



SCHOOL of
GRADUATE STUDIES
EAST TENNESSEE STATE UNIVERSITY

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

Electronic Theses and Dissertations

Student Works

5-2015

A Hierarchical Graph for Nucleotide Binding Domain 2

Samuel Kakraba

East Tennessee State University

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

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

Recommended Citation

Kakraba, Samuel, "A Hierarchical Graph for Nucleotide Binding Domain 2" (2015). *Electronic Theses and Dissertations*. Paper 2517.
<https://dc.etsu.edu/etd/2517>

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.

A Hierarchical Graph for Nucleotide Binding Domain 2

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

Samuel Kakraba

May 2015

Debra Knisley, Ph.D., Chair

Robert Gardner, Ph.D.

Jeff Knisley, Ph.D.

Nicole Lewis, Ph.D.

Michele Joyner, Ph.D.

Keywords: Cystic Fibrosis, Mutation, Graph-Theoretic Models, Nucleotide Binding

Domain 2, NBD2, Molecular Descriptors, Nested Graph, CFTR, N1303K

ABSTRACT

A Hierarchical Graph for Nucleotide Binding Domain 2

by

Samuel Kakraba

One of the most prevalent inherited diseases is cystic fibrosis. This disease is caused by a mutation in a membrane protein, the cystic fibrosis transmembrane conductance regulator (CFTR). CFTR is known to function as a chloride channel that regulates the viscosity of mucus that lines the ducts of a number of organs. Generally, most of the prevalent mutations of CFTR are located in one of two nucleotide binding domains, namely, the nucleotide binding domain 1 (NBD1). However, some mutations in nucleotide binding domain 2 (NBD2) can equally cause cystic fibrosis. In this work, a hierarchical graph is built for NBD2. Using this model for NBD2, we examine the consequence of single point mutations on NBD2. We collate the wildtype structure with eight of the most prevalent mutations and observe how the NBD2 is affected by each of these mutations.

Copyright by
Samuel Kakraba
All Rights Reserved
May 2015

DEDICATION

This work is dedicated to my lovely wife and daughter (Esther Kakraba & Maame Efua Essumanba Adjei-Kakraba), my amazing host family (Luke Guthrie & Laura Lacey Guthrie and John Luther Guthrie Jr.), and my sweet mother (Madam Ekuwa Awotwe).

ACKNOWLEDGMENTS

My utmost thanks go to the Almighty God for his strength, protection, knowledge and wisdom imparted unto me. I am highly indebted to my supervisor, Prof. Debra Knisley of Department of Mathematics and Statistics at East Tennessee State University, and committee members for the wonderful support while undertaking this study. I am thankful to my wife, Esther Kakraba, for all the support. Special thanks go to my host family (Luke Guthrie and Laura Lacey Guthrie) for their tremendous support while in USA for my graduate studies. My sincere gratitude goes to following; Kakraba family of Akotokyir (Cape Coast in Ghana), Andzie-Ghansah family of Cape Coast (Ghana), Yankson, Fletcher, Fiifi Mensah, Amuah, Sokpe, Prah, Yakubu and Naandam family (all of UCC, Ghana), Saah family (Ghana), Price family (ETSU), Bussey family (SC. and MD.), Harley family (Anaji, Takoradi - Ghana), Dadzie family of Takoradi (Ghana), Anna Braxton Duncan, Nick Varakin, Lacey and Duncan family (SC), Varakin family of Maryland, Bardoner family (TN), Kofi Essuman family (Cape Coast, Ghana), Clement Aayire Yadem of Finland, Barbara Awuku, Julius Kwesi Aayire Kunweleyil (Fargo, ND), Cory Ball (Kingsport), Kimberly Brockman (Graduate School, ETSU), Prof. Jamie McGill Whittimore, Prof. Teresa Haynes and Prof. Daryl Stephens (all of ETSU), and the entire faculty of the Mathematics Department of ETSU, Christa Seyler of Union University, Joshua Davis (ETSU), Chris Morkle (Suhum, Ghana), Baidoo family of Akotokyir, my fellow graduate students at ETSU., Geoffrey Mensah, Emmanuel Kwesi Mensah (Nyinasin, Ghana), and all who have in one way or other contributed to the life of my family.

TABLE OF CONTENTS

ABSTRACT	2
DEDICATION	4
ACKNOWLEDGMENTS	5
LIST OF TABLES	8
LIST OF FIGURES	9
1 INTRODUCTION	10
1.1 Roles of Protein	10
1.2 Importance of Structure Related to the Function	11
1.3 A Mathematical Model of Protein to Characterize Functions	17
2 GRAPH-THEORETIC MODELS OF PROTEINS	19
2.1 Terms and Definition of Graph Theory	19
2.2 A Survey of Graph Models in the Literature	22
3 GRAPHICAL INVARIANTS AS AMINO ACID DESCRIPTORS	28
3.1 Explanation of Graph Invariants	28
3.2 Molecular Descriptors or Combinatorial Descriptors of Amino Acids	30
3.3 Table of Molecular Descriptors for the 20 Most Essential Amino Acids	33
4 THE HIERARCHICAL/ NESTED GRAPH MODEL OF NBD2	37
4.1 Cystic Fibrosis and CFTR	37
4.2 The Model for Nucleotide Binding Domain 2 (NBD2)	39

5	THE EFFECT OF SINGLE-POINT MUTATIONS ON NBD2 AS SHOWN BY THE MODEL	44
5.1	Some Known Mutation in NBD2 and Association with Cystic Fibrosis	44
5.2	Application of a Single Point Mutation on the Model for NBD2	45
5.3	Clustering of Single Point Mutations/ Results of Single Point Mutations	47
5.4	Discussion of Results	47
6	CONCLUSION	49
6.1	Linking Findings to Existing Literature on Graph-Theoretic Models	49
6.2	Future Research Directions/ Open Problems	49
6.3	Summary	50
	BIBLIOGRAPHY	52
	VITA	59

LIST OF TABLES

1	Molecular Descriptors of the 20 Most Essential Amino Acids	33
2	Molecular Descriptors of the 20 Most Essential Amino Acids Continued 1	34
3	Molecular Descriptors of the 20 Most Essential Amino Acids Continued 2	35
4	Subdomain, Subsequence, Amino Acid Sequence	40
5	Subdomain, Subsequence, Reason	40
6	Classification of Mutations Based on Cystic Fibrosis [41, 25, 40, 43] .	44
7	Top Level Graph Molecular Descriptors for Single Point Mutations .	46

LIST OF FIGURES

1	Protein Structure Showing all Four Levels of Protein, Pearson Inc.,(2010)	13
2	Protein Folding [12]	14
3	Effect of Mutation on the Structure of DNA, [9, 8]	15
4	Nested Graph-Theoretic Model for NBD1 by Knisley et al.[36]	26
5	Subdomain graph of G2 with 508 <i>F</i> and without 508 <i>F</i> , [36]	27
6	Graph to Illustrate Some Standard Definitions	29
7	Graph to Illustrate Some Adopted Definitions	31
8	Graph-Theoretic Model for Tryptophan	32
9	Midlevel Graph for Subdomain S5 (on left) and S4 (on right)	41
10	Hierarchical Graph for NBD2	41
11	Clustering of Mutations, Output from R [4]	47

1 INTRODUCTION

In this chapter, we discuss the roles of proteins, the importance of structure related to the function of a protein, and how a single point mutation in the protein sequence can prevent the protein from undertaking the normal functions. A brief note on mathematical models of protein to characterize functions is also presented in this chapter.

1.1 Roles of Protein

Proteins, as large complex molecules play very important roles in the body. Each protein performs a specific function in the cells. Foreign invaders such as bacteria, viruses among others, are defended from the body by specialized proteins called *antibodies*. Proteins like *myosin* and *actin* (known as *contractile proteins*) function in muscle contraction and movement. Other form of proteins are enzymes. Enzymes (often referred to as catalysts) like *lactase* break down the sugar lactose found in milk while *pepsin* is a digestive enzyme in the stomach that breaks down proteins in food. Some proteins also serve hormonal functions. *Oxytocin*, *insulin* and *somatotropin* are examples of hormonal proteins. These forms of proteins are called messenger proteins. They are specialized in helping to coordinate certain bodily activities. Illustratively, *somatotropin* is a growth hormone that stimulates protein production in muscle cells while *insulin* is noted to regulate glucose metabolism through controlling blood-sugar concentration. Contractions in females during childbirth (useful for safe labor) are stimulated by *oxytocin* [11, 47, 29, 13, 3, 6]. Proteins like *collagen*, *elastin* and *keratin* are often termed *structural proteins*. They are fibrous and provide support. Connec-

tive tissues like ligaments and tendons, derive their support from *elastin* and *collagen*, while protective coverings like beaks, horns, quills and feathers obtain their strength from *keratin*. Specialized proteins known as *transport proteins* carry other proteins and compounds throughout the body. Hemoglobin, found in red blood cells is a typical example of a transport protein. Transportation of oxygen from the lungs to all tissues and cells as well as carriage of carbon dioxide (a metabolic waste product) back to the lungs for excretion from the body are all functions of hemoglobin. When our bodies need energy in the absence or depletion of carbohydrates, energy from proteins is obtained for use by the body, by the degradation of proteins into their component amino acids and subsequently, oxidization processes analogous to glucose take place, thereby creating energy for the body [11, 47, 29, 33, 13, 3, 6].

