



SCHOOL of  
GRADUATE STUDIES  
EAST TENNESSEE STATE UNIVERSITY

East Tennessee State University  
**Digital Commons @ East  
Tennessee State University**

---

Electronic Theses and Dissertations

Student Works

---


5-2017

# Differentiating Between a Protein and its Decoy Using Nested Graph Models and Weighted Graph Theoretical Invariants

Hannah E. Green

*East Tennessee State University*

Follow this and additional works at: <https://dc.etsu.edu/etd>

 Part of the [Discrete Mathematics and Combinatorics Commons](#), and the [Other Applied Mathematics Commons](#)

---

## Recommended Citation

Green, Hannah E., "Differentiating Between a Protein and its Decoy Using Nested Graph Models and Weighted Graph Theoretical Invariants" (2017). *Electronic Theses and Dissertations*. Paper 3248. <https://dc.etsu.edu/etd/3248>

This Thesis - Open Access is brought to you for free and open access by the Student Works at Digital Commons @ East Tennessee State University. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons @ East Tennessee State University. For more information, please contact [digilib@etsu.edu](mailto:digilib@etsu.edu).

Differentiating Between a Protein and its Decoy Using Nested Graph Models and  
Weighted Graph Theoretical Invariants

---

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Hannah Green

May 2017

---

Debra Knisley, Ph.D., Chair

Jeff Knisley, Ph.D.

Teresa Haynes, Ph.D.

Keywords: graph theory, computational biology, proteins, invariants.

## ABSTRACT

Differentiating Between a Protein and its Decoy Using Nested Graph Models and  
Weighted Graph Theoretical Invariants

by

Hannah Green

To determine the function of a protein, we must know its 3-dimensional structure, which can be difficult to ascertain. Currently, predictive models are used to determine the structure of a protein from its sequence, but these models do not always predict the correct structure. To this end we use a nested graph model along with weighted invariants to minimize the errors and improve the accuracy of a predictive model to determine if we have the correct structure for a protein.

Copyright by Hannah Green 2017

All Rights Reserved

## ACKNOWLEDGMENTS

I would like to start by thanking my committee members. Dr. Debra Knisley for the idea behind the project, Dr. Jeff Knisley for the creation of the nested graph model, and Dr. Teresa Haynes for inspiring a love of graph theory in me.

I would to thank the University Advancement office for giving me a job as an undergraduate that helped me get through school, and then again by offering me a graduate assistantship after I decided to get my master's degree.

Lastly, I would like to thank the friends I have made over the past few years, who have shown acceptance and given me support when I needed it the most, and who have encouraged and aided me in my pursuits.

## TABLE OF CONTENTS

ABSTRACT . . . . .	2
ACKNOWLEDGMENTS . . . . .	4
LIST OF TABLES . . . . .	8
LIST OF FIGURES . . . . .	10
1 BACKGROUND . . . . .	11
1.1 Proteins . . . . .	11
1.2 Graph Theory . . . . .	14
1.3 Biological Networks . . . . .	19
1.4 Computational Methods . . . . .	20
2 METHODS . . . . .	22
3 RESULTS . . . . .	25
4 CONCLUSIONS . . . . .	28
BIBLIOGRAPHY . . . . .	29
Appendices . . . . .	33
A Domain and Top Level Graphs of Each Protein and Decoy . . . . .	33
A.1 2cro . . . . .	33
A.2 2croon1sn3 . . . . .	36
A.3 2croon2ci2 . . . . .	39
A.4 2ci2 . . . . .	42
A.5 2ci2on1sn3 . . . . .	45
A.6 2ci2on2cro . . . . .	48
A.7 1sn3 . . . . .	51

A.8	1sn3on2ci2 . . . . .	54
A.9	1sn3on2cro . . . . .	57
B	Tables of vertex weights and weighted invariants for Domain and Top Level Graphs . . . . .	60
VITA	. . . . .	83

## LIST OF TABLES

1	Amino Acid Descriptors . . . . .	23
2	Proteins with multiple decoys . . . . .	25
3	Amino Acid Descriptors (Part 1) . . . . .	60
4	Amino Acid Descriptors (Part 2) . . . . .	61
5	Amino Acid Descriptors (Part 3) . . . . .	62
6	Amino Acid Descriptors (Part 4) . . . . .	63
7	2cro Top Level (Part 1) . . . . .	64
8	2cro Top Level (Part 2) . . . . .	65
9	2croon1sn3 Top Level (Part 1) . . . . .	66
10	2croon1sn3 Top Level (Part 2) . . . . .	67
11	2croon2ci2 Top Level (Part 1) . . . . .	68
12	2croon2ci2 Top Level (Part 2) . . . . .	69
13	2ci2 Top Level (Part1) . . . . .	70
14	2ci2 Top Level (Part2) . . . . .	71
15	2ci2on1sn3 Top Level (Part1) . . . . .	72
16	2ci2on1sn3 Top Level (Part2) . . . . .	73
17	2ci2on2cro Top Level (Part1) . . . . .	74
18	2ci2on2cro Top Level (Part2) . . . . .	75
19	1sn3 Top Level (Part1) . . . . .	76
20	1sn3 Top Level (Part2) . . . . .	77
21	1sn3on2ci2 Top Level (Part1) . . . . .	78
22	1sn3on2ci2 Top Level (Part2) . . . . .	79



23	1sn3on2cro Top Level (Part1) . . . . .	80
24	1sn3on2cro Top Level (Part2) . . . . .	81
25	Combinded PCA Values . . . . .	82

## LIST OF FIGURES

1	Amino Acid Base Structure . . . . .	12
2	A simple graph . . . . .	15
3	A complete graph and a cycle . . . . .	15
4	An isomorphism of $C_5$ . . . . .	16
5	Illustration of a dominating set . . . . .	17
6	Illustration of a vertex cover set . . . . .	17
7	Combined Dendrogram . . . . .	26
8	Combined dendrogram with 1sn3 and related decoys omitted . . . . .	27
9	Domain Graphs 1 through 4 of protein 2cro . . . . .	33
10	Domain Graphs 5 through 7 and Top Level Graph for protein 2cro . . . . .	34
11	2cro Contact Map . . . . .	35
12	Domain Graphs 1 through 4 of Decoy 2croon1sn3 . . . . .	36
13	Domains 5 through 7 and Top Level for 2croon1sn3 . . . . .	37
14	2croon1sn3 Contact Map . . . . .	38
15	Domain Graphs 1 through 4 of Decoy 2croon2ci2 . . . . .	39
16	Domain Graphs 5 through 7 and Top Level Graph for protein 2cro . . . . .	40
17	2croon2ci2 Contact Map . . . . .	41
18	Domain Graphs 1 through 4 of protein 2ci2 . . . . .	42
19	Domain Graphs 5 through 7 and Top Level Graph for protein 2ci2 . . . . .	43
20	2ci2 Contact Map . . . . .	44
21	Domain Graphs 1 through 4 of decoy 2ci2on1sn3 . . . . .	45
22	Domains 5 through 7 and Top Level for 2ci2on1sn3 . . . . .	46

23	2ci2on1sn3 Contact Map . . . . .	47
24	Domain Graphs 1 through 4 of decoy 2ci2on2cro . . . . .	48
25	Domains 5 through 7 and Top Level for 2ci2on2cro . . . . .	49
26	2ci2on2cro Contact Map . . . . .	50
27	Domain Graphs 1 through 4 of protein 1sn3 . . . . .	51
28	Domain Graphs 5 through 7 and Top Level Graph for protein 1sn3 . . . . .	52
29	1sn3 Contact Map . . . . .	53
30	Domain Graphs 1 through 4 of decoy 1sn3on2ci2 . . . . .	54
31	Domains 5 through 7 and Top Level for 1sn3on2ci2 . . . . .	55
32	1sn3on2ci2 Contact Map . . . . .	56
33	Domain Graphs 1 through 4 of decoy 1sn3on2cro . . . . .	57
34	Domains 5 through 7 and Top Level for 1sn3on2cro . . . . .	58
35	1sn3on2cro Contact Map . . . . .	59

## 1 BACKGROUND

Our goal is to distinguish correct protein structures from incorrect protein structures using graph theoretical methods. Thus, it is necessary to have an understanding of proteins (including their structure, function, and importance), graph theory, and the way in which graph theory can be used to model proteins. In addition, a brief overview of some of the other methods used in this research will be given.

### 1.1 Proteins

An *amino acid* is a molecule composed of an amino group and a carboxyl (acid) group, which form the “base”, and residue group [1, 2, 3, 4, 5]. In any two amino acids, the base is the same while the residue group can differ, and so the residue group determines the amino acid. While there are many forms the residue group can take, there are 20 residue groups that are crucial in the formation of proteins. We call the amino acids determined by those 20 residues, the protogenic amino acids [3, 4, 5]. The structure of a general amino acid can be seen below in Figure 1, where  $R$  represents the residue group.

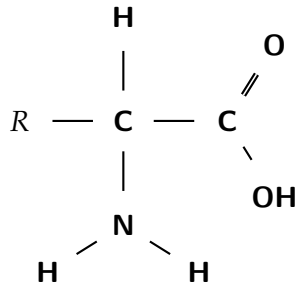


Figure 1: Amino Acid Base Structure

When the amino group of an amino acid and the carboxyl group of another react it forms what is called a peptide bond. A *protein* is defined “a polypeptide chain of amino acid residues linked together in a definite sequence” [6]. Each distinct sequence of amino acids forms a unique protein, with the polypeptide chain forming the backbone of the protein [1, 3, 6]. The sequence is the first of four layers of protein structure, and so it is also called the *primary structure* of the protein. The other three layers are the secondary, tertiary, and quaternary structures. The *secondary structure* consists of  $\alpha$ -helices and  $\beta$ -strands, the *tertiary structure* is how the protein is folded in 3-dimensional space, and the *quaternary structure* is the common configuration of multiple proteins found in nature [3].  $\alpha$ -helices and  $\beta$ -strands are common structures found in every protein, they are determined by a definite subsequence of amino acids and form from hydrogen bonds within this subsequence. The tertiary structure of the protein is the most important layer in determining the function of a protein, as the 3-dimensional structure of a protein determines how it interacts with other substances [7, 1, 3, 4, 5].

The 3D structure of a protein is discovered by x-ray crystallography, a method

which uses the diffraction of x-rays by crystals to determine the structure of a molecule. Obtaining a crystal of a protein is a complicated and intricate process which requires a dependable source of a protein and the ability to purify this sufficiently enough to produce a usable solution (“high quality and homogeneous”) [8, 9]. This solution must then be crystallized, which is a complicated and intricate process in itself, with no guarantee of producing a usable crystal. This makes x-ray crystallography a long and slow, but necessary, process to have a 100% accurate 3D model of a protein.

