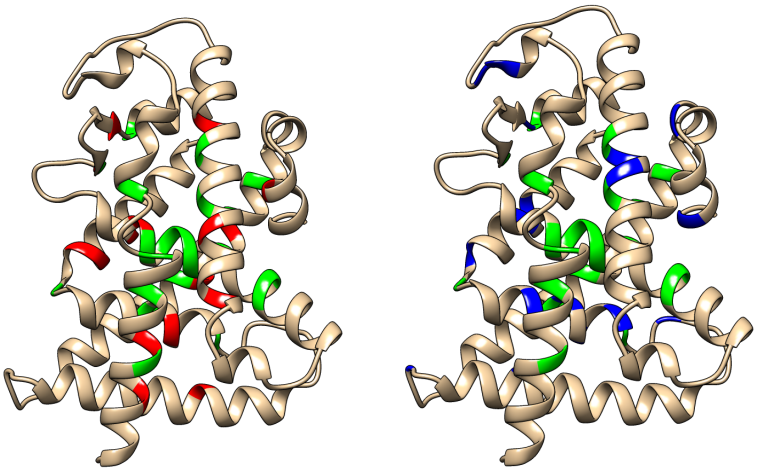


Figure 3.14: BC distributions and scatter plots for CAR and RNR. BC distributions for **(a)** CAR and **(b)** RNR. The scatter plots for **(c)** mCAR and TCP-mCAR and **(d)** 000-RNR and dAT0-RNR.

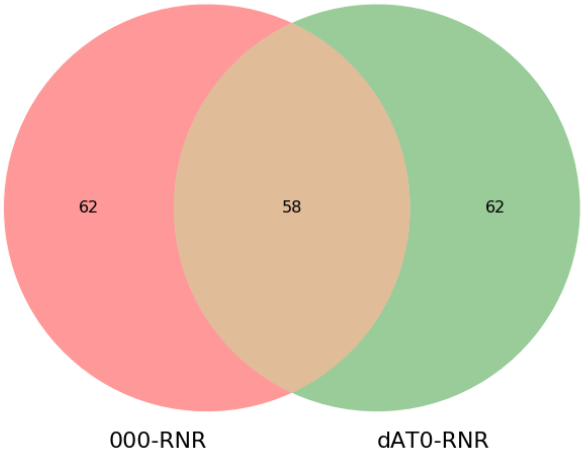


(a)

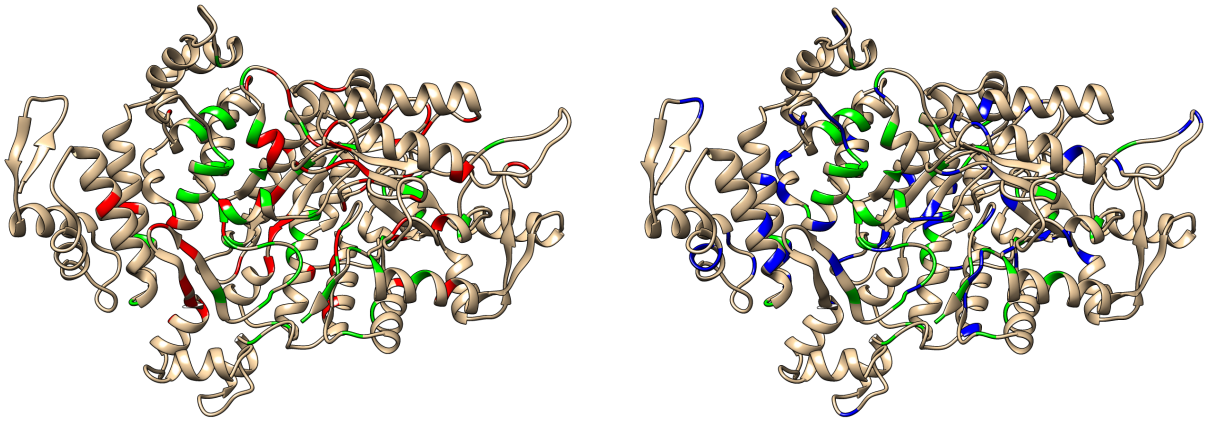


(b) **(c)**

Figure 3.16: Venn diagrams of top 16% of residues by BC for mCAR. (a) The intersection of the top 16% of residues between mCAR and TCP-mCAR. The intersections of the two sets are shown in brown. Red indicates the set of residues from the mCAR system outside the intersection. Green indicates the set of residues from the TCP-mCAR system outside the intersection. **(b,c)** The intersection of the top 16% of residues between mCAR and TCP-mCAR mapped onto the protein structure. The intersecting set of common residues between mCAR and TCP-mCAR are indicated by green. **(b)** The set of residues for mCAR that are outside the intersection are indicated by red. **(c)** The set of residues for TCP-mCAR that are outside the intersection are indicated by blue.