1.2 Importance of Structure Related to the Function

Polymer-sequences, made up of several amino acids, form proteins. With the exception of proline, each amino acid has the same fundamental structure, differing only in the side-chain, designated the R-group. Research has found that protein chain is estimated to have approximately in the range of 50 to 2000 amino acid residues. During the process of chemical combination of amino acids, water molecule is lost. The peptide chain then forms after the water molecule is lost. Therefore, a peptide chain is made up of the residues of amino acid or amino acid residues. In view of this, each unit of protein is called an amino acid residue. Proteins have four (4) structural levels namely; *primary*, *secondary*, *tertiary* and *quaternary* structure. The linear sequence or order of covalently-linked specific amino acids in the polypeptide chain is the *pri-*

primary structure of a protein. By well-established standards, the primary structure of a protein is thought of to start from the amino-terminal (N) end to the carboxyl-terminal (C) end. The unique sequence of a protein accounts for the structure and function of that protein. The primary structure of each protein is unique, owing to both the different ordering or arrangement of the amino acids in the polypeptide and the total number of amino acids constituting the protein molecule. The *secondary structure* of protein is defined by the patterns of hydrogen bonds between backbone amino and carboxyl groups. A secondary structure of a protein pertains to the folding of a polypeptide chain. The folding of the polypeptide chain results in either an alpha helix, beta strand or a random coil structure, which characterize the secondary structure of protein. Nucleic acids like the clover leaf structure of *tRNA* is a typical example of secondary structure of protein [5]. By tertiary structure of a protein, we refer to the protein's three-dimensional structure by complete folding of the sheets and helices of a secondary structure. The tertiary structure is held in position by hydrophobic and hydrophilic interactions [3, 6, 2, 5]. Figure 1 depicts the primary, secondary, tertiary and quaternary structure of protein.

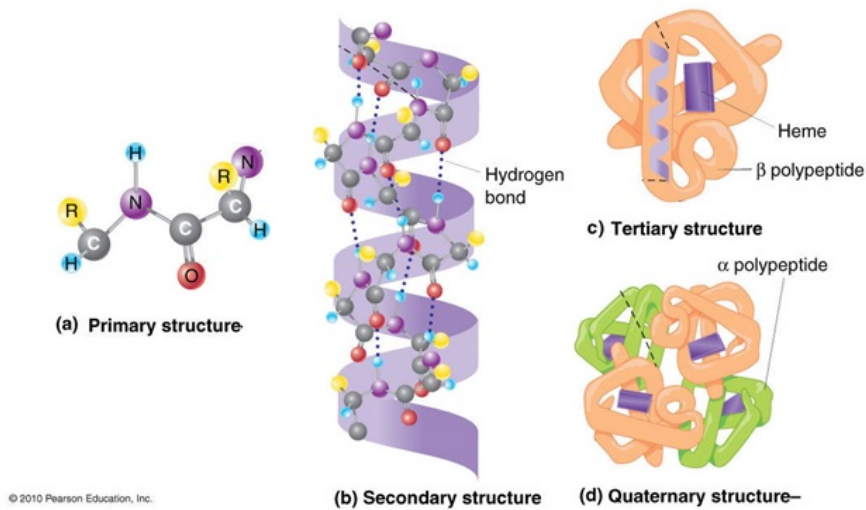


Figure 1: Protein Structure Showing all Four Levels of Protein, Pearson Inc.,(2010)

The process by which the protein structure takes on its functional shape or conformation is termed as *protein folding*. Protein folding is a physical process by which a polypeptide folds into its characteristic and functional three-dimensional structure from random coil [17]. Before protein folding takes place, each protein portrays or exists as an unfolded polypeptide or random coil when translated from a sequence of *mRNA* to a linear chain of amino acids. The unfolded polypeptide or random coil is unstable (long-lasting) three-dimensional structure. The interaction between amino acids forms a well-defined three dimensional structure which is termed the folded protein. Amino acids interact with each other to produce a well-defined three-dimensional structure, the *folded protein* termed as *native state*. The amino acid sequence or order dictates what type of three-dimensional structure results from the protein folding. The process of protein folding starts by the N-terminus of the protein folding while the C-terminal portion of the protein is still undergoing synthesis by the *ribosome*. These processes occur concurrently. Specialized proteins called *chaperones* are known to assist in the folding of other proteins. The shape, size and function of

a particular protein are determined by the three-dimensional structure of the protein in question. Figure 2 shows different proteins folding into diverse shapes that are function-specific.

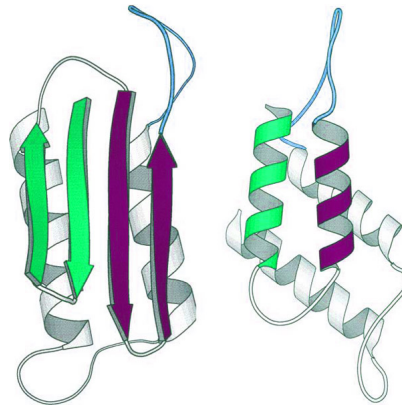


Figure 2: Protein Folding [12]

Mutation is the permanent change of the structure of a gene. Mutations result in a variant form of structure of genes that may be passed onto future generations of the organism. Unfortunately, some mutations damage the DNA structure thereby significantly changing the genetic information. Mutations can be accounted for by several factors. Errors that arise in DNA replication or from the damaging effects of mutagens, such as chemicals and radiation, which react with DNA and change the structures of individual nucleotides, can lead to mutations. Illustratively, during DNA replication, an organic base may be paired incorrectly within the newly forming strand, or some extra organic bases may be built into its structure. Alternatively, some portions or sections of DNA strands may be moved to other regions of the molecule, or deleted, or even attached to other chromosomes. Should either be the case, it results in the genetic information being changed. The molecular structure of a protein constructed from this new genetic information that results from this mutation,

will likely be faulty and either malfunctioning, or not function at all, in some extreme cases. Most mutations that occur are *point mutations*. It is well established fact that point mutations are known to replace one nucleotide with another; even though other forms of mutations involve insertion or deletion of one or a few nucleotides. Figure 3 depicts how a mutation might change the structure of the DNA molecule [13, 3, 6, 2, 5, 17, 19].

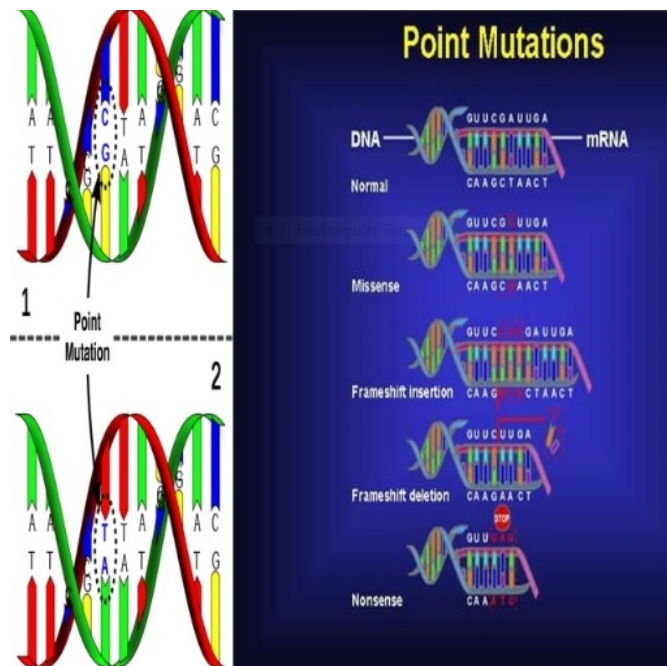


Figure 3: Effect of Mutation on the Structure of DNA, [9, 8]

Scientists like biologists, in particular computational biologists still battle with the seemingly incomprehensible thought of how mutations in the gene can cause specific change in structure and in the long run prevent the protein from undertaking its normal function (cause the protein to dysfunction), despite all the efforts to understand the complexities in systems biology being made. That is to say, not so much understanding has been gained on how a domain of the protein can be significantly

affected by a mutation in some part of the said domain. In this thesis, a mathematical model using graph theory to help predict the effect of a mutation on a protein known to cause a disease, namely cystic fibrosis, is presented. It is hypothesized that graph theory can be used to measure change in a protein domain caused by a mutation and therefore assist us in our examination of how the protein domain in which a mutation occurs will respond to the respective mutation. Even though Knisley et al. [36] were the only people to use a hierarchical graph as a mathematical model for the study of effect of mutation on the NBD1 for CFTR, their model was only for NBD1. Currently, no literature exists on using theoretical nested graphs in studying the effect of mutations on the protein structure in NBD2, thereby begging researchers to investigate further. Despite that most mutations do occur in NBD1, a number of them also occur in NBD2. There are seventeen mutations in the LSGGQ sequence and Walker B motif of NBD1 which cause CF, while there are four mutations in respective region of NBD2. Authenticated research has found out that whereas there is only one mutation in the Walker A motif of NBD1 causing cystic fibrosis, we have as many as five of these mutations taking place in NBD2. In view of the the fact that mutations that results in cystic fibrosis can equally occur in NBD2, it is appropriate to make an effort to gain understanding on how mutations in NBD2 can impact significantly on NBD2 [36, 32, 44, 37, 34, 38, 23]. In this thesis, we present a mathematical model for NBD2 of CFTR, using graph theory to help study how NBD2 is affected by a mutation known to cause cystic fibrosis.