Recent research into protein structure has provided evidence that while the size of the sequence space of proteins is immense, the structure space is a small finite set. This is to say that a protein with an unknown structure is likely to have a structure that has already been observed. Since we know that proteins with similar sequences tend to have similar structures, we can predict the unknown structures of proteins using the known structures of similar sequences and filling in the gaps with predictive modeling [10]. However, it is possible for these predictive models to be wrong. To aid in this issue, Samudrala and Levitt created the well known and often used “Decoys R Us” database used to house several decoys of a protein [10, 11, 12, 13]. These decoys are common errors produced by predictive models, and so are useful in determining the accuracy of any predictive model of protein structure. The database is a large data set that can be broken down into three main subsets: multiple, single, and loop. For our project, we will focus on the single subset of the decoy set, the purpose of which is to separate valid protein structures from invalid structure. This set can be broken down further into the specific type

of decoy, which for us will be the misfold decoy set [14].

## 1.2 Graph Theory

First, we must define the basic concept of a graph. A *graph*, or *network*,  $G$ , is a finite nonempty set  $V$  of objects called *vertices*, also called *nodes*, together with a possibly empty set,  $E$  of 2-element subsets of  $V$  called *edges*. For notational purposes, to refer to a graph,  $G$ , with vertex set  $V$  and edge set  $E$ , we generally say  $G = (V, E)$ . A sample graph can be seen in figure 2. Let  $G = (V, E)$  and let  $m, n$  be nonnegative integers, we say that  $|V| = n$  and  $|E| = m$ , that is,  $G$  has  $n$  vertices and  $m$  edges. We call  $n$  the *order* of  $G$ , and we call  $m$  the *size* of  $G$ . If an edge exists between two vertices, we say those two vertices are *adjacent*; consider  $u, v$  vertices of a graph  $G$ , if  $u$  and  $v$  are adjacent, then an edge exists between them, and we call this edge  $uv$ . A vertex and the edge connecting it to another vertex are said to be *incident*. Let  $v$  be a vertex in a graph  $G$ , the number of vertices adjacent to  $v$  is the degree of vertex  $v$ , usually denoted  $\deg(v)$ . The maximum degree of vertices in  $G$  is denoted  $\Delta(G)$ , and the minimum degree is denoted  $\delta(G)$ . Consider  $G = (V, E)$  and let  $v \in V(G)$ , the *closed neighborhood* of vertex  $v$ , denoted  $N[v]$ , is a subset of  $V$  which contains the vertex  $v$  and all vertices adjacent to  $v$ ; the *open neighborhood* of  $v$ ,  $N(v)$  is  $N[v] \setminus \{v\}$  [15].

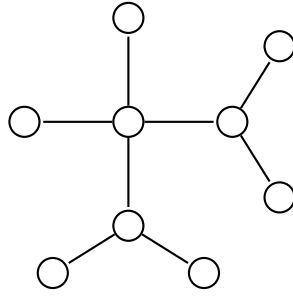
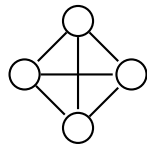
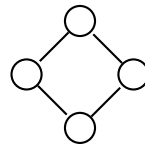


Figure 2: A simple graph

Next, we define some special classes of graphs. For  $n$  a nonnegative integer, a *complete graph*, denoted  $K_n$ , is a graph on  $n$  vertices in which every two distinct vertices are adjacent. For  $n \geq 3$ , a *cycle*, denoted  $C_n$ , is a graph on  $n$  vertices and  $n$  edges, with vertices that can be labeled  $v_1, v_2, \dots, v_n$  and edges  $v_1v_n$  and  $v_1v_{i+1}$  for  $i = 1, 2, \dots, n - 1$ . Examples of both a complete graph and a cycle are given below. Consider two graphs,  $H$  and  $G$ , if  $V(H) \subseteq V(G)$  and  $E(H) \subseteq E(G)$ , then, we say  $H$  is a *subgraph* of  $G$ , denoted  $H \subseteq G$ .  $H$ , a subgraph of  $G$ , is said to be *induced* if for each  $v \in V(H)$ , all edges incident with  $v$  in  $G$  are present in  $H$ . Note that for any positive integer  $n$ ,  $K_n$  is an induced subgraph of  $K_{n+1}$  [15].



(a)  $K_4$



(b)  $C_4$

Figure 3: A complete graph and a cycle

We say two graphs,  $G$  and  $H$  are *isomorphic* if there exists a bijective function  $\varphi : V(G) \rightarrow V(H)$  such that  $u$  and  $v$  are adjacent in  $G$  if and only if  $\varphi(u)$  and



$\varphi(v)$  are adjacent in  $H$  [15]. We call the function  $\varphi$  an isomorphism. This is to say that isomorphic graphs have the same structure, and thus  $\varphi$  preserves structure. A graph theoretical *invariant* is a property of graph that does not change under an isomorphism. The simplest invariant is the degree of a vertex. This is made clear when one realizes that the degree relies on the structure of the graph, and as such, does not change under isomorphism (this is illustrated in the example below).

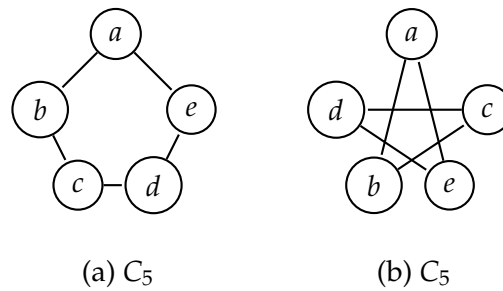


Figure 4: An isomorphism of  $C_5$

Along with the degree of a vertex, another invariant is the domination number of a graph. We say that vertex  $v \in V(G)$  *dominates* its closed neighborhood. Thus, a *dominating* set  $S \subset V(G)$ , is a set such that every vertex of  $G$  is dominated by at least one vertex in set  $S$ . The minimum cardinality of such dominating sets of  $G$  is called the *domination number* of  $G$ , denoted  $\gamma(G)$ . In the graph given below, the black vertices make a minimum dominating set, thus we can see that  $\gamma(G) = 3$ . It is important to note that while the vertices below do form a minimum domination set, this is not the only minimum dominating set.

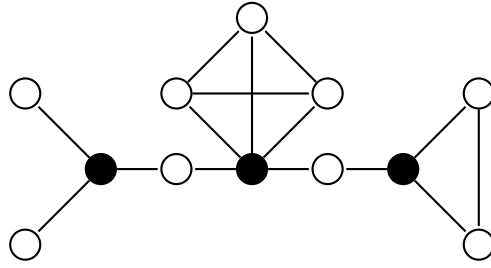


Figure 5: Illustration of a dominating set

Another invariant is the clique number of a graph. A *clique* of  $G$  is a complete subgraph of  $G$ . The *clique number* of a graph  $G$ , denoted  $\omega(G)$ , is the order of the largest clique (complete subgraph) of  $G$ . In the graph above,  $\omega(G) = 4$ . The last invariant we will discuss is the vertex covering number. A vertex and an incident edge are said to *cover* each other. A *vertex cover* is a set of vertices of a graph that cover all the edges of that graph. The *vertex covering number*, denoted  $\beta(G)$ , is the minimum cardinality of vertex covers. In the example below, one can see that the vertex cover set has cardinality 3, thus  $\beta(G) = 3$ .

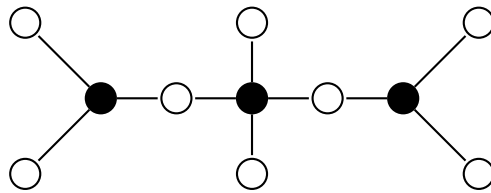


Figure 6: Illustration of a vertex cover set

For any graph, a real number can be assigned to each edge. This is called a *weighted graph* [15]. For our purposes, we want to adjust this definition and assign a real number to each vertex, creating a *vertex weighted graph*. We can now adjust the invariants that were previously discussed to incorporate weights, which will

be called *weighted invariants*. All of these weighted invariants will use the idea of the sum of weights of a set of vertices. Once again, the degree of a vertex can be considered the simplest invariant. Above, the degree of a vertex was determined by the number of nodes adjacent to it, or, stated another way, the cardinality of its open neighborhood. To apply the idea of vertex weights to the degree of a vertex, we simply take the sum of the weights in the open neighborhood of a vertex, and we call this the *vertex weighted degree*.

Next, consider domination as discussed above. It is possible to have multiple minimum dominating sets, which are indistinguishable from one another in a simple graph with no other information. However, when we have weighted vertices, we can further quantify these sets. Across all dominating sets with minimum cardinality, we can sum the weight of the vertices in these sets to weight the sets themselves, and we now have a way of distinguishing between the sets of the same cardinality. We call the weight of such a set a *weighted domination number*. However, unlike above we do not only consider the minimum this time, as it may be more applicable in some circumstances to calculate the maximum weights of such sets. Note that when using the maximum, we only want to find the maximum weighted minimum dominating set, as, by definition, the maximum dominating set is the entire graph.

To weight the clique number we find the largest clique and sum the weight of its nodes. When we have multiple cliques of the same size, it is beneficial to determine both the maximum and minimum weighted cliques. The weighted vertex cover number is similar to the weighted domination number.

For a highly connected network with many nodes, it can be hard to determine what certain invariants are. Thus, it is beneficial to break it down into smaller graphs that are more manageable. A *nested graph* is a graph where node itself represents a graph. This allows us to break down the original graph into smaller subgraphs, called *domains*, and then create another graph, which we will call a *top level graph*, where each vertex represents one of these domains. For our purposes, adjacency in the top level graph will be determined by the amount of edges that connect one domain to another in the original graph; we will set a threshold of edges and if there exist more edges than the threshold, we will say the domains are adjacent [16].

### 1.3 Biological Networks

Notice that, if the double bond is ignored, the chemical diagram in Figure 1 and the simple graph in Figure 2 are the same. Thus, we can say that chemical diagrams are networks, where atoms are the nodes and bonds are the edges. We can extend this network approach from chemical diagrams to entire biological systems. In fact, this approach to biological systems in terms of networks is not new and has been in use for many years to determine different interactions in these systems. What is new, however, is using a network to study the structure of biological systems [17]. In 2002, Vishveshwara et al compiled a survey of the different ways networks have been used to study proteins, with most applications being used to determine the structure of proteins. Once the graph for a protein's structure was acquired, different graph theoretical measures were able to be applied to analyze

each protein [18]. The application of graph theory to the structure of proteins is still an active, booming field of research. The critical assessment of methods of protein structure prediction (CASP) is a recurring “experiment to assess the state of the art in protein structure prediction” [10, 19, 20]. As CASP is recurring and the models are ever improving, it becomes important to improve the way models are assessed for accuracy. For this, Chatterjee, Ghosh, and Vishveshwara created graph theoretic models of proteins, which were quantified by various measures, and used these models to train a support vector machine, which was in turn used to assess the accuracy of the most recent CASP models [10].

In 2013, Knisley, Knisley, and Herron used graph theory to model the protein CFTR, or cystic fibrosis membrane conductance regulator. Specifically, a nested graph model was used to model mutations of the NBD1 domain of this protein, with different graph theoretical measures and weights associated with each mutation [16].

## 1.4 Computational Methods

Although we are concerned with the 3-dimensional structure of a protein, we need the data we gather to be 2-dimensional; we need to know which protein we are dealing with and information about that protein. To tackle this, when we have data about each protein that has more than two dimensions, we will use principal component analysis to reduce the dimensionality of the data set down to one dimension. Principal component analysis (PCA) is a method that takes a dataset and determines which data points are the most important to the data set,

and which are largely irrelevant [21].