(a)



(b) **(c)**

Figure 3.15: Venn diagrams of top 16% of residues by BC for RNR. (a) The intersection of the top 16% of residues between 000-RNR and dAT0-RNR. The intersections of the two sets are shown in brown. Red indicates the set of residues from the 000-RNR system outside the intersection. Green indicates the set of residues from the dAT0-RNR system outside the intersection. **(b,c)** The intersection of the top 16% of residues between 000-RNR and dAT0-RNR mapped onto the protein structure. The intersecting set of common residues between 000-RNR and dAT0-RNR are indicated by green. **(b)** The set of residues for 000-RNR that are outside the intersection are indicated by red. **(c)** The set of residues for dAT0-RNR that are outside the intersection are indicated by blue.

supported by the figures 3.16b and 3.16c as well as 3.15b and 3.15c which show that the residues outside the BC intersection for both CAR and RNR are interspersed throughout the protein structures and generally physically contiguous to the residues inside the intersection.

The similarity of the BC distributions across all seven systems suggests that ligand-binding does not result in major changes in the distribution of BC values. Also, the reduction in the intersection of the top 16% of residues combined with the variance in the scatter plots suggests substantial contact rearrangements following ligand-binding. This is also reflected in how the BC values are mapped to the protein structure. Also, the similarity between the DC distributions and BC distributions for both the CAR and RNR systems suggests a potential dependency of BC on DC. Taken together, these results support the first alternative outlined in 1.1b.

3.6 Eigenvector Centrality: The Twilight Zone of Centralities

As described previously, EC is a measure that is intrinsically dependent on DC. Therefore, since substantial contact rearrangements were observed among the top 16% of residues by DC in the CAR and RNR proteins, there is a corresponding expectation that reordering of the top 16% of EC values will also occur. Recall that The EC of a vertex is defined by the eigenvector of the largest eigen value λ of the adjacency matrix A. Because EC scales by λ^{-1} this has a tendency to generate long-tails in a pronounced right-skewed distribution [40].

On visual inspection, the EC distributions for the seven systems are different. Although the EC distributions for mCAR and TCP-mCAR are highly similar (figure 3.17a), the EC distribution for CIT-hCAR is quite different. Given the consistency of the BC, CC, and DC of CIT-hCAR with the mCAR orthologs one might expect that the EC would also resemble the mCAR EC distributions. Coupled with the dependence of EC on the DC it might appear odd at first blush that the EC distribution would exhibit this variation. Because the ligand-free PSN was for hCAR is not part of the analysis in this work, it is impossible to interpret the