1.3 A Mathematical Model of Protein to Characterize Functions

Modeling plays a key role in all aspects of life. By way of definition, a *model* is any simplification, substitute or stand-in for what we are really studying or predicting. Scientists use models to gain a better understanding of systems that cannot be studied in real life or that would be too complicated to study. Models are used because they are convenient substitutes, the way that a recipe is a convenient aid in cooking. The main aim of systems biology is to make the interactions of cellular components in a systemic manner to be understandable to the intelligent mind. Interestingly, theoretically and practically, mathematical modeling plays a crucial role integrating and testing models. Illustratively, modeling of biological systems permit us to simulate the way in which such systems work or function and respond (react) to some treatments, test or stimuli. Obviously, it is much easier to undertake such tests by use of models than performing such tests on living organisms or systems all the time. When results from model prove useful and workable for a particular test (or treatment/conditions or stimuli), we can then apply the result in a real life setting. Models are also convenient to use for instances where we can never directly test otherwise in real life [24, 26, 22].

In the past, several scientists used physical and chemical properties in modeling of biological systems in an attempt to characterize functions. Although the principles of graph theory were earlier used in the study of fields like computer networks and telecommunication, transportation services such as airline reservation, electrical engineering among others, it was not until recently that the field of graph theory found its place in modeling biological systems. In particular, graph-theoretic models have

proven to be an indispensable mathematical tools for investigating protein structure, folding, and to characterize protein function [30]. In this way, by the use of graph-theoretic models, meaningful insight into protein structures is being gained. In this thesis, we use a graph-theoretic model to build a hierarchical graph for NBD2 and use it to examine the impact of cystic fibrosis causing-mutations on the NBD2. Knisley et al. built a nested graph for NBD1 and used it to predict the effect of mutations on NBD1. Details of the work of Knisley et al. are discussed in the literature review. Even though the method of this research is analogous to that used by Knisley et al., two main differences exist between this work and their work. Knisley et al. were concerned with cystic fibrosis causing-mutations in NBD1 and it's resulting impact on NBD1. However, in this thesis, we are concerned with mutations that results in cystic fibrosis in NBD2. In view of this, we build a hierarchical or nested graph and use it to examine the impact of mutations that cause cystic fibrosis on NBD2. Another difference arises from the improved molecular descriptors that will be used to build the nested graph for NBD2. We will restrict ourselves to the mutations that occur in the part of the protein that we model in this work [36] .

2 GRAPH-THEORETIC MODELS OF PROTEINS

This chapter addresses some basic terms and definitions in graph theory that is essential to this work. The chapter also reviews literature on graph-theoretic models relevant to this research.

2.1 Terms and Definition of Graph Theory

Graph theory is a branch of discrete mathematics. In discrete mathematics, objects such as integers, graphs, and statements of logic are studied. Irrespective of the fact that the history of graph theory may be specifically traced to 1735, when the Swiss mathematician, Leonhard Euler, solved the Königsberg bridge problem. Unlike many branches of mathematics that date back to time immemorial, graph theory is new since the most parts has been developed since 1890. Below are some standard definitions in graph theory that are useful for this thesis. These definitions and discussion below are discernible from [26, 22, 46, 27, 16].

A Graph, G is a finite nonempty set V of objects called *vertices* (the singular is *vertex*) together with a possibly empty set E of 2- element subsets of V called *edges*. Links and lines are synonymous to edges while points and nodes can be used in place of vertices. By way of convention, we write $G = (V(G), E(G))$ to mean that a graph G has *vertex set* $V(G)$ and *edge set* $E(G)$. We consider only simple graphs in this work. By simple graphs, we refer to graphs with no multiple edges or loops. Initially though, graphs were called *linkages* by some mathematicians until James Joseph Sylvester (1814-1897) introduced the idea of *graphs* in place of linkages. The *order* of a graph G denoted by $n(G)$ is the total number of vertices in graph G .

The *size* of a graph denoted by $m(G)$ refers to the number of edges or links in the graph G . The *degree* of a vertex v in a graph G is the number of edges in G that are adjacent to vertex v . In other words, the degree of v is the number of vertices in its neighborhood $N(v)$. Similarly, the degree of v is the number of edges that are incident to v . We refer to the largest degree among the vertices of graph G as the *maximum degree* and call the least or smallest degree among the vertices of graph G as the *minimum degree*. We denote the maximum degree of a graph G by $\Delta(G)$ and represent the minimum degree of a graph G by $\delta(G)$.

The eccentricity $e(v)$ of a vertex in a connected graph G is the distance between v and a vertex farthest from v in G . The greatest eccentricity among the eccentricities of all vertices of G is called the diameter $diam(G)$, while the smallest eccentricity among all the eccentricities of the vertices of G is called the radius $rad(G)$.

A vertex v in a graph G is said to dominate itself and each of its neighbors, that is v dominates the vertices in its closed neighbourhood $N[v]$. A set S of vertices of graph G is a *dominating set* of G if every vertex of G is dominated by at least one vertex of S . In other words, a set S of vertices of a graph G is a *dominating set* if every vertex in $V(G) - S$ is adjacent to at least one vertex in S . The minimum cardinality among the dominating sets of G is called the *domination number* of G and is represented by $\gamma(G)$.

We will denote the adjacency matrix of the graph G by $A(G)$. The *adjacency matrix* of graph G denoted by $A(G)$ and given by $A(G) = \begin{cases} 1 & \text{if } v_i v_j \in E(G) \\ 0 & \text{otherwise} \end{cases}$ where $v_i v_j$ denotes an edge in G . The *Laplacian matrix* of the weighted graph is

given by $L(G) = D(G) - A(G)$.

A scalar λ is called an *eigenvalue* of the $n \times n$ matrix A if, there is a nontrivial solution x of $Ax = \lambda x$. Such an x is called an eigenvector corresponding to the eigenvalue λ .

Atomic number is defined as the number of protons in the nucleus of an atom. Atomic number determines the chemical properties of an element and its place in the periodic table. Conforming to accepted standards, the atomic number is represented with the symbol Z .

Total weighted degree is the summation of weights assigned to each vertex in the weighted graph denoted by d_{wt} . The domination number of the weighted graph is called *weighted domination number* and is denoted by $\gamma_w(G)$.

Weighted eccentricity, $e_w(G)$: This is the set containing eccentricity of each vertex in the weighted graph, G . We term the minimum number in this set, as the *weighted radius*, $rad_w(G)$ of the weighted graph. The maximum value among the set containing the eccentricity of each vertex in the weighted graph of G is termed the *weighted diameter* and denoted by $diam_w(G)$. Normalized eccentricity of the weighted graph denoted by $e_{wn}(G)$ refers to the sum of the eccentricity of each vertex divided by order of the weighted graph (number of vertices in the weighted graph).

Weighted adjacency matrix, $A_w(G)$ is the adjacency matrix of the weighted graph while *weighted diagonalized matrix*, $D_w(G)$ is the diagonalised matrix of the weighted graph. We use *weighted Laplacian matrix* to represent the Laplacian matrix of the weighted graph and denote it by $L_w(G)$. From the standard definition of Laplacian matrix, it follows that:

$$L_w(G) = D_w(G) - A_w(G)$$

A scalar λ_{wi} is called the *weighted eigenvalue* of the $n \times n$ matrix A obtained from the weighted graph, if there is a nontrivial solution x of $Ax = \lambda x$. Such an x is called an eigenvector corresponding to the eigenvalue λ_{wi} . The maximum value among the *eigenvalues* of the weighted graph is denoted by λ_{wmax} , and called the *maximum weighted eigenvalue* while the minimum value among the eigenvalues, denoted by λ_{wmin} of the weighted graph is called the *minimum weighted eigenvalue* of the weighted graph.