## 2 METHODS

For this project, I utilized a jupyter notebook, created by Dr. Jeff Knisley, that reads a PDB file of a protein and generates a networkx graph of the protein structure, called the contact map, where each vertex represents an amino acid. After the contact map has been generated, it is possible to create a nested graph model for the protein. Recall, from above, that to create a nested graph from an existing graph, we take the domain graphs to be subgraphs of the existing graph; so, here, we need our domain graphs to be subgraphs of the contact map. To accomplish this, we partition the sequence of the protein into smaller intervals, with the size of each dependent on the number of intervals we wish to have. While each interval should ideally be of the same size, this is not always possible as we may not have a number of intervals that divides the sequence length. Since each vertex represents an amino acid, each interval is a set of vertices and, so, we take our domain graphs to be the subgraph of the contact map induced by each interval. For notational purposes, we refer to each domain graph as  $D_{\text{graphs}}[i, i]$ , where  $i$  is the interval containing the vertices of the graph. To avoid confusion, we shorten this notation to  $D_i$  when referring to the vertex in the top level graph that represents domain graph  $D_{\text{graphs}}[i, i]$ .

To determine adjacency in the top level graph, we first create a new set of graphs called joint graphs. Each joint graph,  $D[i, j]$  where  $i \neq j$ , is the induced subgraph of the contact map whose vertices come from intervals  $i$  and  $j$ . Clearly, both  $D_{\text{graphs}}[i, i]$  and  $D_{\text{graphs}}[j, j]$  are subgraphs of  $D_{\text{graphs}}[i, j]$ ; so, if more than two edges join vertices from  $D_{\text{graphs}}[i, i]$  and  $D_{\text{graphs}}[j, j]$ , we say that  $D_i$

and  $D_j$  are adjacent in the top level graph.

Table 1: Amino Acid Descriptors

nme	AA1	AA3	G	g	d	...	vanderWaal	...	EIIP
A	A	ALA	12	12	2	...	2.50E-02	...	3.73E-02
R	R	ARG	54	36	9	...	0.2	...	9.59E-02
N	N	ASN	42	24	5	...	0.1	...	3.60E-03
D	D	ASP	44	24	5	...	0.1	...	0.1263
C	C	CYS	12	12	3	...	0.1	...	8.29E-02
E	E	GLU	42	24	6	...	0.1	...	7.61E-02
Q	Q	GLN	44	24	6	...	0.1	...	5.80E-03
G	G	GLY	0	0	0	...	2.50E-02	...	5.00E-03
H	H	HIS	40	36	6	...	0.1	...	2.42E-02
I	I	ILE	24	24	5	...	0.19	...	0
L	L	LEU	36	24	5	...	0.19	...	0
K	K	LYS	38	24	8	...	0.2	...	3.71E-02
M	M	MET	44	24	6	...	0.19	...	8.23E-02
F	F	PHE	36	24	7	...	0.39	...	9.46E-02
P	P	PRO	24	12	4	...	0.17	...	1.98E-02
S	S	SER	12	12	3	...	2.50E-02	...	8.29E-02
T	T	THR	12	12	3	...	0.1	...	9.41E-02
W	W	TRP	62	48	9	...	0.56	...	5.48E-02
Y	Y	TYR	52	24	8	...	0.39	...	5.16E-02
V	V	VAL	12	12	3	...	0.15	...	5.70E-03

To weight our nodes in the domain graphs, we used many different descriptors, some of which can be seen above in Table 1 (the full table of descriptors is given in appendix 3). To provide weights for the vertices of the top level graph, we apply the four weighted invariants discussed in the previous chapter (degree, domination, clique, and vertex cover) to the domain graphs. For the sake of clarity, we select an arbitrary domain graph, say  $D_i$ , and an arbitrary descriptor, say  $x$ . A pre-existing script in the jupyter notebook that gave the weighted degree of



each vertex was altered into two functions, one that returned the maximum such weighted degree, and another that returned the minimum. For the other three invariants, existing networkx algorithms were slightly altered to find the applicable sets and then another function was written for each to sum the weight of these sets [22]. For each descriptor and each domain graph, the minimum and maximum degree, domination number, and clique number were found and stored in a dataframe, as was the minimum vertex covering number. This led to a total of 147 weights for each vertex in the top level graph. For the weighted invariants of the top level graph, we only used those weights produced by the same algorithm we wished to run; i.e., to get the minimum weighted domination number of the top level graph produced by weight  $x$ , we only used the minimum weighted domination number produced by weight  $x$  of the domain graph. After doing this for all invariants and descriptors, we were left with a table of size  $21 \times 7$ . We used the principal component analysis algorithm from scikit-learn to reduce the dimensionality of this dataset from  $21 \times 7$  to  $1 \times 7$  [23]. We reduced the dimensionality in such a way that we can produce a two dimensional data set, telling us which protein or decoy we have and information about that protein or decoy.

Once we had the principal component values for the protein and each decoy, MATLAB was then used to cluster the data and produce a dendrogram [24]

### 3 RESULTS

Table 2: Proteins with multiple decoys

Protein	Decoy1	Decoy2
2cro	2croon2ci2	2croon1sn3
2ci2	2ci2on2cro	2ci2on1sn3
1sn3	1sn3on2cro	1sn3on2ci2

To analyze and cluster proteins in a meaningful way, we need proteins with more than one misfold decoy. There are three such proteins in the Decoys R Us database, shown in the table above [14]. For each protein and decoy, we used the method described in the previous section. Since each of these has a sequence of length 65, we used five domains of 9 amino acids, and two domains of 10 amino acids for a total of seven domains. Given below is a MATLAB dendrogram for all of proteins and decoys together.

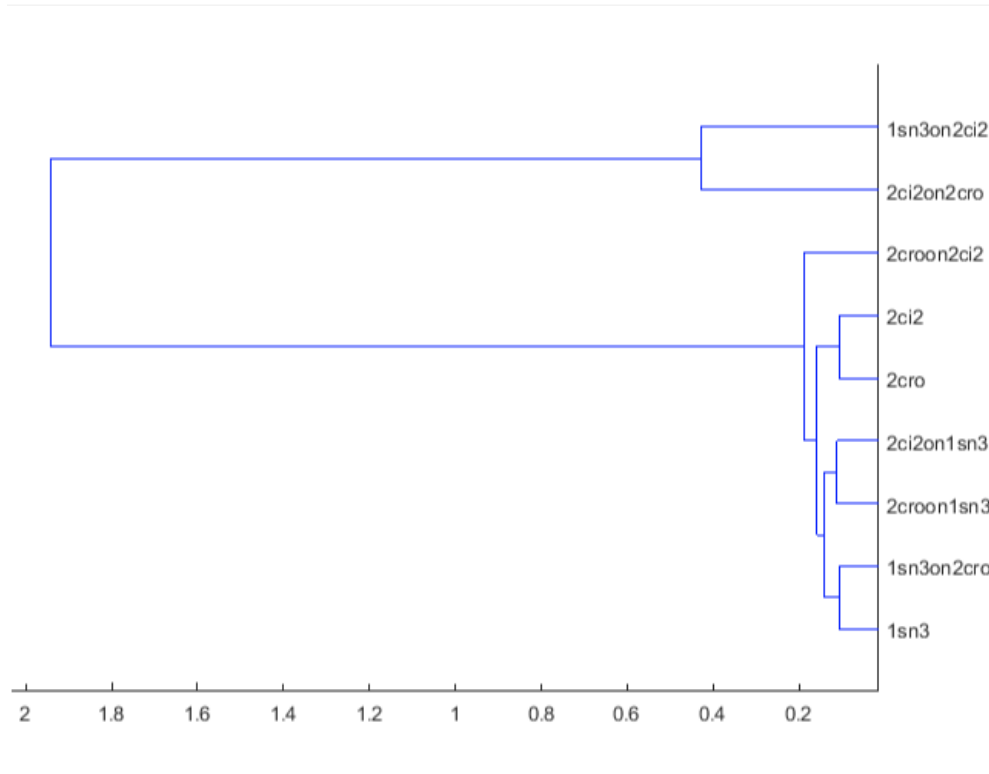


Figure 7: Combined Dendrogram

As can be seen, the proteins 2cro and 2ci2 have been grouped together, while the protein 1sn3 has been grouped with a decoy. According to the Protein Data Bank, 1sn3 is now an obsolete protein that has been replaced by 2sn3, and so its structure is not the correct structure of a valid protein. This is to say that 1sn3 is more similar to a decoy than it is to a valid protein.

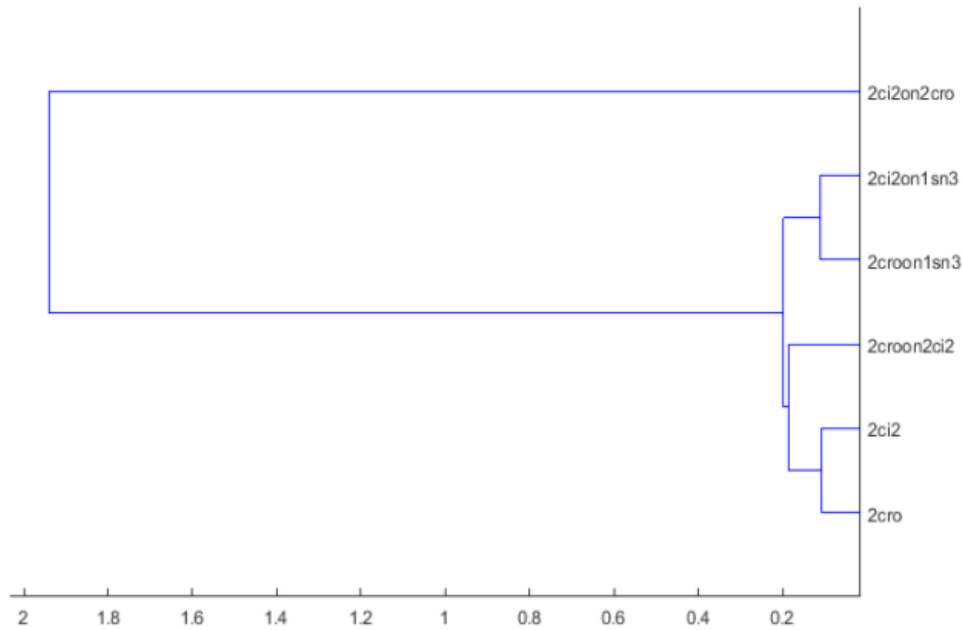


Figure 8: Combined dendrogram with 1sn3 and related decoys omitted

The next step was to remove the data for 1sn3 and its decoys and cluster the data again, producing a new dendrogram. As can be seen from the dendrogram in figure 8 above, the removal of the data for 1sn3 did not alter the results in any way. Thus, we have succeeded in our short term goal to separate valid protein structures from decoys.

## 4 CONCLUSIONS

We have demonstrated that a vertex weighted nested graph model was able to quantify the structure of a protein in such a way that a MATLAB clustering algorithm was able to separate the decoys and the valid proteins into different clusters.

Future work that remains to be done is to check this nested graph model for consistency among different clustering methods. If our method of quantification allows for correct and incorrect protein structures to be consistently separated, then we have created useful tool for determining and thus improve the accuracy of predictive models of proteins.