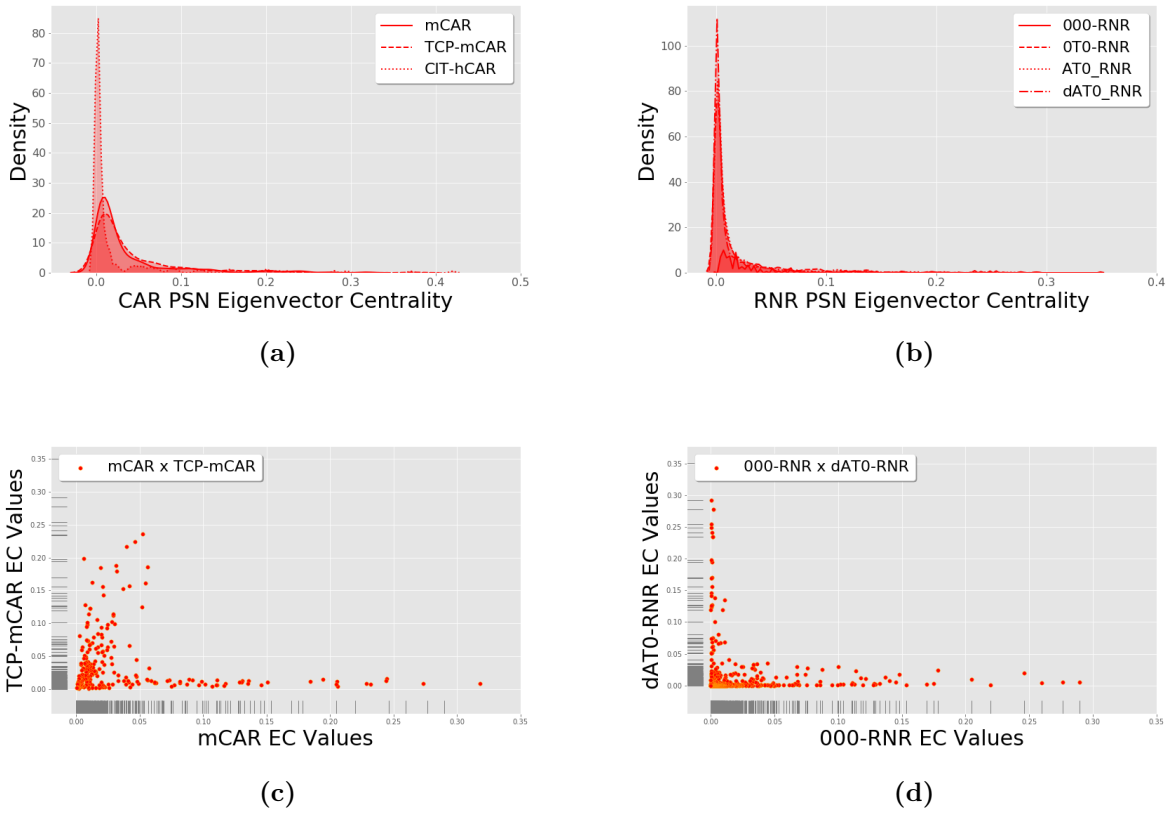


Figure 3.17: EC distributions and scatter plots for CAR and RNR. EC distributions for **(a)** CAR and **(b)** RNR. The scatter plots for **(c)** mCAR and TCP-mCAR and **(d)** 000-RNR and dAT0-RNR.

significance of this apparent divergence. Furthermore, the expectation that the EC would yield heavily-skewed right-tailed distributions is confirmed by the scatter plot in figure 3.17c.

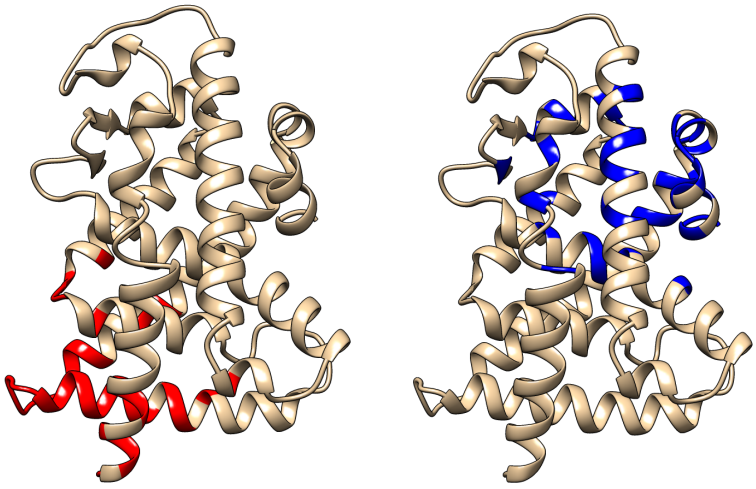
On visual inspection, the EC distributions for the RNR systems are similar. The exception here is 000-RNR. The distribution appears 'truncated'. Here, again, what makes this striking is the dependence of EC on the DC. The EC distributions for 0T0, AT0, and dAT0 RNR exhibit the Heavy right-skewed distributions one might expect given the DC distributions (figure 3.17b). Considering the consistency of the BC, CC, and DC of 000-RNR with with the centrality distributions of the other RNR systems, it seems reasonable to expect that the EC would be similar to the EC distributions as well. The scatter plot in figure 3.17d is consistent with the EC distributions in figure 3.17b.

Furthermore, the Venn diagrams for mCAR/TCP-mCAR indicate a complete reduction of CAR residues in the intersection of the top 16% by EC (figure 3.18a). It is striking to observe that the EC values for the CAR protein structures show no overlapping regions and the grouping of residues occurs in distinct physical locations in the protein. The top 16% of EC values for mCAR are localized at the N-terminal domain and the H9-H10 loop (3.18b). However, the top 16% of EC values for TCP-mCAR are focused at the C-terminal domain at the AF2 region located in H11, HX, and H12 (3.18c). Residues in the top 16% for EC are also found on H2' and H3. It is worth noting that these regions are key to modulating the activity of CAR. H3 and the H9-H10 loop are involved in the formation of the homodimer interface and the stabilization of H11, HX, and H12 subsequent to agonist binding is critical to CAR activation [66, 43].

The Venn diagrams for 000-dAT0 RNR indicate a dramatic reduction the intersection of residues in the top 16% by EC (figure 3.19a). This is consistent with the mapping of the EC values to the RNR protein structures. The residues of the top 16% of EC values for 000-RNR that are outside the intersection (figure 3.19b) are concentrated toward the specificty site. However, the top 16% of EC values for TCP-mCAR are concentrated about the activity site (figure 3.19c). Significant redistribution of EC values upon ligand-binding has been reported previously [39].