2.2 A Survey of Graph Models in the Literature

A survey of graph-theoretic models in the literature reveals interesting work done over the past years. Some literature related to this thesis work is discussed below. Gil Amitai, Arye Shemesh, Einat Sitbon, Maxim Shklar and Dvir Netanel [21] in their work on *Network Analysis of Protein Structures Identifies Functional Residues* developed a method for changing protein structures into interaction graphs for the residue. They used CSU program to find all inter-atomic contacts for each protein chain. They then incorporated the atomic contacts found for each amino acid residue. Edges represented interaction between residues while vertices represented the connected residue of RIG. The interactions took a number of things into consideration, including backbone peptide bonds as well as non-covalent bonds (such as hydrogen and hydrophobic interactions). In their quest to gain a meaningful insight into protein structures, protein structures were also drawn from the Protein Data Bank dataset. During the examination of their structural set, the method that was explained by Thornton et

al.,[18] was used. The structure set used in their work was without similar (homologous) pairs and took into consideration all six top-level enzyme classification (EC) numbers. With the use of the the *NACCESS program*, they successfully computed the *residue relative accessibility*. PyMol program was used to exemplify protein structures [46, 20, 39]. Samudrala and Mouth used clique-finding algorithm of a graph-theoretic model in their attempt to investigate the side chain conformational space in a comparative modeling of proteins. Weighted vertices and edges were used in their work. With the exception of vertices that were from the same side chain and those that gave rise to steric clashes, edges were drawn between all nodes in the graph considered for the study. With the use of appropriate interaction scales for weighted edges between the nodes that thrived on algorithms that found the cliques in the graph, weighted edges were obtained for edges in the graph. Upon the constructing of the entire graph, computations involving finding clique numbers were employed to find all the maximal set of completely connected vertices. Depending on the vertex and edge weights, a rating scale was adopted that was representative of the computed clique scores. This algorithm was employed in building a comparative model for the side chains, segments of main-chain and mix and match between different homologues in context sensitive manner [27, 16, 45]. By way of vertex-weighted hierarchical (nested) graph, Knisley et al.[36], successfully modeled NBD1 of CFTR for the study of effect of mutations that cause cystic fibrosis in NBD1. Like Samudrala and Mouth, the graph-theoretic model (nested graph) built by Knisley et al. for NBD1 was weighted graph. By way of explanation of nested graph, if a conceptual graph G1 is *nested* inside a concept C, it means that: either G1 is directly part of the referent of C or G1 is directly part

of the referent of a concept C_2 which is nested inside C . To start with, they built a graph-theoretic model for each of twenty main amino acids. The backbone and central carbon atom were denoted by a single vertex. A vertex represented each of the atoms in the respective amino acid residue structure. The estimated integer value of the mass of the respective atom stood in as the weight of the vertices in the residue. Interestingly, the edges of the weighted graph-theoretic model symbolized molecular bonds. Because the hydrogen atom was common among each of the amino acid structures, their work did not give consideration to the hydrogen atom found in the amino acids. Couple of respective vectors of descriptors obtained from some graph theoretic measures included weighted domination, weighted diameter, circumference. Characteristic of the work of Knisley et al. was the eight subsequence partitions which they denoted by $S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8$, that was appropriate for the sequence of CFTR that matched up to the NBD1. The secondary structures of protein served as the guiding rule in determining the eight subsequences mentioned above, such that each subsequence had only one kind of secondary structure. In other words, each of the eight subsequence could only contain a beta strand or an alpha helix, or a loop. No subsequence contained more than one of those secondary structures. In this work, our partitioning will equally be guided by the secondary sequence as did Knisley et al. The graph-theoretic model for NBD1 CFTR by Knisley et al. had edges that depended on the closeness of measure with the distance end point being determined by a threshold distance between any two residue of each subdomain. Noteworthy also is the fact that that three layers were characteristic features of the hierarchical graph for NBD1 of CFTR with the lowest level having an assemblage of twenty small

vertex-weighted graphs, with each depicting one of the twenty typical amino acids. A group of eight graphs with their vertices weighted in which each vertex was a depiction of an amino acid found at the middle level of the hierarchical graph for NBD1 of CFTR. The combinatorial descriptors of the amino acid graphs at the lowest level served as the weights assigned to the vertices of each of the eight midlevel graphs. The weighted graph at the highest level was a pictorial description of the NBD1. Each of the vertices at the highest level of the graph stood in for one and only one of the subdomains G_i with the attributed weights obtained from the vertex-weighted graphs of each subdomain G_i . Using a measure of nearness or distance (proximity measure) of 8 angstroms, the vertices were connected. That is to say that, two vertices in the midlevel graph were connected with an edge if they are 8 angstroms from each other. Their nested graph for NBD1 is shown in Figure 4.

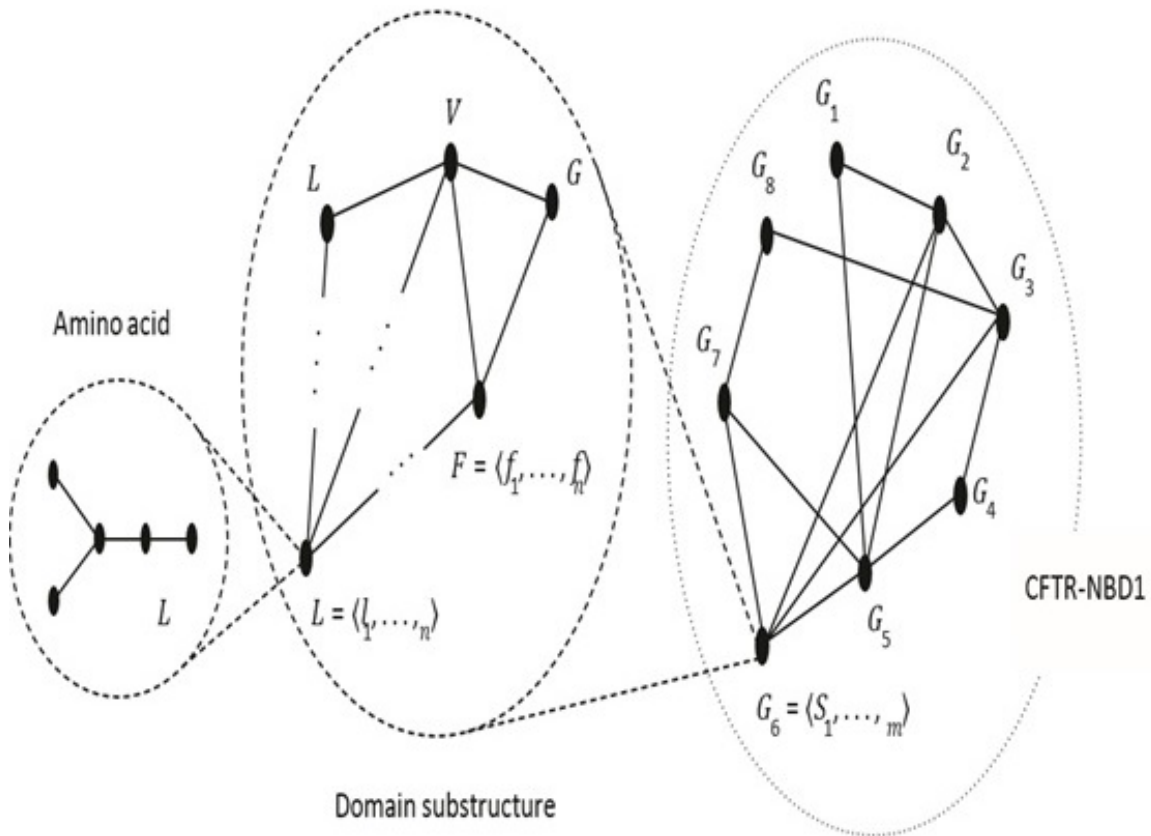


Figure 4: Nested Graph-Theoretic Model for NBD1 by Knisley et al.[36]

In an attempt to examine the impact on the NBD1 on occurrence of mutations, Knisley et al.[36] selected 8 diseases associated with some mutations in the Cystic Fibrosis Mutation Databank that occur in NBD1 after gathering a set of measures for Wildtype NBD1. With these chosen mutations to be included in the model, a set of graph-theoretic measures for each mutation was captured following the procedure described below. Overall structural impact of a single mutation on the NBD1 was captured by effecting a change in the interrelated residue level. One and only one subdomain S_i is aroused by the change that occurred in the interdependent residue level. An online protein folding server called I-TASSER was used to obtain the new

subdomain G_i of the affected subdomain. Illustratively, Figure 5 depicts G_2 that containing $F508$ and the graph with the predicted structural changes upon a consequence of deleting $F508$.

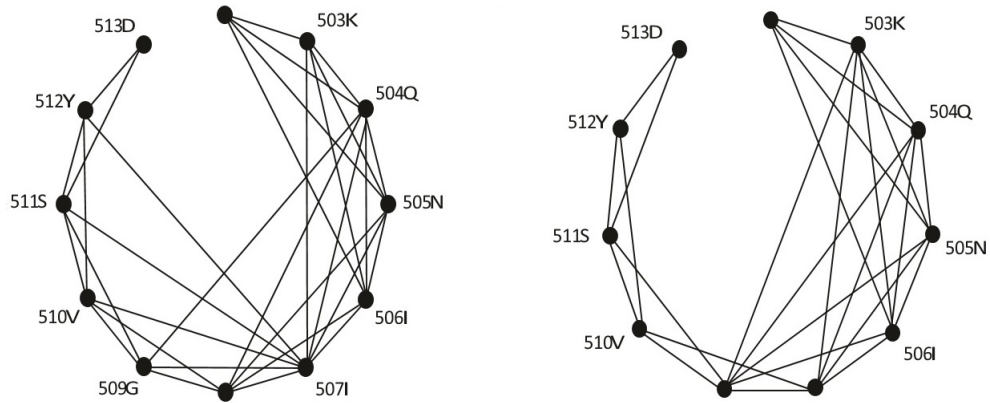


Figure 5: Subdomain graph of G_2 with $508F$ and without $508F$, [36]

Using new set of combinatorial descriptors, Knisley et al. had a dendrogram clustering for the mutations to ascertain how the various studied mutations clustered themselves along the wildtype mutation. This thesis work extends the approach in [36] to building a hierarchical graph for nucleotide binding domain 2. The difference however arises from the improved descriptors that will be used.