## BIBLIOGRAPHY

- [1] Gideon E Nelson, Gerald G Robinson, and Richard A Boolootian. *Fundamental Concepts of Biology*. Wiley, 1967.
- [2] John W Kimball. *Biology*. Addison-Wesley Pub. Co., 1965.
- [3] Gregory A Petsko and Dagmar Ringe. *Protein structure and function*. New Science Press, 2004.
- [4] Anders Liljas, Lars Liljas, Jure Piskur, Göran Lindblom, Poul Nissen, and Morten Kjeldgaard. *Textbook of structural biology*. World Scientific, 2009.
- [5] Joseph I Routh. *Introduction to Biochemistry*. W. B. Saunders Company, 1971.
- [6] Jane S Richardson. The anatomy and taxonomy of protein structure. *Advances in protein chemistry*, 34:167–339, 1981.
- [7] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. URL: [www.rcsb.org](http://www.rcsb.org).
- [8] MS Smyth and JHJ Martin. x ray crystallography. *Journal of Clinical Pathology*, 53(1):8, 2000.
- [9] Yigong Shi. A glimpse of structural biology through x-ray crystallography. *Cell*, 159(5):995–1014, 2014.

- [10] S Chatterjee, S Ghosh, and S Vishveshwara. Network properties of decoys and casp predicted models: a comparison with native protein structures. *Molecular BioSystems*, 9(7):1774–1788, 2013.
- [11] Jianhong Zhou, Wenying Yan, Guang Hu, and Bairong Shen. Amino acid network for the discrimination of native protein structures from decoys. *Current Protein and Peptide Science*, 15(6):522–528, 2014.
- [12] Britt Park and Michael Levitt. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *Journal of molecular biology*, 258(2):367–392, 1996.
- [13] Mathew C Lee and Yong Duan. Distinguish protein decoys by using a scoring function based on a new amber force field, short molecular dynamics simulations, and the generalized born solvent model. *Proteins: Structure, Function, and Bioinformatics*, 55(3):620–634, 2004.
- [14] Ram Samudrala and Michael Levitt. Decoys rus: a database of incorrect conformations to improve protein structure prediction. *Protein science*, 9(7):1399–1401, 2000.
- [15] Gary Chartrand, Linda Lesniak, and Ping Zhang. *Graphs & digraphs*, volume 39. CRC Press, 2010.
- [16] Debra J Knisley, Jeff R Knisley, and Andrew Cade Herron. Graph-theoretic models of mutations in the nucleotide binding domain 1 of the cystic fibrosis

transmembrane conductance regulator. *Computational Biology Journal*, 2013, 2013.

[17] Björn H Junker and Falk Schreiber. *Analysis of biological networks*, volume 2. John Wiley & Sons, 2011.

[18] Saraswathi Vishveshwara, KV Brinda, and N Kannan. Protein structure: insights from graph theory. *Journal of Theoretical and Computational Chemistry*, 1(01):187–211, 2002.

[19] John Moult, Krzysztof Fidelis, Adam Zemla, and Tim Hubbard. Critical assessment of methods of protein structure prediction (caspl)-round v. *Proteins: Structure, Function, and Bioinformatics*, 53(S6):334–339, 2003.

[20] John Moult, Krzysztof Fidelis, Andriy Kryshchak, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (caspl)round x. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):1–6, 2014.

[21] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[22] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, August 2008.

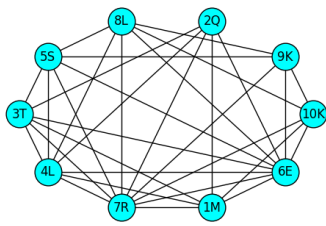


- [23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [24] MATLAB Users Guide. The mathworks. *Inc., Natick, MA*, 5:333, 1998.

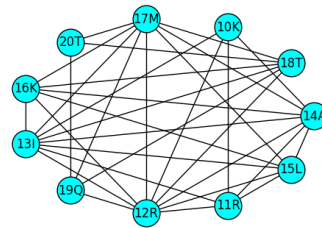
## APPENDICES

### A Domain and Top Level Graphs of Each Protein and Decoy

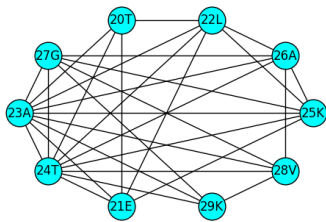
#### A.1 2cro



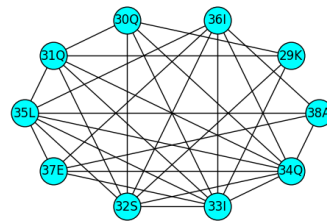
(a) 2cro D1



(b) 2cro D2

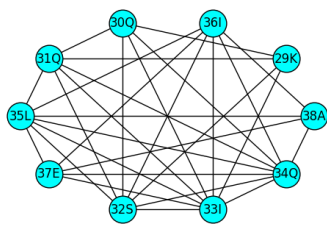


(c) 2cro D3

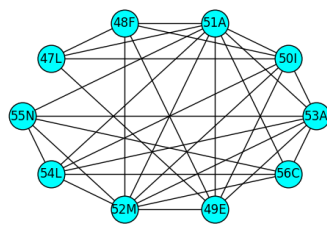


(d) 2cro D4

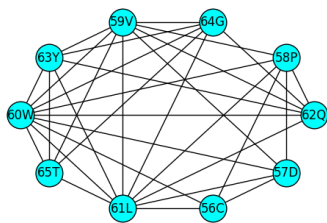
Figure 9: Domain Graphs 1 through 4 of protein 2cro



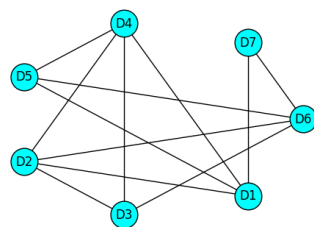
(a) 2cro D5



(b) 2cro D6



(c) 2cro D7



(d) 2cro Top Level Graph

Figure 10: Domain Graphs 5 through 7 and Top Level Graph for protein 2cro

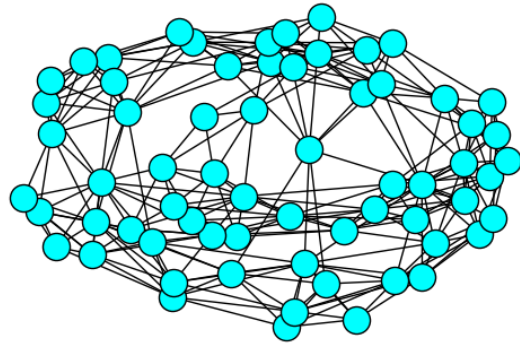
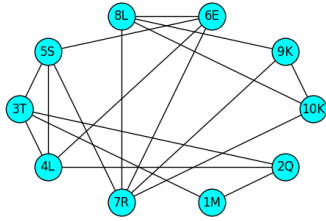
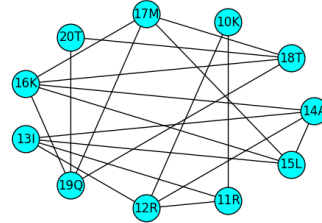


Figure 11: 2cro Contact Map

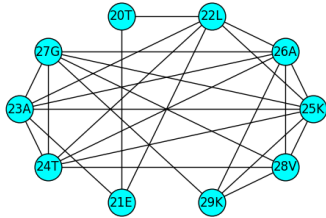
## A.2 2croon1sn3



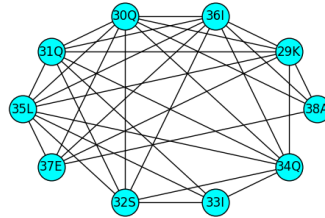
(a) 2croon1sn3 D1



(b) 2croon1sn3 D2

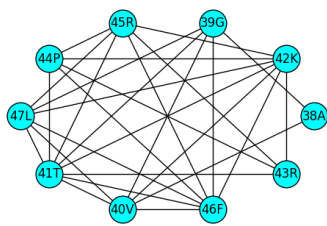


(c) 2croon1sn3 D3

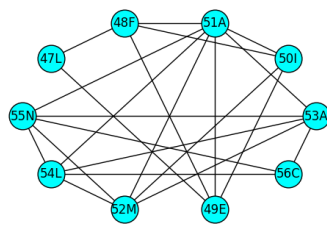


(d) 2croon1sn3 D4

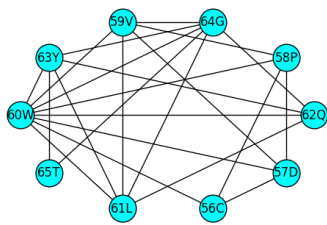
Figure 12: Domain Graphs 1 through 4 of Decoy 2croon1sn3



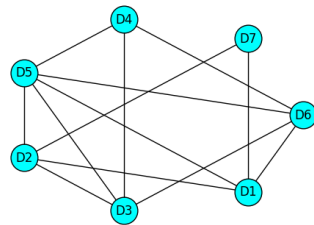
(a) 2croon1sn3 D5



(b) 2croon1sn3 D6



(c) 2croon1sn3 D7



(d) 2croon1sn3 Top Level Graph

Figure 13: Domain Graphs 5 through 7 and Top Level Graph for Decoy 2croon1sn3

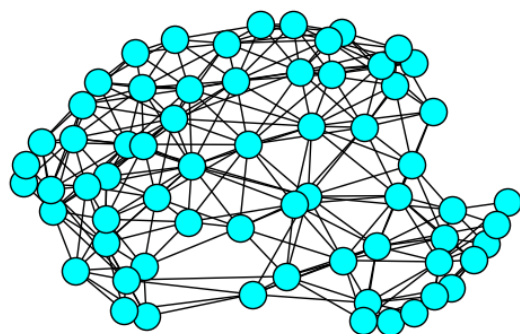
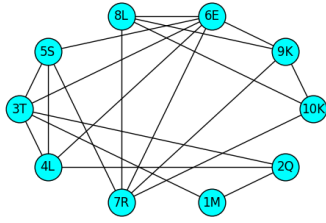
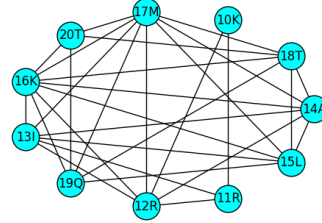