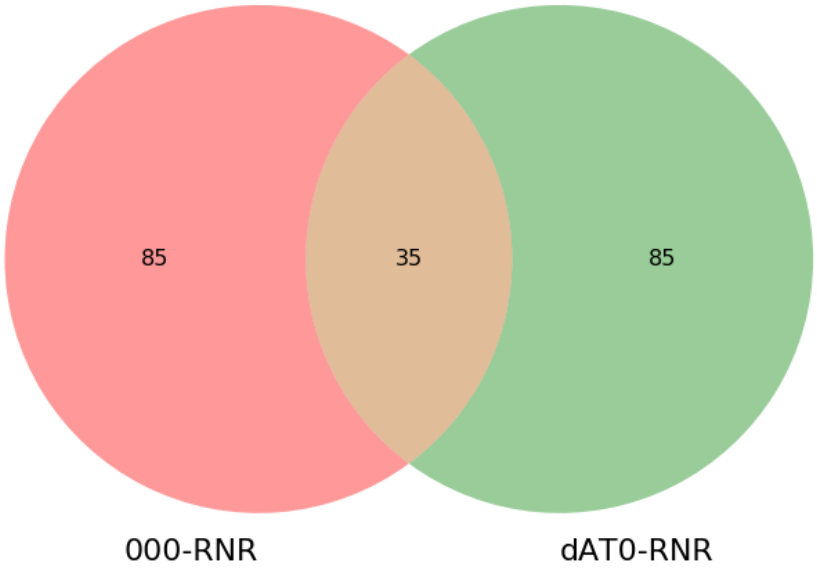
The evidence from EC supports two of the alternatives outlined in 1.1. First, in the case of CAR, the EC distributions are very similar and there is a complete reduction in the intersection of the sets of top 16% of residues by EC centrality. It is likely that this is due to scaling by the inverse of the largest eigen value, λ. If the residues with the highest degree are in different physical locations in the protein structure between mCAR and TCP-mCAR, then scaling by λ^{-1} will 'force' adjacent residues to 1 in mCAR while 'forcing' the EC values of those same residues toward 0 in TCP-mCAR, and vice-versa. Second, in the case of RNR, the EC distributions are not similar and the intersection of the set of top 16% of residues by EC are substantially reduced. In this instance, the results support the third alternative outlined in 1.1d.

(a)

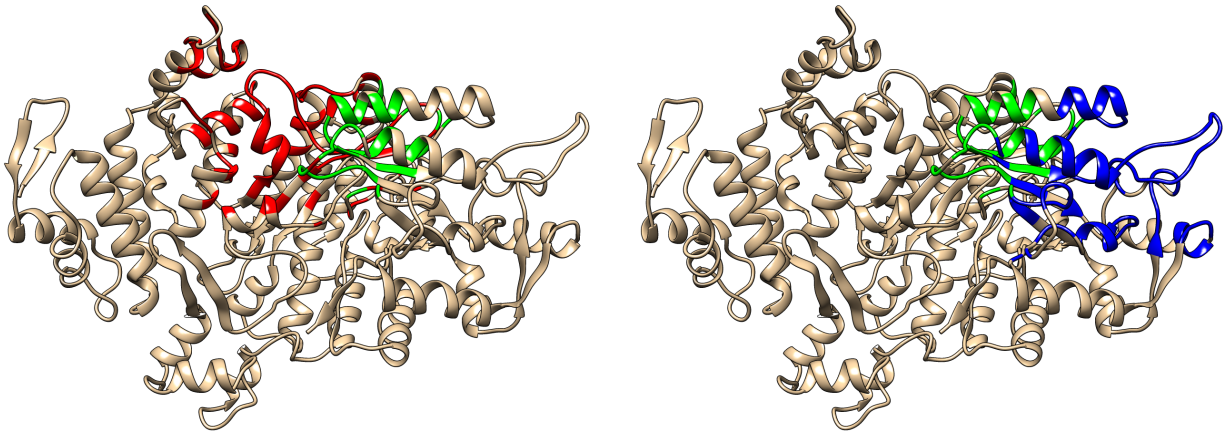


(b) **(c)**

Figure 3.18: Venn diagrams of top 16% of residues by EC for mCAR. (a) The intersection of the top 16% of residues between mCAR and TCP-mCAR. The intersections of the two sets are shown in brown. Red indicates the set of residues from the mCAR system outside the intersection. Green indicates the set of residues from the TCP-mCAR system outside the intersection. **(b,c)** The intersection of the top 16% of residues between mCAR and TCP-mCAR mapped onto the protein structure. The intersecting set of common residues between mCAR and TCP-mCAR are indicated by green. **(b)** The set of residues for mCAR that are outside the intersection are indicated by red. **(c)** The set of residues for TCP-mCAR that are outside the intersection are indicated by blue.