3 GRAPHICAL INVARIANTS AS AMINO ACID DESCRIPTORS

A discussion of how invariants of a weighted differ from invariants as applied to unweighted graph is presented in this chapter. The chapter also throws more light on how some combinatorial descriptors or molecular descriptors for the first 20 most essential amino acids were computed. Tables of values for several computed molecular descriptors for these amino acids are also found in this chapter.

3.1 Explanation of Graph Invariants

It is a well established practice in mathematics that we associate numbers with mathematical objects in various ways. Illustratively, a *determinant* (a number) is associated with a matrix, degree (a number) is associated with a polynomial, *dimension* (number) is associated with a space, length (a number) is associated with a vector among others. Several numeric values can also be associated with graphs as well. Usually, such numbers or descriptors are called *graph invariants*. Properties or measures, numbers (descriptors) that are associated with graphs are called “graph invariants” if these numbers or descriptors (quantitative values) are invariant, invariable, constant, changeless, or unchanging under graph isomorphisms: each is a function f , such that $f(G1) = f(G2)$ whenever graphs $G1$ and $G2$ are isomorphic graphs. An isomorphism s from graph G into H is a bijective mapping: that is, $s : V(G) \rightarrow V(H)$ and that preserves adjacency: that is $u \sim v$ if and only if $s(u) \sim s(v)$. In other words, two graphs $G1$ and $G2$ are said to be isomorphic, if they have the same number of vertices, the same number of edges, the same degrees for corresponding vertices, the same number of connected components, the same size of largest clique and smallest circle,

the same number of loops and that adjacency relationship is preserved and so on. In a nutshell, graph invariant is a property, quantitative measure or number assigned to a graph, that is preserved under an isomorphism. Some examples of graph invariants, include the number of vertices (termed the order of the graph), the number of edges (called size of the graph), edge chromatic number (the minimum number of colours needed to color the edges of the graph such that no two adjacent edges have the same colors), genus number, clique number, domination number among others are found below. Figure 6 is an illustrative example of computations from standard definitions of some graph invariants.

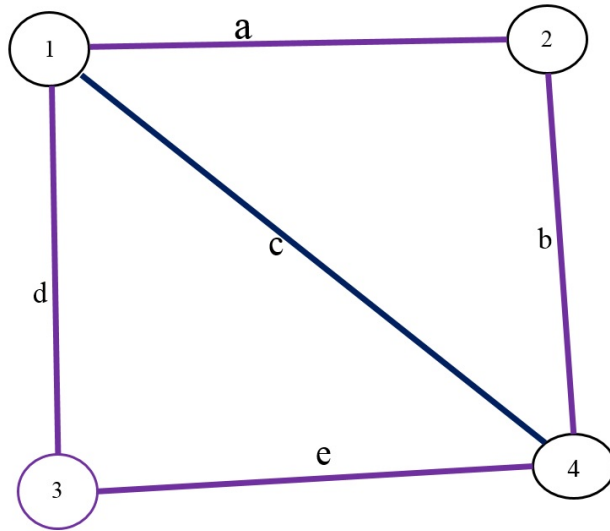


Figure 6: Graph to Illustrate Some Standard Definitions

The *vertex set* of the graph in Figure 6 is $\{1,2,3,4\}$, *edge set* is $\{a,b,c,d,e\}$, order of the graph (number of vertices in the graph) is 4, Size of the graph (number of edges in the graph) is 5, degree of vertex(1) as the number of vertices adjacent to (shares edges with) vertex(1) is 3. The minimum cardinality among the dominating set is called the domination number of the graph. *Dominating sets* are S_1 , S_2 and

$S3$ where $S1$ is $\{\text{vertex}(1)\}$, $S2$ is $\{\text{Vertex}(4)\}$, $S3$ is $\{\text{Vertex}(2), \text{Vertex}(3)\}$. It can be seen from the above that sets $S1$ and $S2$ have the minimum cardinality with the cardinality being 1, hence the domination number of the graph, denoted by $\gamma(G)$ is 1. In other words, to dominate the graph, we need to select only one vertex either $\text{vertex}(4)$ or $\text{vertex}(1)$.

3.2 Molecular Descriptors or Combinatorial Descriptors of Amino Acids

Following earlier successful efforts to model proteins as network with graphs by Knisely et al., and other researches in computational biology and bioinformatics [36, 40], we build a graph-theoretic model for each of the amino acids and then assign quantitative values (molecular descriptors) for each of them. The procedures for finding the molecular descriptors are consistent with all amino acids. While Haynes et al., introduced the use of the domination number of a graph to quantitatively describe a biomolecule [31], Knisley et al., in earlier work on NBD1 [36] and predicting protein-protein interaction [34] used the domination number, coupled with other graph invariants, as a numerical assignment to the amino acid residue structures and built a predictive model for protein-ligand binding affinity. Irrespective of the fact that both of these were successful, the authors were very quick to note the flaw of graphical invariants as molecular descriptors when examining weighted graphs. As it has always been the case when graph invariants are considered, the weights of the vertices are taken to be one. No wonder these measures or estimates are termed invariants since they are invariant or unchanging or changeless under isomorphism. This fact is highly incompatible with weighted graph. As noted by Knisley

et al.[36, 35], and as in the case of weighted graphs studied in this work, we need to modify the definition of graph invariants. If we incorporate the vertex weights for two graphs with isomorphic non-weighted structures, the “invariants” computed for these two graphs will no longer be invariants but will vary considerably based on the weights assigned. With this fact in view, the measures or descriptors we define, although derived from well-established graphical invariants or standard definitions are no longer invariant under isomorphism, since the weights of the vertices are factored into the definition of the measure. Henceforth, we have adopted the term *molecular descriptors* or *combinatorial descriptors* for these values in this work. An illustrative example of a weighted graph with computed molecular descriptors is shown in Figure 7 from definitions adopted by the researcher.

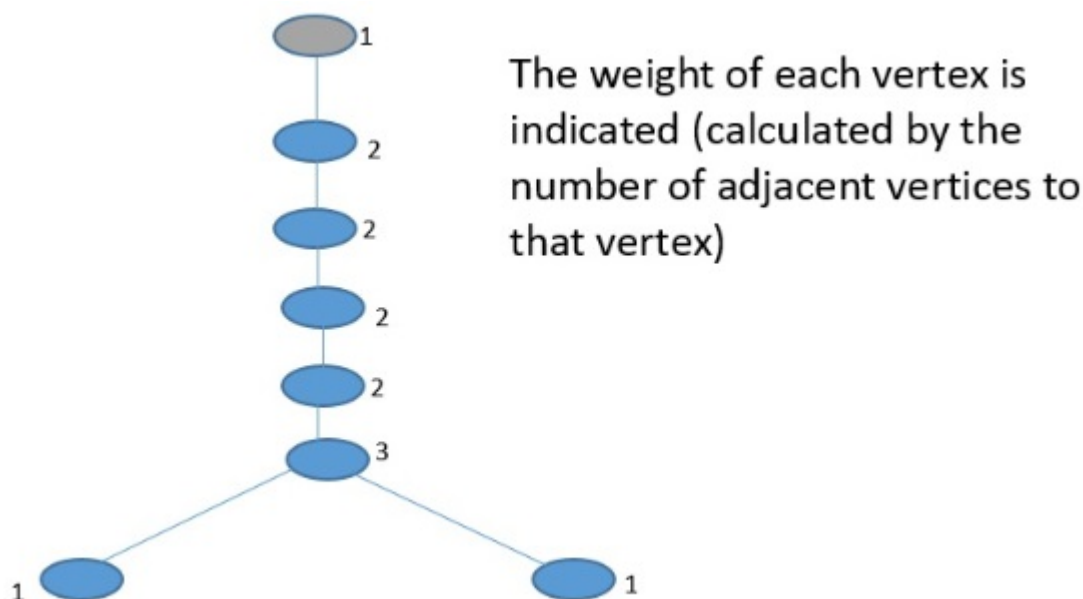


Figure 7: Graph to Illustrate Some Adopted Definitions

The weight of each vertex is indicated (calculated by the number of adjacent vertices to that vertex). The eccentricity of a vertex (u) in a graph is the maximum

distance away from the farthest vertex (k) in the graph. The minimum eccentricity is called a radius but since we are dealing with weighted graphs, we will call it weighted radius. We shall also use weighted eccentricity since we are working with weighted graphs. Se denotes the eccentricity sequence of the graph-theoretic model of Argine. Weighted eccentricity sequence, $Se = \{12, 12, 12, 10, 9, 8, 7, 6\}$, average weighted eccentricity = $(12+12+12+12+10+9+8+7+6)/8 = 9.5$, weighted diameter, $D = \text{maximum value in } Se = 12$, weighted radius, $r = \text{Minimum value in } Se = 6$, average weighted degree = 1.75 (obtained by adding the all weighted degree and dividing by number of vertices). Graph-theoretic model for each of the amino acid using their atomic numbers as weights were obtained and molecular descriptors were obtained for each of the amino acids. Figure 8 depicts a graph-theoretic model based on atomic number assignment as weights to each of the vertices in Tryptophan.