Figure 14: 2croon1sn3 Contact Map

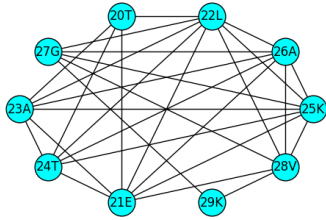
### A.3 2croon2ci2



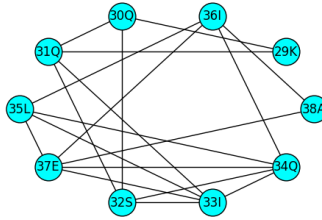
(a) 2croon2ci2 D1



(b) 2croon2ci2 D2



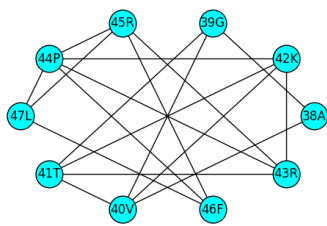
(c) 2croon2ci2 D3



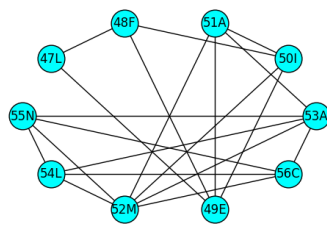
(d) 2croon2ci2 D4

Figure 15: Domain Graphs 1 through 4 of Decoy 2croon2ci2

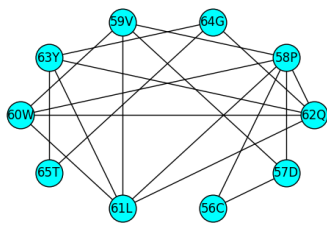




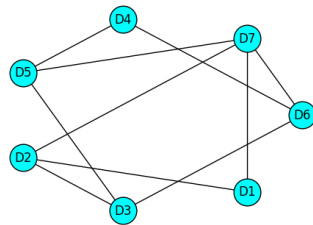
(a) 2croon2ci2 D5



(b) 2croon2ci2 D6



(c) 2croon2ci2 D7



(d) 2croon2ci2 Top Level Graph

Figure 16: Domain Graphs 5 through 7 and Top Level Graph for protein 2cro

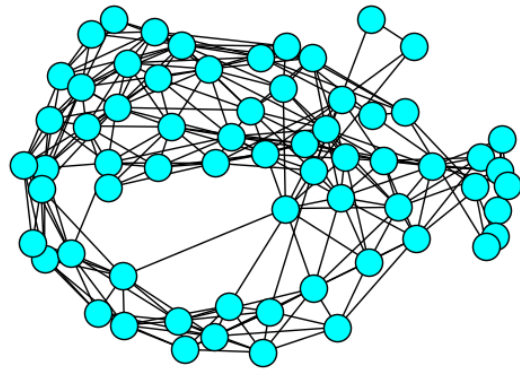
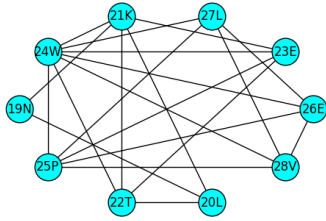
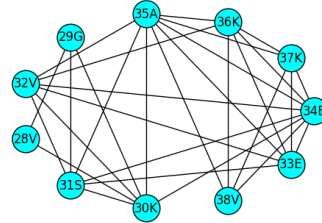


Figure 17: 2croon2ci2 Contact Map

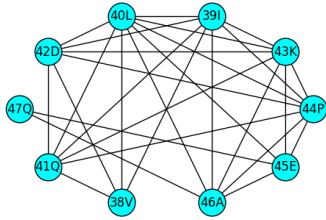
A.4 2ci2



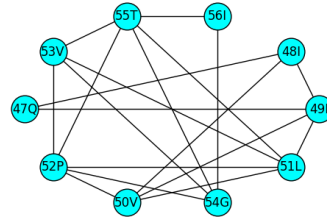
(a) 2ci2 D1



(b) 2ci2 D2

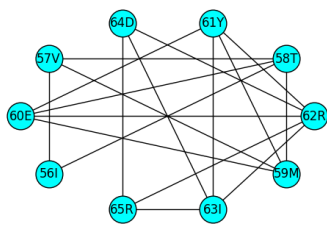


(c) 2ci2 D3

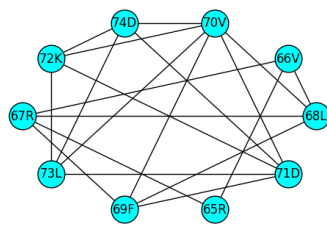


(d) 2ci2 D4

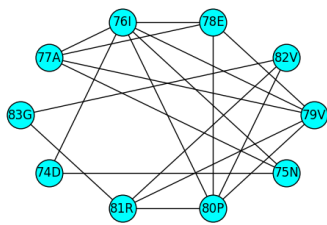
Figure 18: Domain Graphs 1 through 4 of protein 2ci2



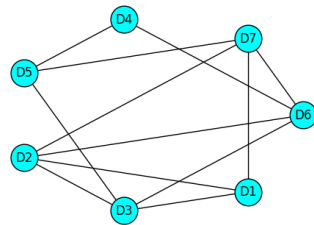
(a) 2ci2 D5



(b) 2ci2 D6



(c) 2ci2 D7



(d) 2ci2 Top Level Graph

Figure 19: Domain Graphs 5 through 7 and Top Level Graph for protein 2ci2

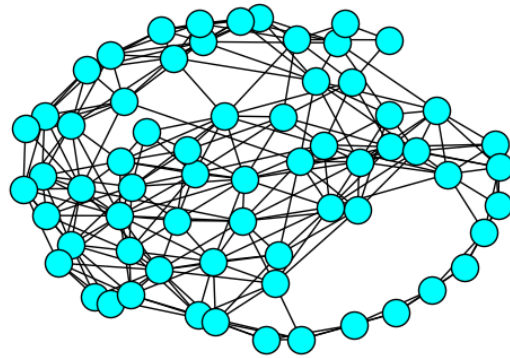
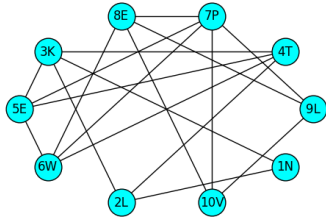
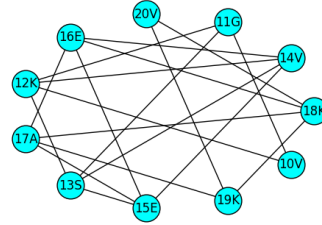


Figure 20: 2ci2 Contact Map

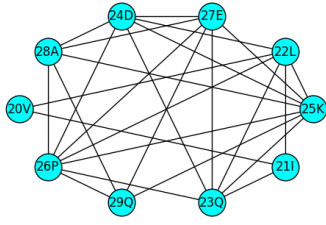
A.5 2ci2on1sn3



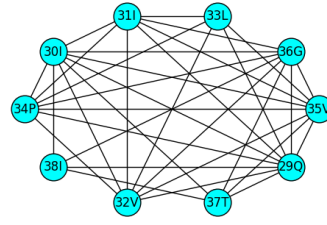
(a) 2ci2on1sn3 D1



(b) 2ci2on1sn3 D2

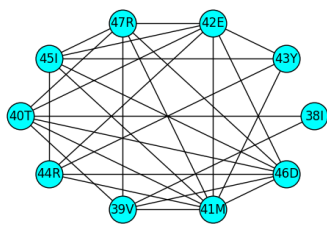


(c) 2ci2on1sn3 D3

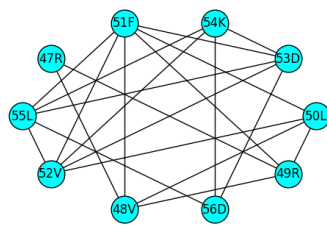


(d) 2ci2on1sn3 D4

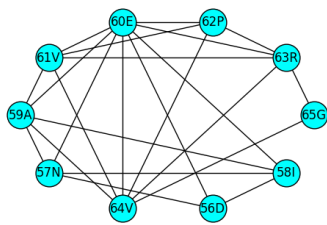
Figure 21: Domain Graphs 1 through 4 of decoy 2ci2on1sn3



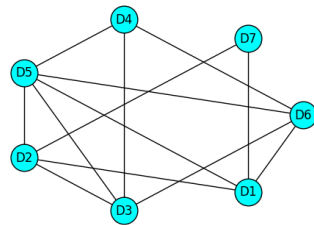
(a) 2ci2on1sn3 D5



(b) 2ci2on1sn3 D6



(c) 2ci2on1sn3 D7



(d) 2ci2on1sn3 Top Level Graph

Figure 22: Domain Graphs 5 through 7 and Top Level Graph for decoy 2ci2on1sn3

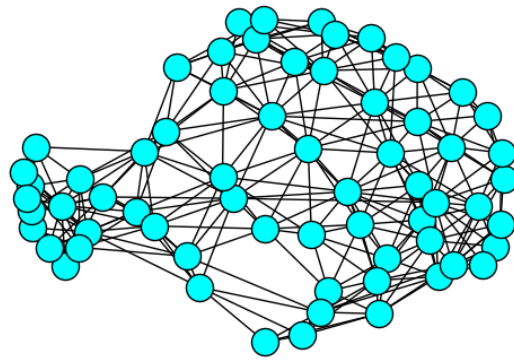
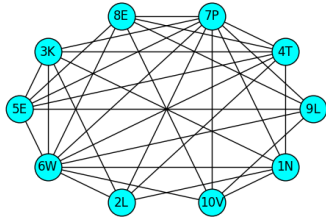


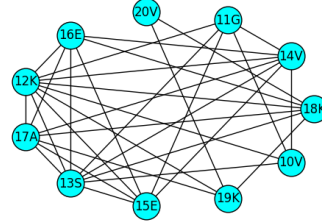
Figure 23: 2ci2on1sn3 Contact Map



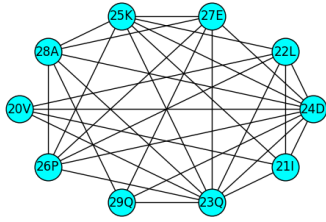
A.6 2ci2on2cro



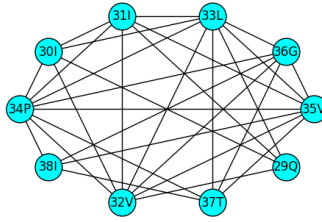
(a) 2ci2on2cro D1



(b) 2ci2on2cro D2

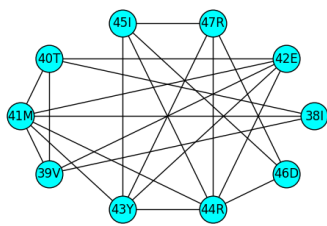


(c) 2ci2on2cro D3

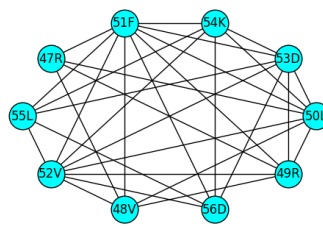


(d) 2ci2on2cro D4

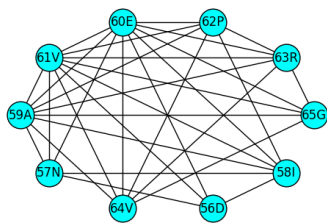
Figure 24: Domain Graphs 1 through 4 of decoy 2ci2on2cro