(a)



(b) **(c)**

Figure 3.19: Venn diagrams of top 16% of residues by EC for RNR. (a) The intersection of the top 16% of residues between 000-RNR and dAT0-RNR. The intersections of the two sets are shown in brown. Red indicates the set of residues from the 000-RNR system outside the intersection. Green indicates the set of residues from the dAT0-RNR system outside the intersection. **(b,c)** The intersection of the top 16% of residues between 000-RNR and dAT0-RNR mapped onto the protein structure. The intersecting set of common residues between 000-RNR and dAT0-RNR are indicated by green. **(b)** The set of residues for 000-RNR that are outside the intersection are indicated by red. **(c)** The set of residues for dAT0-RNR that are outside the intersection are indicated by blue.

3.7 Discussion and Conclusions

At the outset we outlined four possible results from the work reported here. The first possibility we proposed is that a ligand-binding event induces no substantial change in the shape of the centrality distributions and that the intersection of top residues by centrality would remain approximately the same. In some sense, this serves as a null hypothesis: There is no substantial difference in topology between the ligand-free and ligand-bound PSN. Put in other words, there is either only trivial differences or the significant differences are so subtle that the methods available to network analysis are too coarse to make the discrimination. The first alternative we propose is that the ligand-binding event creates no substantial change in the shape of the centrality distributions, but the intersection of the top residues by centrality is reduced. In the first instance we propose that no difference implies that the intersection remains unchanged. But, it means more than merely unchanged. 'No substantial difference in topology' implies that the intersection of the top 16% by centrality would be complete, i.e. there would be no residues outside the intersection, or at least no more than would be accounted for at random. Therefore, one alternative to the null case is that the intersection is reduced–less than complete. The second alternative proposed is that the ligand-binding event shifts the centrality distributions, but the intersection of the top residues by centrality remains approximately the same. A shift in the distributions may imply a more or a less compact protein structure, but substantial intersection implies that the core topological features are unchanged. The last alternative proposed is that the ligand-binding event shifts the ensemble of centrality distributions, and the intersection of the top residues by centrality is reduced. This would be a strong indication that substantial contact rearrangements were induced upon ligand-binding.

The evidence presented here supports the first alternative as described above. The distributions for BC, CC, and DC are highly similar for both the CAR and RNR systems. Additionally, the intersections and scatter plots for mCAR and TCP-mCAR and 000-RNR and dAT0-RNR show substantial contact rearrangements. For CAR the sizes of the intersecting sets of residues for CC, DC, and BC are 22, 23, and 23; and for RNR they are 86, 56, and 58, respectively. Taken together these data point to ligand-dependent contact

rearrangements that either do not perturb the network topology or subsequently results in a highly equivalent topology. One explanation for this may be that a stable, well-folded three-dimensional structure constrains the motion and the number of potential contacts that can be made by any given residue such that the making and breaking of contacts averages out over the global structure.

Chapter 4

Summary and Future Work

Network analysis is widely used for interrogating biological systems. One reason for this is the potential network analysis holds for identifying features in data that are essential to the dynamics and function of complex systems. Networks can also be used to make predictions about how the system may function or how dynamics may change under perturbation. Here we treated residues composing the three-dimensional protein structure as a set of nodes and mean contacts derived from MD simulations as the potential set of edges between residues. A mean contact is the average "contact" between two residues over the time course of MD simulation. A "contact" between two residues is said to occur whenever any two atoms between two residues are within a defined Euclidean distance or cutoff. The PSNs used here were created from residues (nodes) and mean contacts (edges).

4.1 Summary

This research was conducted to determine how contact rearrangements between residues subsequent to ligand-binding affects the properties of the PSN. Here we use PSNs created from seven protein systems under different ligand-binding states. We used two proteins as our model systems, the transcription factor constitutive androstane receptor and the enzyme ribonucleotide reductase. We investigated how betweenness, closeness, degree and eigenvector centralities varied between ligand-free and ligand-bound systems.

Ligand-binding induced substantial changes in centrality values for many residues; however, the overall distributions of centrality values across the protein structure were largely unchanged. In fact, many distributions were indistinguishable. One notable exception was the difference between 000-RNR and dAT0-RNR. Moreover, results from this investigation suggested that closeness centrality primarily identifies amino acids that are physically central to the three-dimensional structure. It had been previously reported that closeness centrality identified residues that belong to an enzyme's catalytic active site. However, the results here suggested that this may only be true for "typical" enzymes where the active site is physically central to the protein. Furthermore, our results also showed that the degree distributions in the PSNs are not scale-free, or power-law. This implies that random graph that best 'models' PSNs is either the ER or WS. In summary, this work demonstrated that the centrality distributions for the representative protein systems are invariant under ligand binding for the four classical network centralities, despite substantial shifts in centrality values for individual residues induced upon ligand-binding.