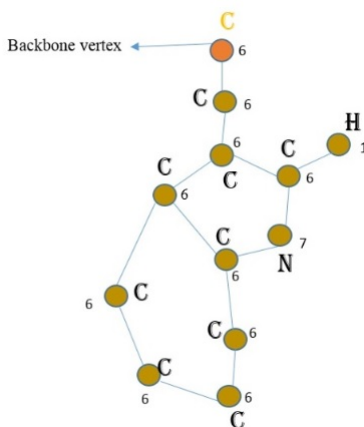


Figure 8: Graph-Theoretic Model for Tryptophan

Molecular descriptors or combinatorial descriptors can be computed by a similar approach by assigning the atomic number as weights in the graph theoretic model as depicted in Figure 8 for the graph-theoretic model for Tryptophan.

3.3 Table of Molecular Descriptors for the 20 Most Essential Amino Acids

Tables 1-3 give some molecular descriptors for the 20 most essential amino acids computed from graph theoretic models of each of these amino acids based on definitions adopted for this study by the researcher.

Table 1: Molecular Descriptors of the 20 Most Essential Amino Acids

Molecule	Symbol	d1	d2	d3	d4	d5	d6	d7	d8
Arginine	R	8.00	7.00	12.00	6.00	8.120	6.00	12.00	1.50
Histidine	H	7.00	6.00	14.00	6.000	6.71	6.000	9.00	2.00
Lysine	K	6.00	5.00	10.00	4.00	7.00	5.00	9.00	1.667
Aspartic Acid	D	5.00	4.00	8.00	4.00	5.17	3.00	6.00	1.60
Glutamic Acid	E	6.00	5.00	10.00	5.00	6.00	4.00	8.00	1.667
Serine	S	3.00	2.00	4.00	2.00	1.670	2.00	3.00	1.333
Threonine	T	4.00	3.00	6.00	3.00	3.250	1.00	4.00	1.50
Asparagine	N	5.00	4.00	8.00	4.00	5.00	3.00	6.00	1.60
Glutamine	Q	6.00	5.00	10.00	4.00	5.860	4.00	8.00	1.667
Cysteine	C	3.00	2.00	4.00	2.00	2.33	1.00	3.00	1.333
Glycine	G	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Proline	P	4.00	4.00	8.00	4.00	4.00	4.00	4.000	2.00
Alanine	A	2.00	1.00	2.00	1.00	1.00	1.00	1.00	1.00
Isoleucine	I	5.00	4.00	8.00	4.00	3.25	3.00	6.00	1.600
Valine	V	4.00	3.00	6.00	3.00	3.25	1.00	4.00	1.50
Leucine	L	5.00	4.00	8.00	4.00	5.00	3.00	6.00	1.60
Methionine	M	5.00	4.00	8.00	4.00	5.40	3.00	7.00	1.60
Phenylalaine	F	8.00	8.00	14.00	6.00	7.00	6.000	11.000	1.750
Tyrosine	Y	9.00	9.00	18.00	7.00	8.88	6.000	13.000	2.000
Tryptophan	W	11.00	12.00	24.00	8.00	11.10	9.000	14.000	2.182

Table 2: Molecular Descriptors of the 20 Most Essential Amino Acids Continued 1

Molecule	Symbol	<i>d</i> 9	<i>d</i> 10	<i>d</i> 11	<i>d</i> 12	<i>d</i> 13	<i>d</i> 14	<i>d</i> 15	<i>d</i> 16
Arginine	R	12.499	-4.307	3.500	-2.590	19.00	31.444	20.00	38.00
Histidine	H	12.876	-3.721	4.286	-1.185	15.00	23.10	18.00	31.00
Lysine	K	10.363	-3.151	3.00	-0.536	12.00	24.50	18.00	31.00
Aspartic Acid	D	11.539	-4.178	3.20	0.528	12.00	16.40	12.00	20.00
Glutamic Acid	E	11.530	-3.425	3.333	-0.538	12.00	21.00	14.00	26.00
Serine	S	5.00	1.00	2.00	2.00	6.00	13.33	8.00	20.00
Threonine	T	9.928	-3.928	3.00	3.00	6.00	12.40	8.00	14.00
Asparagine	N	11.539	-4.178	3.20	0.528	12.00	16.50	14.00	20.00
Glutamine	Q	12.207	-4.255	3.333	-1.043	12.00	21.167	15.00	24.00
Cysteine	C	6.243	-2.243	2.00	2.00	6.00	16.670	12.00	22.00
Glycine	G	0.00	0.00	0.00	0.00	1.00	3.50	1.00	6.00
Proline	P	12.00	-4.00	4.00	4.00	12.00	12.00	12.00	12.00
Alanine	A	2.00	0.00	1.00	2.00	6.00	6.00	6.00	6.00
Isoleucine	I	10.851	-6.085	1.80	-1.517	12.00	15.60	12.00	18.00
Valine	V	9.928	-3.928	3.00	3.00	6.00	10.50	6.00	12.00
Leucine	L	11.029	-4.729	3.20	1.052	12.00	15.60	12.00	18.00
Methionine	M	9.49	-2.812	2.80	0.678	18.00	27.20	18.00	34.00
Phenylalaine	F	14.851	-4.801	4.25	-1.672	18.00	23.25	18.00	24.00
Tyrosine	Y	12.868	-4.793	4.333	-2.054	18.00	27.78	20.00	38.00
Tryptophan	W	13.511	-6.324	4.00	-2.576	24.00	27.50	18.00	36.00

Table 3: Molecular Descriptors of the 20 Most Essential Amino Acids Continued 2

Molecule	Symbol	<i>d17</i>	<i>d18</i>	<i>d19</i>	<i>d20</i>	<i>d21</i>	<i>d22</i>
Arginine	R	45.00	5.00	23.343	0.00	10.667	4.20
Hisitidine	H	47.00	4.70	24.243	-1.734	10.400	1.605
Lysine	K	37.00	6.17	22.739	-0.179	10.167	1.372
Aspartic Acid	D	34.00	6.80	28.634	0.00	10.40	2.969
Glutamic Acid	E	40.00	6.67	28.731	0.00	10.667	1.822
Serine	S	22.00	7.33	20.00	0.00	8.667	6.00
Threonine	T	27.00	5.40	23.819	-4.227	9.00	6.00
Asparagine	N	33.007	6.60	27.708	0.00	10.00	3.00
Glutamine	Q	39.00	6.50	27.831	0.00	10.50	1.849
Cysteine	C	28.00	9.33	28.00	0.00	11.333	6.00
Glycine	G	7.00	3.50	7.00	0.00	3.50	0.00
Proline	P	24.00	6.00	24.00	0.00	12.00	12.00
Alanine	A	12.00	6.00	12.00	0.00	6.00	0.00
Isoleucine	I	30.00	6.00	24.841	-1.641	9.60	3.373
Valine	V	24.007	6.00	24.00	0.00	9.00	6.00
Leucine	L	30.00	6.00	25.021	0.00	9.60	3.113
Methionine	M	40.00	8.00	31.344	0.00	13.60	2.656
Phenylalaine	F	48.00	6.00	26.993	0.00	12.00	2.026
Tyrosine	Y	56.00	6.22	28.252	-0.96	12.222	1.599
Tryptophan	W	68.00	5.667	29.778	0.211	12.75	2.044

The molecular descriptors of combinatorial descriptors in the Tables 1, 2, 3 were computed from a graph-theoretic model based on weighted degree and assignment of atomic numbers as degrees of each vertex. **Keys:** *d1* = number of vertices (order of the graph), *d2* = number of edges (size of the graph), *d3* = total weighted degree of the graph (obtained by adding all the weights of all the vertices), *d4* = weighted domination number, *d5* = average eccentricity, *d6* =radius (minimum eccentricity), *d7* = diameter (maximum eccentricity), *d8* = average weighted degree (total degree

divided by the number of vertices), $d9$ = maximum eigenvalue of the weighted Laplacian matrix of the graph, $d10$ = minimum eigenvalue of the weighted Laplacian matrix of the graph, $d11$ = Average eigenvalue of the Laplacian matrix of the the graph, $d12$ = second smallest eigenvalue of the Laplacian matrix of the graph. Using the atomic numbers as weights of vertices in the graph theoretic model of each of the amino acids, we obtain the following descriptors in Tables 2-3: $d13$ = weighted domination number using the atomic number, $d14$ = average weighted eccentricity based on the the atomic number, $d15$ = weighted radius based on the atomic number (minimum eccentricity), $d16$ = weighted diameter based on the atomic number (maximum eccentricity), $d17$ = total weighted atomic number of the graph (obtained by summing all the atomic number of each of the vertices in the graph), $d18$ = average weighted atomic number or degree based on atomic number in the graph. Descriptors $d19$ through $d22$ in the Tables 1, 2 and 3 were obtained from weighted Laplacian matrix, $d19$ = weighted maximum eigenvalue based on atomic number, $d20$ = weighted minimum eigenvalue based on the atomic numbers, $d21$ = weighted average eigenvalue based on the atomic numbers, and $d22$ = weighted second smallest eigenvalue of the weighted Laplacian matrix.