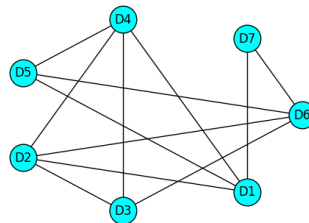
(a) 2ci2on2cro D5



(b) 2ci2on2cro D6



(c) 2ci2on2cro D7



(d) 2ci2on2cro Top Level Graph

Figure 25: Domain Graphs 5 through 7 and Top Level Graph for decoy 2ci2on2cro

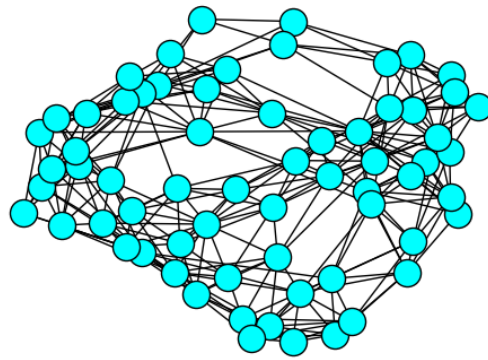
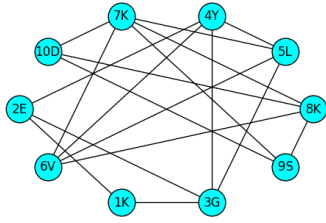
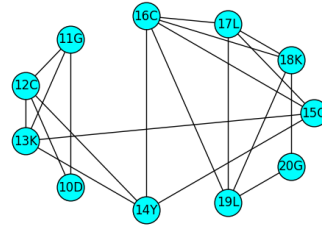


Figure 26: 2ci2on2cro Contact Map

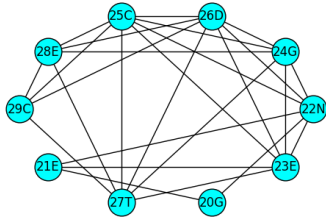
A.7 1sn3



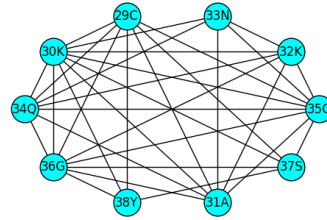
(a) 1sn3 D1



(b) 1sn3 D2

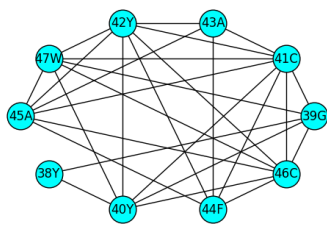


(c) 1sn3 D3

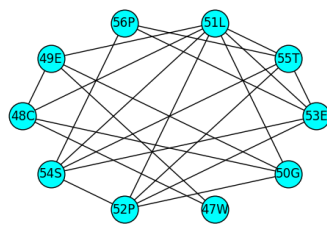


(d) 1sn3 D4

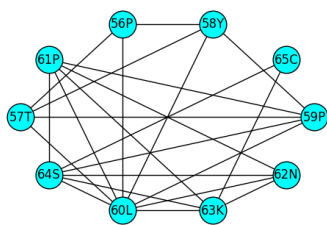
Figure 27: Domain Graphs 1 through 4 of protein 1sn3



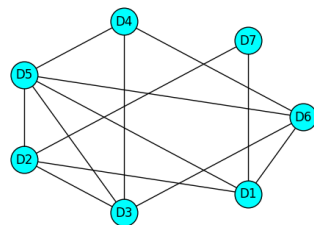
(a) 1sn3 D5



(b) 1sn3 D6



(c) 1sn3 D7



(d) 1sn3 Top Level Graph

Figure 28: Domain Graphs 5 through 7 and Top Level Graph for protein 1sn3

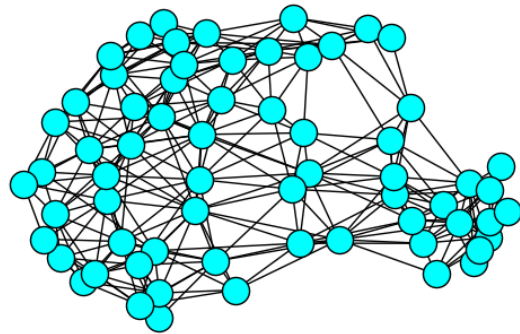
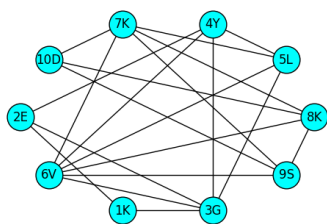
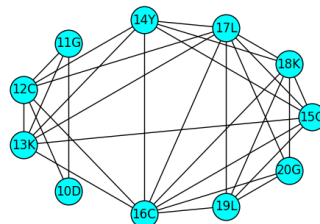


Figure 29: 1sn3 Contact Map

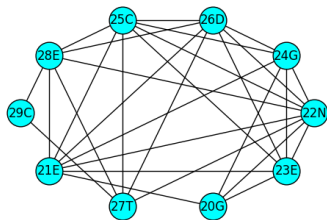
A.8 1sn3on2ci2



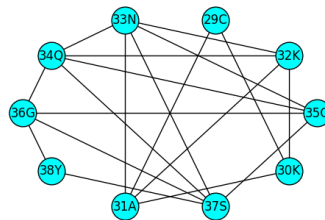
(a) 1sn3on2ci2 D1



(b) 1sn3on2ci2 D2

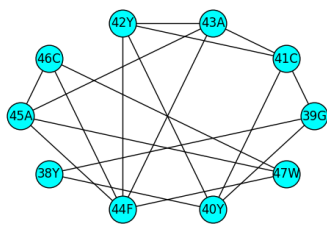


(c) 1snon2ci23D3

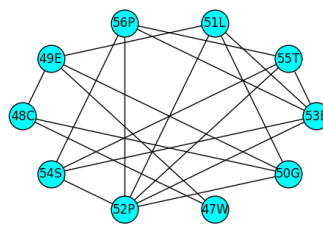


(d) 1sn3on2ci2 D4

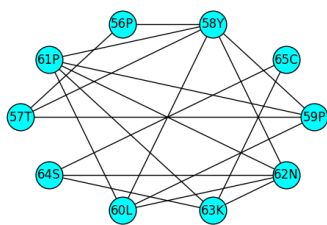
Figure 30: Domain Graphs 1 through 4 of decoy 1sn3on2ci2



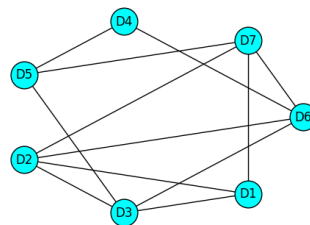
(a) 1sn3on2ci2 D5



(b) 1sn3on2ci2 D6



(c) 1sn3on2ci2 D7



(d) 1sn3on2ci2 Top Level Graph

Figure 31: Domain Graphs 5 through 7 and Top Level Graph for decoy 1sn3on2ci2



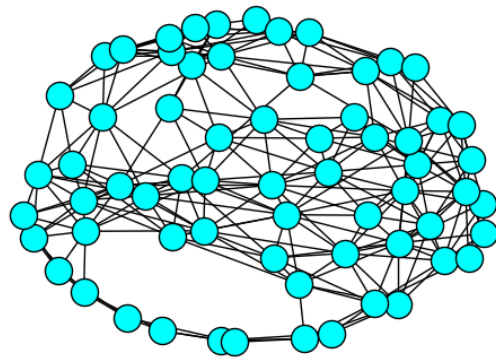
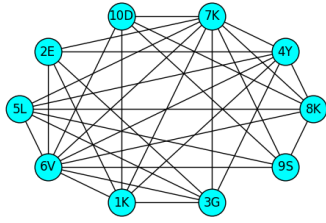
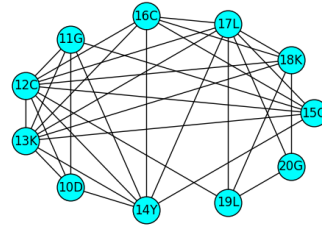


Figure 32: 1sn3on2ci2 Contact Map

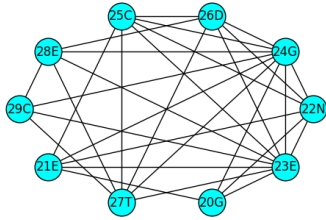
A.9 1sn3on2cro



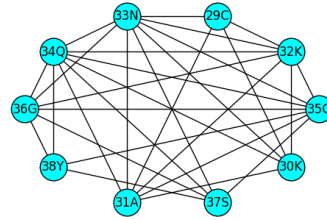
(a) 1sn3on2cro D1



(b) 1sn3on2cro D2

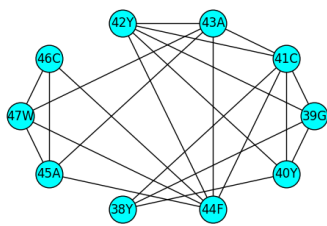


(c) 1sn3on2croD3

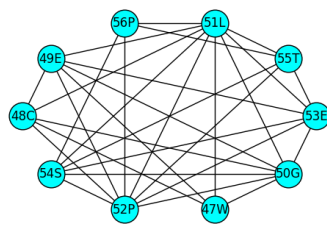


(d) 1sn3on2cro D4

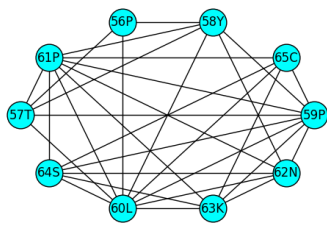
Figure 33: Domain Graphs 1 through 4 of decoy 1sn3on2cro



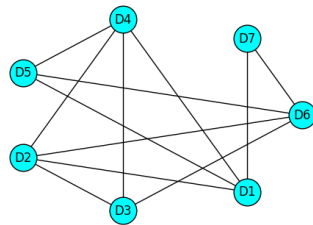
(a) 1sn3on2cro D5



(b) 1sn3on2cro D6



(c) 1sn3on2cro D7



(d) 1sn3on2cro Top Level Graph

Figure 34: Domain Graphs 5 through 7 and Top Level Graph for decoy 1sn3on2cro

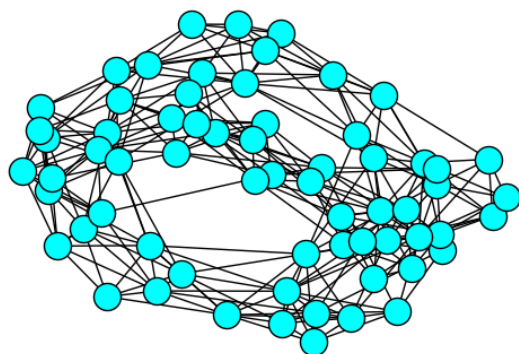


Figure 35: 1sn3on2cro Contact Map

B Tables of vertex weights and weighted invariants for Domain and Top Level

Graphs

Table 3: Amino Acid Descriptors (Part 1)

nme	AA1	AA3	G	g	d	c	m	p	Plr	Chrg	Hydpthy
A	A	ALA	12	12	2	0	1	12	0	0	1.8
R	R	ARG	54	36	9	0	1.75	14	1	1	-4.5
N	N	ASN	42	24	5	0	1.6	15	1	0	-3.5
D	D	ASP	44	24	5	0	1.6	16	1	-1	-3.5
C	C	CYS	12	12	3	0	1.333	32	0	0	2.5
E	E	GLU	42	24	6	0	1.667	16	1	-1	-3.5
Q	Q	GLN	44	24	6	0	1.667	15	1	0	-3.5
G	G	GLY	0	0	0	0	0	0	0	0	-0.4
H	H	HIS	40	36	6	5	2	14	1	1	-3.2
I	I	ILE	24	24	5	0	1.6	12	0	0	4.5
L	L	LEU	36	24	5	0	1.6	12	0	0	3.8
K	K	LYS	38	24	8	0	1.667	14	1	1	-3.9
M	M	MET	44	24	6	0	1.6	12	0	0	1.9
F	F	PHE	36	24	7	6	2	12	0	0	2.8
P	P	PRO	24	12	4	4	2	12	0	0	-1.6
S	S	SER	12	12	3	0	1.333	16	1	0	-0.8
T	T	THR	12	12	3	0	1.5	14	1	0	-0.7
W	W	TRP	62	48	9	9	2.182	12	0	0	-0.9
Y	Y	TYR	52	24	8	6	2	16	1	0	-1.3
V	V	VAL	12	12	3	0	1.5	12	0	0	4.2