4.2 Future Work

The results generated here can be extended to the future study of protein dynamic networks (PDNs). The PDN is a set of principal components provides information regarding the magnitude and synchronization of residue fluctuations about the mean contact. The initial principal components describe larger, global fluctuations while higher order PCs describe smaller, localized motions. Additionally, these fluctuations may be positive or negative indicating that contacts are either correlated or anti-correlated. In the short-term, integrating an analysis of the PSNs with the PDNs will allow us to investigate how correlated anti-correlated motion is associated with residues in the top 16% by centrality. This will enable us to investigate the potential role of this subset of residues in the dynamics and function of the protein. In the long term, this may allow us to build a network analytical pipeline that will provide insight into allosteric regulation of protein dynamics and function.

Bibliography

[1] Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3):291–294. 22

[2] Allain, A., Chauvot de Beauchene, I., Langenfeld, F., Guarracino, Y., Laine, E., and Tchertanov, L. (2014). Allosteric pathway identification through network analysis: from molecular dynamics simulations to interactive 2D and 3D graphs. *Faraday Discuss.*, 169:303–321. 2

[3] Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., and Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, 344(4):1135–1146. 3, 36

[4] Atilgan, A. R., Akan, P., and Baysal, C. (2004). Small-world communication of residues and significance for protein dynamics. *Biophys. J.*, 86(1 Pt 1):85–91. 36

[5] Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512. 33

[6] Barzel, B. and Barabasi, A. L. (2013). Universality in network dynamics. *Nat Phys*, 9. 6

[7] Bassett, D. S., Zurn, P., and Gold, J. I. (2018). On the nature and use of models in network neuroscience. *Nat. Rev. Neurosci.*, 19(9):566–578. 6

[8] Bhalla, U. S. and Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387. 6

[9] Breitling, R. (2010). What is systems biology? *Frontiers in Physiology*, 1:159. 6

[10] Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nat Commun*, 10(1):1017. 33

[11] Buchenberg, S., Sittel, F., and Stock, G. (2017). Time-resolved observation of protein allosteric communication. *Proc. Natl. Acad. Sci. U.S.A.*, 114(33):E6804–E6811. 1

[12] Chea, E. and Livesay, D. R. (2007). How accurate and statistically robust are catalytic site predictions based on closeness centrality? *BMC Bioinformatics*, 8:153. 36

[13] Chi, P. B. and Liberles, D. A. (2016). Selection on protein structure, interaction, and sequence. *Protein Sci.*, 25(7):1168–1178. 2

[14] Clarke, D., Sethi, A., Li, S., Kumar, S., Chang, R. W. F., Chen, J., and Gerstein, M. (2016). Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. *Structure*, 24(5):826–837. 2

[15] Cusack, M. P., Thibert, B., Bredesen, D. E., and Del Rio, G. (2007). Efficient identification of critical residues based only on protein structure by network analysis. *PLoS ONE*, 2(5):e421. 36

[16] De Ruvo, M., Giuliani, A., Paci, P., Santoni, D., and Di Paola, L. (2012). Shedding light on protein-ligand binding by graph theory: the topological nature of allostery. *Biophys. Chem.*, 165-166:21–29. 3

[17] del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006a). Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein Sci.*, 15(9):2120–2128. 3, 36

[18] del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006b). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol.*, 2:2006.0019. 36

[19] Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., and Giuliani, A. (2013). Protein contact networks: an emerging paradigm in chemistry. *Chem. Rev.*, 113(3):1598–1613. 1, 2

[20] Doncheva, N. T., Klein, K., Domingues, F. S., and Albrecht, M. (2011). Analyzing and visualizing residue networks of protein structures. *Trends Biochem. Sci.*, 36(4):179–182. 6

[21] Doshi, U., Holliday, M. J., Eisenmesser, E. Z., and Hamelberg, D. (2016). Dynamical network of residue-residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. *Proc. Natl. Acad. Sci. U.S.A.*, 113(17):4735–4740. 1, 3

[22] Fokas, A. S., Cole, D. J., Ahnert, S. E., and Chin, A. W. (2016). Residue Geometry Networks: A Rigidity-Based Approach to the Amino Acid Network and Evolutionary Rate Analysis. *Sci Rep*, 6:33213. 3, 36, 39

[23] Glasscock, C. J., Lucks, J. B., and DeLisa, M. P. (2016). Engineered Protein Machines: Emergent Tools for Synthetic Biology. *Cell Chem Biol*, 23(1):45–56. 22

[24] Greene, L. H. and Higman, V. A. (2003). Uncovering network systems within protein structures. *J. Mol. Biol.*, 334(4):781–791. 6

[25] Gros, C. (2015). *Complex and Adaptive Dynamical Systems: A Primer*. Springer. 10, 16, 43

[26] Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786. 2

[27] Hofer, A., Crona, M., Logan, D. T., and Sjoberg, B. M. (2012). DNA building blocks: keeping control of manufacture. *Crit. Rev. Biochem. Mol. Biol.*, 47(1):50–63. 23