4 THE HIERARCHICAL/ NESTED GRAPH MODEL OF NBD2

A discussion on cystic fibrosis and CFTR is presented in this chapter. The discussion includes the prevalence of cystic fibrosis, how the disease comes about, and how cystic fibrosis affects the function of several organs as well. How a single point mutation in the NBD2 of CFTR has such structural consequences for the domain is well elaborated in this chapter. The chapter also offers explanation to how we modeled NBD2 with a hierarchical graph.

4.1 Cystic Fibrosis and CFTR

One of the most prevalent inherited diseases is cystic fibrosis. This disease is caused by a mutation in a membrane protein, the cystic fibrosis transmembrane conductance regulator (CFTR) [11]. The most prevalent genetic disorder among the Caucasian population (Europe, North America, among others) is cystic fibrosis. Available statistics from Cystic Fibrosis Foundation indicates that about 30,000 people (adults and children) in the United States and 70,000 worldwide have cystic fibrosis with 1000 new cases diagnosed each year in United State of America [7]. People who have CF inherited a defective gene. A single point mutation in the CFTR protein causes cystic fibrosis (CF). When a severe mutation occurs in CFTR protein, this can affect the transportation of water and salt thereby causing the mucus that found in the tube of several organs like the lungs, pancreas and reproductive organs to thicken. When the mucus thickens resulting from severe mutation in CFTR protein, this harbors infections especially respiratory infections occurring with several clinical consequences including the malfunctioning of these organs. Even though the two major systems af-

affected are the lungs and the gastrointestinal tract, several other organs of the human body such as pancreas, reproductive organs, liver, gall bladder, salivary gland and the colon are affected, due to occurrence of a mutation in this membrane protein. Even though more than one thousand nine hundred different mutations of CFTR, with various levels of severity of clinical consequence are reported, an estimated 5% of the Caucasian population are affected by mutation in the CFTR [7]. Despite the large number of reported mutations of CFTR, the deletion of phenylalanine at position 508 ($\Delta F508$) occurs in more than 90% of the CF population, while substitution of Lysine with Asparagine at position 1303 (N1303K) accounts for estimated 2.5% of all the CF population. The N1303K as a mutation, is linked with defective protein processing and results in the absence CFTR on the surface, its subsequent effect on the entire protein domain. N1303K mutation results in one of the more severe phenotypes [7, 48, 1, 10, 28]. Irrespective of the fact that there have been substantial advances in science and medical researches, we still lack an adequate understanding of how just a single point mutation in this membrane protein can have such a devastating effect on this protein domain. Currently, no literature exists on using graph-theoretic model (nested graph) in studying the effect of single point mutation on the NBD2. In view of this, this work is the first literature on using a graph-theoretic model for NBD2 of CFTR to study the effect of a single point mutation on NBD2. In this study, a mathematical model using graph theory to exam the impact of a single point mutation known in NBD2 to cause cystic fibrosis is presented. We compare the wild type structure with eight of the most prevalent mutations. Using the graph-theoretic model for NBD2, we can gain a meaningful insight into how NBD2 is affected by an

occurrence of a single point mutation in this domain protein. In other words, by way of hierarchical graph, we sort to probe into how a single point mutation of NBD2 of CFTR can affect the structure and function of this protein domain, NBD2.

4.2 The Model for Nucleotide Binding Domain 2 (NBD2)

If two vertices share an edge, they are adjacent. In real life application, this can describe a affiliation or association among alike entities. For example, we might say that if two people stay about 8 miles apart, then they are neighbors or friends. In which case, in the graph-theoretic model, an edge will be incident to these two people. An immediacy or simply proximity graph is created where the vertices harmonize or coincides with objects (amino acids) should they be within a given distance from one another, the said vertices under consideration are said to be adjacent. The *3GD7* [14] file from the Protein Data Bank [15] was used. Using the amino acid sequence, NBD2 was enclosed and captured with the subsequence 1209 – 1394. Partition or stratification of the amino acids into even smaller subdomains was obtained from this. The subdomains notably *S1*, *S2*, *S3*, *S4*, *S5*, *S6*, *S7*, *S8*, and *S9* are sequences of amino acids that differ on the existence of alpha helices and beta strands within their structure. The subdomains mostly differ in length from 10 to 18 amino acids with the exception of cases where there were unique reasons to violate this length. Tables 4 and 5 explain our partitioning and reasoning employed for the choice of this partition. From the subdomain, a proximity graph, with a threshold of 8 angstroms, was created. The procedure employed in the study is analogous to that used by Knisley et al. their work on NBD1 [36]. However, the differences arise from the

improved molecular descriptors (graph invariants) used for this study and the fact that our graph-theoretic model is for NBD2, instead of NBD1. Besides, the mutations studied in this work are all found in NBD2 while Knisley et al. concerned themselves with mutations in NBD1.

Table 4: Subdomain, Subsequence, Amino Acid Sequence

Subdomain	Subsequence	Amino Acid Sequence
S1	1209..1224	QMTVKDLTAKYTEGGN
S2	1225..1238	AILENISFSISPGQ
S3	1239..1261	RVGLLGRTGSGKSTLLSAFLRL
S4	1262..1277	NTEGEIQIDGVSWDSI
S5	1278..1305	TLEQWRKAFGVIPQKVFIFSGTFRKNLD
S6	1306..1324	PNAAHSDQEIWKVADEVGL
S7	1325..1340	RSVIEQFPGKLDVFLV
S8	1341..1364	DGGCVLSHGKQLMCLARVLSKA
S9	1365..1391	KILLDEPSAHLDPVTYQIIRRTLKQA

Table 5: Subdomain, Subsequence, Reason

Subdomain	Subsequence	Reason
S1	1209..1224	beta strand, binding site, turn, bend
S2	1225..1238	binding site, beta strand, turn
S3	1239..1261	beta strand, binding site, turn, bend, alpha-helix
S4	1262..1277	bend, beta strand, turn
S5	1278..1305	binding site, alpha helix, bend, turn, beta strand
S6	1306..1324	turn, bend, alpha helix
S7	1325..1340	bend, turn, alpha helix, beta strand, 3/10 -alpha helix
S8	1341..1364	turn, bend, alpha helix, beta strand
S9	1365..1391	alpha helix, beta strand, bend

Figure 9 depicts the midlevel graphs for corresponding to subdomain *S5*, and *S4* at Threshold of 8 angstroms.

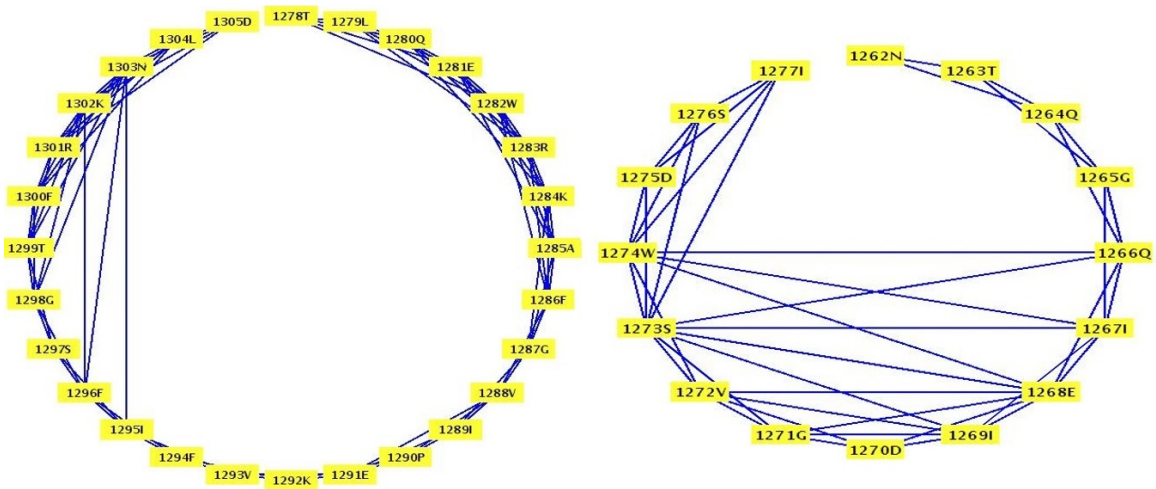


Figure 9: Midlevel Graph for Subdomain S5 (on left) and S4 (on right)

The graph in Figure 10 is the hierarchical or nested graph for Nucleotide Binding Domain 2 (NBD2) of Cystic Fibrosis Transmembrane Conductance.