Table 4: Amino Acid Descriptors (Part 2)

nme	stably	ss-stability	vanderWaal	chargetransf	chargedonar
A	2.18	9.8	2.50E-02	0	0
R	2.71	7.3	0.2	0	1
N	1.85	3.6	0.1	1	1
D	1.75	4.9	0.1	1	0
C	3.89	3	0.1	0	1
E	1.89	4.4	0.1	1	0
Q	2.16	2.4	0.1	0	1
G	1.17	0	2.50E-02	1	0
H	2.51	11.9	0.1	0	1
I	4.5	17.2	0.19	0	0
L	4.71	17	0.19	0	0
K	2.12	10.5	0.2	0	1
M	3.63	11.9	0.19	0	1
F	5.88	23	0.39	0	1
P	2.09	15	0.17	0	0
S	1.66	2.6	2.50E-02	0	0
T	2.18	6.9	0.1	0	0
W	6.46	24.2	0.56	0	1
Y	5.01	17.2	0.39	0	1
V	3.77	15.3	0.15	0	0

Table 5: Amino Acid Descriptors (Part 3)

nme	averhydrophocitiy	coilconformation	IsoElectric	Balaban
A	2.00E-02	0.71	6	0
R	-0.42	1.06	10.76	6216.573
N	-0.77	1.37	5.41	455.375
D	-1.04	1.21	2.77	464.711
C	0.77	1.19	5.05	22
E	-1.14	0.84	3.22	1306.932
Q	-1.1	0.87	5.65	1302.743
G	-0.8	1.52	5.97	0
H	0.26	1.07	7.59	1857.46
I	1.81	0.66	6.02	496.7265
L	1.14	0.69	5.98	418.2822
K	-0.41	0.99	9.74	3288.873
M	1	0.59	5.74	794.2392
F	1.35	0.71	5.48	3492.442
P	-9.00E-02	1.61	6.3	58.78775
S	-0.97	1.34	5.68	14
T	-0.77	1.08	5.66	127.363
W	1.71	0.76	5.89	9654.332
Y	1.11	1.07	5.66	5722.512
V	1.13	0.63	5.96	117.5755

Table 6: Amino Acid Descriptors (Part 4)

nme	RofGyr	ShapeIndex	EIIP
A	0.77	1.28	3.73E-02
R	2.38	2.34	9.59E-02
N	1.45	1.6	3.60E-03
D	1.43	1.6	0.1263
C	1.22	1.77	8.29E-02
E	1.77	1.56	7.61E-02
Q	1.75	1.56	5.80E-03
G	0.58	0	5.00E-03
H	1.78	2.99	2.42E-02
I	1.56	4.19	0
L	1.54	2.59	0
K	2.08	1.89	3.71E-02
M	1.8	2.35	8.23E-02
F	1.9	2.94	9.46E-02
P	1.25	2.67	1.98E-02
S	1.08	1.31	8.29E-02
T	1.24	3.03	9.41E-02
W	2.21	3.21	5.48E-02
Y	2.13	2.94	5.16E-02
V	1.29	3.67	5.70E-03



Table 7: 2cro Top Level (Part 1)

Dscrptr	min wt dom	max wt dom	min wt deg	max wt deg
G	48	602	296	1090
g	24	288	204	720
d	9	74	47	184
c	0	25	0	29
m	3.667	27.849	14	51
p	26	212	120	483
Plr	0	8	3	20
Chrg	-3	10	-3	12
Hydpthy	-68.6	26.4	-37	33
stablty	4.07	36.87	27	104
ss-stability	14.8	154.3	92	349
vanderWaal	0.1	3.83	0	4
chargetransf	0	6	0	6
chargedonar	0	15	3	17
averhydrophocitiy	-13.21	8.99	-6	10
coilconformation	1.55	18.71	6	30
IsoElectric	12.44	164.46	51	235
Balaban	14	49718.2857	6172	67661
RofGyr	3.39	21.39	14	53
ShapeIndex	2.84	36.78	18	75
EIIP	0	1.2788	0	0

Table 8: 2cro Top Level (Part 2)

Dscrptr	max clique wt	min clique wt	min wt vert cov
G	498	118	1440
g	336	84	1080
d	80	22	265
c	21	0	0
m	23.001	7.334	80.252
p	190	64	695
Plr	9	2	20
Chrg	4	-2	2
Hydpthy	14.1	-27.4	-40
stablty	48.11	11.99	153.26
ss-stability	153.6	36.5	524.5
vanderWaal	2.355	0.55	7.595
charge transf	3	0	0
chargedonar	8	1	20
averhydrophocitiy	6.7	-4.82	-7.15
coilconformation	13.29	3.89	49.59
IsoElectric	93.16	32.35	343.89
Balaban	26187.7886	2235.39515	33009.20405
RofGyr	23.79	7.6	80.92
ShapeIndex	35.25	10	116.59
EIIP	0.6581	0.1622	1.7706

Table 9: 2croon1sn3 Top Level (Part 1)

Dscrptr	min wt dom	max wt dom	min wt deg	max wt deg
G	36	846	112	984
g	24	684	72	660
d	14	169	18	168
c	0	25	0	31
m	3.682	43.251	6	45
p	24	365	57	425
Plr	0	19	1	19
Chrg	-2	10	-4	8
Hydpthy	-61.6	25	-45	31
stablty	6.21	78.34	8	90
ss-stability	16.7	267.2	18	322
vanderWaal	0.325	5.09	0	3
chargetransf	0	4	0	6
chargedonar	0	18	0	15
averhydrophocitiy	-14.32	12.3	-11	11
coilconformation	3.37	24.94	2	26
IsoElectric	32.95	177.25	22	200
Balaban	36	71449.22465	1535	58492
RofGyr	3.65	44.41	4	47
ShapeIndex	3.49	61.11	8	67
EIIP	0.0058	1.6273	0	0

Table 10: 2croon1sn3 Top Level (Part 2)

Dscrptr	max clique wt	min clique wt	min wt vert cov
G	546	146	1246
g	360	96	984
d	93	23	225
c	13	0	0
m	27.234	7.767	69.884
p	261	70	619
Plr	10	2	22
Chrg	6	-2	4
Hydpthy	17.6	-24.9	-50.8
stablty	49.46	12.31	141.46
ss-stability	172.7	30.5	453.7
vanderWaal	2.555	0.515	6.77
charge transf	4	0	0
chargedonar	8	0	20
averhydrophocitiy	5.82	-3.71	-10.75
coilconformation	19.2	4.49	38.83
IsoElectric	129.66	26.19	291.25
Balaban	39134.45095	2402.4349	54573.7766
RofGyr	27.91	7.38	70.18
ShapeIndex	38.78	10.05	102.75
EIIP	0.8806	0.202	1.4768

Table 11: 2croon2ci2 Top Level (Part 1)

Dscrptr	min wt dom	max wt dom	min wt deg	max wt deg
G	84	682	90	790
g	36	528	60	552
d	10	130	16	134
c	0	29	0	19
m	6.6	36.586	4	37
p	36	267	40	350
Plr	0	14	0	13
Chrg	-4	12	-3	8
Hydpthy	-50.1	55.2	-34	24
stablty	9.58	91.8	8	79
ss-stability	22.1	322	31	272
vanderWaal	0.45	5.41	0	2
chargetransf	0	6	0	4
chargedonar	0	13	0	14
averhydrophocitiy	-13.69	15.67	-6	8
coilconformation	10.02	27.54	3	21
IsoElectric	41.74	141.37	19	174
Balaban	139.5755	70607.00365	1351	56260
RofGyr	5.78	37.43	4	38
ShapeIndex	7.28	70.71	6	57
EIIP	0.0115	1.5998	0	0

Table 12: 2croon2ci2 Top Level (Part 2)

Dscrptr	max clique wt	min clique wt	min wt vert cov
G	276	36	1206
g	192	36	900
d	45	7	252
c	15	0	8
m	14.734	3.5	80.318
p	131	36	681
Plr	6	0	16
Chrg	4	-2	-2
Hydpthy	15.3	-20	-33.5
stablty	30.28	7.97	156.12
ss-stability	111.4	21.2	445.8
vanderWaal	1.58	0.225	6.735
charge transf	2	0	0
chargedonar	5	0	10
averhydrophocitiy	6.81	-4.21	-9.34
coilconformation	8.04	2.54	41.15
IsoElectric	61.15	14.88	337.56
Balaban	19128.4675	117.5755	26821.98105
RofGyr	14.31	3.41	80.69
ShapeIndex	23.79	5.46	115.11
EIIP	0.5305	0.0853	1.2269

Table 13: 2ci2 Top Level (Part1)

Dscrptr	min wt dom	max wt dom	min wt deg	max wt deg
G	60	908	90	890
g	72	552	72	564
d	19	159	18	148
c	0	23	0	27
m	9.45	39.901	6	42
p	48	264	52	370
Plr	0	17	1	17
Chrg	-4	8	-8	7
Hydpthy	-33	50.9	-40	27
stablty	15.27	85.96	11	74
ss-stability	38	312	40	288
vanderWaal	0.825	3.98	0	3
chargetransf	0	12	0	9
chargedonar	0	12	0	10
averhydrophocitiy	-9.46	18.68	-7	10
coilconformation	5.22	19.06	2	25
IsoElectric	41.36	124.3	27	163
Balaban	411.51425	53872.6592	1709	58868
RofGyr	4.15	41.17	5	41
ShapeIndex	12.71	70.64	8	67
EIIP	0.0266	1.3507	0	0

Table 14: 2ci2 Top Level (Part2)

Dscrptr	max clique wt	min clique wt	min wt vert cov
G	418	72	1544
g	264	72	1092
d	64	16	258
c	13	0	0
m	19.218	6.2	82.254
p	171	50	685
Plr	8	1	20
Chrg	4	-3	-5
Hydpthy	14.4	-14.5	-36.8
stablty	34.78	10.84	150.51
ss-stability	137.3	32.5	543.6
vanderWaal	1.825	0.59	7.325
charge transf	6	0	4
chargedonar	4	0	6
averhydrophocitiy	4.96	-3.69	-10.71
coilconformation	11.41	3.03	39.33
IsoElectric	78.64	20.42	333.35
Balaban	24963.97025	986.6035	31584.59055
RofGyr	19.08	5.65	83.21
ShapeIndex	27.84	7.64	126.88
EIIP	0.6519	0.0465	0.9718



Table 15: 2ci2on1sn3 Top Level (Part1)