[28] Huang, Z., Mou, L., Shen, Q., Lu, S., Li, C., Liu, X., Wang, G., Li, S., Geng, L., Liu, Y., Wu, J., Chen, G., and Zhang, J. (2014). ASD v2.0: updated content and novel features focusing on allosteric regulation. *Nucleic Acids Res.*, 42(Database issue):D510–516. 1

[29] Johnson, Q. R., Lindsay, R. J., and Shen, T. (2018). CAMERRA: An analysis tool for the computation of conformational dynamics by evaluating residue-residue associations. *J Comput Chem*, 39(20):1568–1578. 2, 24

[30] Karain, W. I. and Qaraeen, N. I. (2015). Weighted protein residue networks based on joint recurrences between residues. *BMC Bioinformatics*, 16:173. 3

[31] Karain, W. I. and Qaraeen, N. I. (2017). The adaptive nature of protein residue networks. *Proteins*, 85(5):917–923. 3

[32] Khuri, S. and Wuchty, S. (2015). Essentiality and centrality in protein interaction networks revisited. *BMC Bioinformatics*, 16:109. 39

[33] Kolberg, M., Strand, K. R., Graff, P., and Andersson, K. K. (2004). Structure, function, and mechanism of ribonucleotide reductases. *Biochim. Biophys. Acta*, 1699(1-2):1–34. 23

[34] Krishnan, A., Zbilut, J. P., Tomita, M., and Giuliani, A. (2008). Proteins as networks: usefulness of graph theory in protein science. *Curr. Protein Pept. Sci.*, 9(1):28–38. 1

[35] Li, Y., Li, G., Wen, Z., Yin, H., Hu, M., Xiao, J., and Li, M. (2011). Novel feature for catalytic protein residues reflecting interactions with other residues. *PLoS ONE*, 6(3):e16932. 36

[36] Lindsay, R. J., Siess, J., Lohry, D. P., McGee, T. S., Ritchie, J. S., Johnson, Q. R., and Shen, T. (2018). Characterizing protein conformations by correlation analysis of coarse-grained contact matrices. *J Chem Phys*, 148(2):025101. 2

[37] Mall, R., Cerulo, L., Bensmail, H., Iavarone, A., and Ceccarelli, M. (2017). Detection of statistically significant network changes in complex biological networks. *BMC Syst Biol*, 11(1):32. 5

[38] Martinez, J. H., Lopez, M. E., Ariza, P., Chavez, M., Pineda-Pardo, J. A., Lopez-Sanz, D., Gil, P., Maestu, F., and Buldu, J. M. (2018). Functional brain networks reveal the existence of cognitive reserve and the interplay between network topology and dynamics. *Sci Rep*, 8(1):10525. 6

[39] Negre, C. F. A., Morzan, U. N., Hendrickson, H. P., Pal, R., Lisi, G. P., Loria, J. P., Rivalta, I., Ho, J., and Batista, V. S. (2018). Eigenvector centrality for characterization of protein allosteric pathways. *Proc. Natl. Acad. Sci. U.S.A.*, 115(52):E12201–E12208. 3, 54

[40] Newman, M. (2010). *Networks: An Introduction*. Oxford University Press. 10, 16, 43, 52

[41] Novinec, M. (2017). Computational investigation of conformational variability and allostery in cathepsin K and other related peptidases. *PLoS ONE*, 12(8):e0182387. 1

[42] Park, K. and Kim, D. (2011). Modeling allosteric signal propagation using protein structure networks. *BMC Bioinformatics*, 12 Suppl 1:S23. 1, 2

[43] Pham, B., Arons, A. B., Vincent, J. G., Fernandez, E. J., and Shen, T. (2019a). Regulatory Mechanics of Constitutive Androstane Receptor: Basal and Ligand-directed Actions. *J Chem Inf Model*. 23, 40, 54

[44] Pham, B., Lindsay, R. J., and Shen, T. (2019b). Effector-Binding-Directed Dimerization and Dynamic Communication between Allosteric Sites of Ribonucleotide Reductase. *Biochemistry*, 58(6):697–705. 24

[45] Pincus, D., Resnekov, O., and Reynolds, K. A. (2017). An evolution-based strategy for engineering allosteric regulation. *Phys Biol*, 14(2):025002. 2

[46] Piraveenan, M., Prokopenko, M., and Hossain, L. (2013). Percolation centrality: quantifying graph-theoretic impact of nodes during percolation in networks. *PLoS ONE*, 8(1):e53095. 2

[47] Putz, I. and Brock, O. (2017). Elastic network model of learned maintained contacts to predict protein motion. *PLoS ONE*, 12(8):e0183889. 26, 39

[48] Rajula, H. S. R., Mauri, M., and Fanos, V. (2018). Scale-free networks in metabolomics. *Bioinformation*, 14(3):140–144. 6

[49] Reynolds, K. A., McLaughlin, R. N., and Ranganathan, R. (2011). Hot spots for allosteric regulation on protein surfaces. *Cell*, 147(7):1564–1575. 2