Dscrptr	min wt dom	max wt dom	min wt deg	max wt deg
G	108	884	110	984
g	156	588	72	660
d	25	128	21	159
c	0	23	0	27
m	6.934	40.634	4	46
p	38	372	40	409
Plr	0	19	1	17
Chrg	-8	6	-8	6
Hydpthy	-62.6	54.4	-48	51
stablty	7.54	83.93	9	97
ss-stability	6.8	285.3	31	377
vanderWaal	0.275	2.77	0	3
chargetransf	0	12	0	9
chargedonar	0	12	0	11
averhydrophocitiy	-15.4	19.49	-9	15
coilconformation	4.75	21.09	2	26
IsoElectric	29.57	170.08	30	181
Balaban	484.302	51238.4045	1732	54182
RofGyr	8.55	23.63	5	46
ShapeIndex	3.12	53.66	5	79
EIIP	0.0114	1.4646	0	0

Table 16: 2ci2on1sn3 Top Level (Part2)

Dscrptr	max clique wt	min clique wt	min wt vert cov
G	722	188	1236
g	444	120	924
d	112	33	215
c	20	0	0
m	33.552	10.034	65.686
p	276	79	576
Plr	11	2	22
Chrg	3	-4	-5
Hydpthy	42.9	-23.2	-37.2
stablty	66.08	16.66	150.52
ss-stability	245	68.4	417.9
vanderWaal	3.42	0.935	6.44
charge transf	6	0	8
chargedonar	8	0	8
averhydrophocitiy	14.08	-4.54	-7.13
coilconformation	18.93	5.64	42.15
IsoElectric	122.43	35.93	274.59
Balaban	36021.4414	3398.8727	45707.18125
RofGyr	33.08	9.69	66.81
ShapeIndex	51.38	13.48	126.19
EIIP	0.9598	0.1033	1.5512

Table 17: 2ci2on2cro Top Level (Part1)

Dscrptr	min wt dom	max wt dom	min wt deg	max wt deg
G	24	606	248	1092
g	36	300	180	744
d	9	74	48	176
c	0	25	0	27
m	5.1	25.7	14	53
p	24	273	116	455
Plr	0	15	3	20
Chrg	-6	6	-6	5
Hydpthy	-42	60.9	-60	37
stablty	3.84	38.5	28	111
ss-stability	9.2	228.7	113	404
vanderWaal	0.5	3.76	0	4
chargetransf	0	9	0	11
chargedonar	0	7	2	12
averhydrophocitiy	-13.25	18.71	-8	12
coilconformation	1.63	17.59	6	31
IsoElectric	17.72	120.22	50	212
Balaban	470.302	36940.2857	6362	56617
RofGyr	3.45	23.12	15	52
ShapeIndex	6.28	43.71	22	89
EIIP	0	1.1902	0	0

Table 18: 2ci2on2cro Top Level (Part2)

Dscrptr	max clique wt	min clique wt	min wt vert cov
G	454	118	1252
g	300	84	876
d	73	20	257
c	23	0	0
m	23.349	6.001	67.286
p	177	55	681
Plr	7	1	16
Chrg	2	-3	-4
Hydpthy	24.3	-12.4	-36.8
stablty	49	8.12	125.35
ss-stability	191.1	21	442.7
vanderWaal	2.695	0.325	6.64
chargetransf	4	0	12
chargedonar	6	1	12
averhydrophocitiy	6.82	-3.88	-8.54
coilconformation	14.55	3.26	49.48
IsoElectric	92.53	18.09	327.18
Balaban	29545.22595	2277.5692	55069.06075
RofGyr	22	6.06	81.11
ShapeIndex	40.83	5.96	106.17
EIIP	0.5271	0.0857	1.7504

Table 19: 1sn3 Top Level (Part1)

Dscrptr	min wt dom	max wt dom	min wt deg	max wt deg
G	0	722	54	880
g	0	252	36	540
d	0	112	9	146
c	0	45	0	60
m	0	38.05	2	41
p	0	376	42	499
Plr	0	16	1	20
Chrg	-7	8	-5	5
Hydpthy	-55.4	36.7	-53	24
stablty	8.98	79.38	8	85
ss-stability	0	308.6	7	308
vanderWaal	0.175	5.72	0	2
chargetransf	0	16	0	15
chargedonar	0	14	1	16
averhydrophocitiy	-9.76	14.68	-13	11
coilconformation	3.38	29.89	3	35
IsoElectric	18.39	136.97	20	187
Balaban	0	39586.837	1328	59478
RofGyr	4.25	34.1	3	42
ShapeIndex	0	55.7	2	56
EIIP	0.02	1.3597	0	0

Table 20: 1sn3 Top Level (Part2)

Dscrptr	max clique wt	min clique wt	min wt vert cov
G	524	122	1236
g	348	72	816
d	86	19	216
c	27	0	0
m	27.534	4.934	60.681
p	292	45	620
Plr	12	1	16
Chrg	3	-3	-4
Hydpthy	13.1	-30.4	-27.9
stablty	53.99	8.2	136.38
ss-stability	153.1	18.5	334
vanderWaal	2.765	0.45	5.675
charge transf	7	0	6
chargedonar	10	1	20
averhydrophocitiy	7.41	-8	-6.8
coilconformation	21.55	5.28	51.61
IsoElectric	114.23	28.98	271.26
Balaban	32685.158	4713.3805	40170.12855
RofGyr	27.08	6.46	75.58
ShapeIndex	34.96	5.05	81.21
EIIP	1.1834	0.0659	1.5946

Table 21: 1sn3on2ci2 Top Level (Part1)

Dscrptr	min wt dom	max wt dom	min wt deg	max wt deg
G	12	582	106	766
g	12	492	60	456
d	2	96	17	122
c	0	45	0	36
m	1	31.816	5	35
p	12	309	44	397
Plr	0	11	0	16
Chrg	-8	5	-7	5
Hydpthy	-34.8	21.1	-43	23
stablty	11.87	64.21	9	76
ss-stability	8.6	232.8	24	259
vanderWaal	0.1	4.11	0	3
chargetransf	0	13	0	12
chargedonar	0	14	1	15
averhydrophocitiy	-15.68	9.2	-11	9
coilconformation	8.01	31.14	4	31
IsoElectric	34.58	125.81	19	157
Balaban	0	52725.127	1998	52249
RofGyr	3.86	27.33	4	35
ShapeIndex	2.84	41.74	5	50
EIIP	0.1143	0.9425	0	0

Table 22: 1sn3on2ci2 Top Level (Part2)

Dscrptr	max clique wt	min clique wt	min wt vert cov
G	342	88	1218
g	216	60	804
d	57	16	210
c	25	0	22
m	18.7	4.667	69.714
p	196	48	561
Plr	7	1	16
Chrg	2	-2	-4
Hydpthy	9.2	-23.4	-31.9
stablty	38.42	9.07	133.95
ss-stability	126.7	17.9	374.2
vanderWaal	2.325	0.425	6.47
charge transf	7	0	8
chargedonar	7	1	24
averhydrophocitiy	4.96	-7.48	-12.65
coilconformation	14.24	3.58	50.73
IsoElectric	73.73	22.37	246.91
Balaban	26343.59395	1419.15595	20442.37585
RofGyr	17.51	5.65	73.1
ShapeIndex	26.4	5.22	93.89
EIIP	0.69	0.0705	2.0679



Table 23: 1sn3on2cro Top Level (Part1)

Dscrptr	min wt dom	max wt dom	min wt deg	max wt deg
G	12	508	222	1008
g	12	300	144	612
d	5	100	39	163
c	0	52	0	54
m	2.5	26.298	10	47
p	24	272	108	538
Plr	0	12	4	22
Chrg	-6	6	-7	5
Hydpthy	-54.9	32	-55	17
stablty	4.46	50.92	21	94
ss-stability	2.6	164	59	306
vanderWaal	0.075	3.02	0	3
chargetransf	0	12	0	15
chargedonar	0	14	2	18
averhydrophocitiy	-14.21	7.55	-12	7
coilconformation	1.32	19.64	8	38
IsoElectric	19.73	127.1	53	201
Balaban	14	33785.8342	5663	57158
RofGyr	2.45	21.68	13	48
ShapeIndex	2.59	31.55	13	62
EIIP	0.005	1.5941	0	0

Table 24: 1sn3on2cro Top Level (Part2)

Dscrptr	max clique wt	min clique wt	min wt vert cov
G	422	84	1350
g	252	48	888
d	71	11	217
c	24	0	8
m	18.768	3.267	72.481
p	194	31	746
Plr	9	1	16
Chrg	2	-2	-4
Hydpthy	4.9	-21	-27
stablty	39.63	3.74	144.12
ss-stability	131.2	8	319.4
vanderWaal	2.42	0.2	5.155
chargetransf	7	0	6
chargedonar	9	1	28
averhydrophocitiy	2.64	-5.02	-11.86
coilconformation	17.16	2.21	51.1
IsoElectric	95.37	8.63	305.45
Balaban	30928.3015	714.3135	31564.6668
RofGyr	22.43	3.22	73.64
ShapeIndex	24.65	3.16	93.6
EIIP	0.6062	0.0797	1.8598

Table 25: Combined PCA Values

(a) Values 1 through 4

Protein	Val1	Val2	Val3	Val4n
2cro	-0.00013	-0.528499	-0.065336	-0.718719
2croon1sn3	-0.000286	-0.62593	-0.013377	-0.512119
2croon2ci2	-0.001368	-0.735074	-0.013983	-0.58544
2ci2	-0.004484	-0.602891	-0.019043	-0.658881
2ci2on1sn3	-0.00501	-0.54147	-0.018239	-0.572466
2ci2on2cro	0.005078	0.400395	0.068768	0.613392
1sn3	-0.000023	-0.447986	-0.015019	-0.673331
1sn3on2ci2	-0.000025	0.647992	0.024483	0.641952
1sn3on2cro	-0.000177	-0.422757	-0.070709	-0.715094

(b) Values 5 through 7

Protein	Val5	Val6	Val7
2cro	-0.278085	-0.023637	-0.349247
2croon1sn3	-0.342754	-0.020935	-0.477334
2croon2ci2	-0.19901	-0.001159	-0.277709
2ci2	-0.279353	-0.010936	-0.351934
2ci2on1sn3	-0.380546	-0.035765	-0.482328
2ci2on2cro	0.320241	0.024598	0.596255
1sn3	-0.369956	-0.053339	-0.453867
1sn3on2ci2	0.323701	0.017344	0.249643
1sn3on2cro	-0.38715	-0.008874	-0.393641

VITA

HANNAH GREEN

- Education: B.S. Mathematics, East Tennessee State University,  
Johnson City, Tennessee 2015  
M.S. Mathematical Sciences, East Tennessee  
State University,  
Johnson City, Tennessee 2017
- Professional Experience: Graduate Assistant, East Tennessee State University,  
Johnson City, Tennessee, 2015–2017
- Awards & Honors: Vice President, Kappa Mu Epsilon, 2016-2017  
ETSU Academic Performance Scholarship 2012-2015
- Publications: R. A. Beeler, H. Green, and R. T. Harper, “Peg Solitaire  
on Caterpillars,” *Integers*  
17(2) (2017).
- Computational Skills: Python, R, MATLAB,  $\LaTeX$