[50] Ribeiro, A. A. and Ortiz, V. (2014). Determination of Signaling Pathways in Proteins through Network Theory: Importance of the Topology. *J Chem Theory Comput*, 10(4):1762–1769. 1, 2

[51] Salamanca Viloria, J., Allega, M. F., Lambrughi, M., and Papaleo, E. (2017). An optimal distance cutoff for contact-based Protein Structure Networks using side-chain centers of mass. *Sci Rep*, 7(1):2838. 36

[52] Saldano, T. E., Monzon, A. M., Parisi, G., and Fernandez-Alberti, S. (2016). Evolutionary Conserved Positions Define Protein Conformational Diversity. *PLoS Comput. Biol.*, 12(3):e1004775. 2

[53] Schueler-Furman, O. and Wodak, S. J. (2016). Computational approaches to investigating allostery. *Curr. Opin. Struct. Biol.*, 41:159–171. 1

[54] Shan, L., Vincent, J., Brunzelle, J. S., Dussault, I., Lin, M., Ianculescu, I., Sherman, M. A., Forman, B. M., and Fernandez, E. J. (2004). Structure of the murine constitutive androstane receptor complexed to androstenol: a molecular basis for inverse agonism. *Mol. Cell*, 16(6):907–917. 23

[55] Sheftel, S., Muratore, K. E., Black, M., and Costanzi, S. (2013). Graph analysis of Î2 adrenergic receptor structures: a "social network" of GPCR residues. *In Silico Pharmacol*, 1:16. 3

[56] Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E., and Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–332. 36

[57] Song, C., Havlin, S., and Makse, H. A. (2005). Self-similarity of complex networks. *Nature*, 433(7024):392–395. 6

[58] Spirin, V. and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U.S.A.*, 100(21):12123–12128. 6

[59] Stetz, G. and Verkhivker, G. M. (2017). Computational Analysis of Residue Interaction Networks and Coevolutionary Relationships in the Hsp70 Chaperones: A Community-Hopping Model of Allosteric Regulation and Communication. *PLoS Comput. Biol.*, 13(1):e1005299. 39

[60] Suki, B., Bates, J. H., and Frey, U. (2011). Complexity and emergent phenomena. *Compr Physiol*, 1(2):995–1029. 6

[61] Thomas, W. C., Brooks, F. P., Burnim, A. A., Bacik, J. P., Stubbe, J., Kaelber, J. T., Chen, J. Z., and Ando, N. (2019). Convergent allostery in ribonucleotide reductase. *Nat Commun*, 10(1):2653. 23

[62] Topirceanu, A., Udrescu, M., and Marculescu, R. (2018). Weighted Betweenness Preferential Attachment: A New Mechanism Explaining Social Network Formation and Evolution. *Sci Rep*, 8(1):10871. 3

[63] van Steen, M. (2010). *Graph Theory and Complex Networks: An Introduction*. Maarten van Steen. 16

[64] Wang, W. B., Liang, Y., Zhang, J., Wu, Y. D., Du, J. J., Li, Q. M., Zhu, J. Z., and Su, J. G. (2018). Energy transport pathway in proteins: Insights from non-equilibrium molecular dynamics with elastic network model. *Sci Rep*, 8(1):9487. 3

[65] Wang, Y., Di, Z., and Fan, Y. (2011). Identifying and characterizing nodes important to community structure using the spectrum of the graph. *PLoS ONE*, 6(11):e27418. 5

[66] Warnmark, A., Treuter, E., Wright, A. P., and Gustafsson, J. A. (2003). Activation functions 1 and 2 of nuclear receptors: molecular strategies for transcriptional activation. *Mol. Endocrinol.*, 17(10):1901–1909. 40, 54

[67] Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G., and Shen, B. (2014). The construction of an amino acid network for understanding protein structure and function. *Amino Acids*, 46(6):1419–1439. 36

[68] Zweig, K. A. (2018). *Network Analysis Literacy A Practical Approach to the Analysis of Networks*. Springer Wien. ix, 10, 14

Vita

David Foutch is from Keyesport, IL. He graduated from Centralia High School in 1989. Following High School he began working as an entrepreneur. After many detours and failed attempts, he eventually started Allied Electrical Services Inc. He was a successful electrical contractor for over a decade becoming a licensed master electrician. In the Fall of 2006 he began attending Kaskaskia Community college. After finishing an Associate's degree, he was accepted to Southern Illinois University Carbondale in the Fall of 2008 where his interest in computational neuroscience lead him to finish Bachelor of Arts degrees in Psychology and Mathematics in the Spring of 2013. He continued his interests in computational sciences after being accepted to the Genome Science and Technology program at the University of Tennessee, Knoxville in the Fall of 2015. Foutch earned his Masters of Science degree in May 2020.