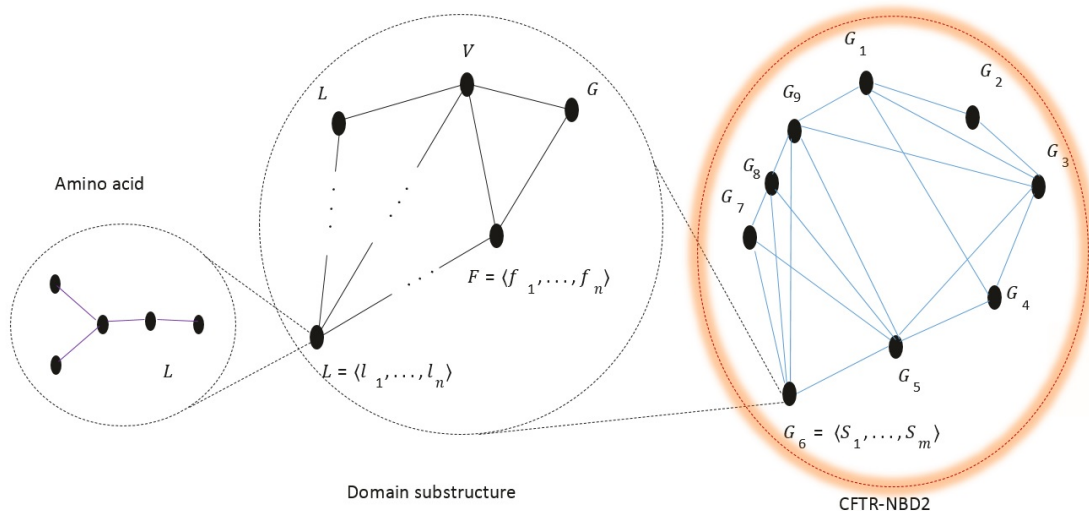


Figure 10: Hierarchical Graph for NBD2

The edges of the NBD2 of CFTR Graph, or commonly referred to as the domain

graph G (refer to Figure 10), are based on an immediacy or closeness measure where the distance end points are decided on basis that if two vertices (two residues of each subdomain) are 8 angstroms away from each other, they are connected with an edge otherwise no edge, otherwise no edge is connected between them. The hierarchical graph for NBD2 of CFTR is symbolized by 3 layers. For the lowest level graph, we have a set of 20 small vertex-weighted graphs, denoting the 20 most main amino acids as depicted in Figure 10. Characteristic of the middle level graph is nine vertex-weighted graphs G_i , with each vertex denoting an amino acid while the weights of the vertices are the combinatorial descriptors or molecular descriptors (invariants) of the amino acid graphs at the lower level. At the top level graph, there are vertex-weighted graph G that exemplify the nucleotide binding domain NBD2. Each of the vertices represent one of the subdomain graphs G_i . It is interesting to note that the weights assigned to these vertices are derived from the vertex-weighted graph-theoretic descriptors adopted by the researcher for this work. First, each vertex in the lower level graph was assigned a weight based on the number of vertices incident to it. We computed e_{aw} , to the nearest 2 decimal place the nearest hundredth, where e_{aw} is the weighted average eccentricity of each vertex in the amino acid. The sum of vertex weights from a vertex u to another vertex v farthest from u determined the weighted eccentricity. The sum excluded the weight of the vertex under consideration since in a simple graph, no vertex is adjacent to itself or there is no self-loop. A list of all weighted eccentricity was obtained for each amino acid in the lower level graph. The weighted average eccentricity was found by summing all the weighted eccentricity of each of the vertex in the lower level graph and dividing by the total number of

vertices (order of the graph) . Second, each vertex in the midlevel graph (labelled as domain substructure in Figure 10) was assigned e_{aw} computed from first step described above. C_w , to the nearest 2 decimal place, was computed for the midlevel graph, where C_w is the total weighted circumference of the midlevel graph. Third, there were 9 vertices in the top level graph (domain graph) depicted in Figure 10. Each vertex in the top level graph represents one and only one subdomain graph (midlevel graph). C_w (described above) that corresponds to each vertex in the top level graph was assigned. Upon the assignment of C_w to each vertex, the weights (d_1) of each vertex was computed by the summing all C_w of its adjacent vertices. Several combinatorial algorithms like total weight (sum of all weights), average weight (sum of weights of the top level graph divided by the number of vertices) were obtained. Weighted adjacency matrix was found for each of the vertices in the top level graph. Suppose the top level graph had only two vertices u and v, and u is adjacent to v in the top level graph, v had a weight of 4.12 while u is of weight 6.19, then their weighted adjacency matrix, T is given as $T = \begin{pmatrix} 0 & 4.12 \\ 6.19 & 0 \end{pmatrix}$.

Other molecular descriptors (measures) like weighted connectivity (obtained by summing all entries on each row of the adjacency matrix, for each of the vertices of the top level graph were also found and incorporated into the molecular descriptors for the top level graph). These weights obtained for the top level graph are the molecular weights or descriptor (invariants) for the wildtype domain graph.

5 THE EFFECT OF SINGLE-POINT MUTATIONS ON NBD2 AS SHOWN BY THE MODEL

How our model is used in studying the effect of a single point mutation on the NBD2 is explained in this chapter. The chapter also presents some brief discussion on existing knowledge of some mutations in NBD2 and how they associate with cystic fibrosis. A dendrogram clustering of single point mutations resulting from application of single point mutation on our model is enshrined in this chapter. Discussion of our results is equally presented in this chapter.

5.1 Some Known Mutation in NBD2 and Association with Cystic Fibrosis

Existing body of knowledge of mutations in NBD2 based on cystic fibrosis, CF [41, 25, 40, 43] puts the following mutations into the categories in the Table 6.

Table 6: Classification of Mutations Based on Cystic Fibrosis [41, 25, 40, 43]

Mutation	CF
Wildtype	No
Y1212G	Mild
G1271E	Mild
S1347R	Mild
I1234V	Mild
D1270N	Mild
V1212W	Mil
S1235R	Mild
N1303K	Severe

5.2 Application of a Single Point Mutation on the Model for NBD2

Using the model built for NBD2, we can study the effect of single point mutation on the entire protein domain. The weights obtained for the top level graph are the molecular weights or descriptor (invariants) for the wildtype domain graph. Since the purpose of building a hierarchical graph for NBD2 is to use the graph to study the effect on the entire domain (NBD2) when a single point mutation takes place, in an attempt to capture the effect of each mutation on the top level graph, the entire process was repeated one at a time for each of these single point mutations (*Y1212G*, *G1271E*, *S1347R*, *I1234V*, *D1270N*, *V1212W*, *S1235R* and *N1303K*) and resulting molecular or combinatorial descriptors (measures or invariants) computed earlier, were recalculated for midlevel graph and subsequently the top level graph. Different set of graph theoretic measures were obtained for each mutation. Illustratively, suppose we examine subdomain 5: the midlevel graph that contains this mutation is shown in Figure 11. After mutation *N1303K* occurs, Asparagine replaces Lysine at position 1303. The structure of the midlevel graph does not change, but the vertex weights for the corresponding vertex at position 1303 changes due to the substitution of Lysine with Asparagine. Eight different mutations of CFTR were used for this study, as can be seen in the Results section. These changes both take place in the midlevel graph and the top level graphs and help to find the resulting structural effect of each of these single point mutations. A dendrogram depicting the clustering of the various mutations mentioned above together with the wildtype is shown in the results. Below is a table of values of some combinatorial descriptors or molecular descriptors of the top level graph after performing each the mutations (one mutation at a time).

Table 7: Top Level Graph Molecular Descriptors for Single Point Mutations

Mutation	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
Wildtype	4.42	2.99	4.93	0.90	8.12	13.95	7.66	18.31	13.20	24.60
Y1219G	4.33	2.90	4.49	0.82	7.37	13.65	6.67	18.00	13.11	24.43
V1212W	3.15	3.60	5.14	0.92	8.23	13.61	6.39	19.24	11.93	24.15
I1234V	4.67	2.99	4.93	0.90	8.10	14.20	7.90	18.30	13.44	24.85
S1235R	3.09	2.99	4.93	0.91	8.18	12.62	6.40	18.38	11.94	23.27
G1271E	4.04	2.99	4.93	0.91	8.23	13.57	7.40	18.55	13.06	23.99
S1347R	4.42	2.99	5.19	0.91	8.18	14.21	7.66	18.64	13.27	24.98
N1303K	4.44	3.01	4.95	0.91	8.14	14.01	7.68	18.35	13.22	27.95
D1270N	4.53	2.99	4.93	0.91	8.23	14.06	7.89	18.54	13.54	24.71

Keys for Table 7:

t1 = average of non-zero numbers in column 3 of the weighted adjacency matrix,
t2 = average of non-zero numbers in column 4 of the weighted adjacency matrix,
t3 = average of non-zero numbers in column 9 of the weighted adjacency matrix,
t4 = average weighted degree of the top level graph (divided by thousand), t5 =
total weighted degree of the top level graph (divided by thousand), t6 = weighted
connectivity for row 1 of the weighted adjacency matrix (obtained by summing all the
numbers on row 1), t7 = weighted connectivity for row 2 of the weighted adjacency
matrix, t8 = weighted connectivity for row 3 of the weighted adjacency matrix, t9 =
weighted connectivity for row 4 of the weighted adjacency matrix, and t10 = weighted
connectivity for row 5 of the weighted adjacency matrix.

5.3 Clustering of Single Point Mutations/ Results of Single Point Mutations

The R statistical Software [4] was used to cluster the mutations using the molecular descriptors (combinatorial descriptors) for the top level graph when the single point mutations were performed. The single linkage function in R was used for our hierarchical clustering because it is less biased. The dendrogram clusters (shown in Figure 13) the wildtype mutation and other mutations using the combinatorial or molecular descriptors from Table 6.

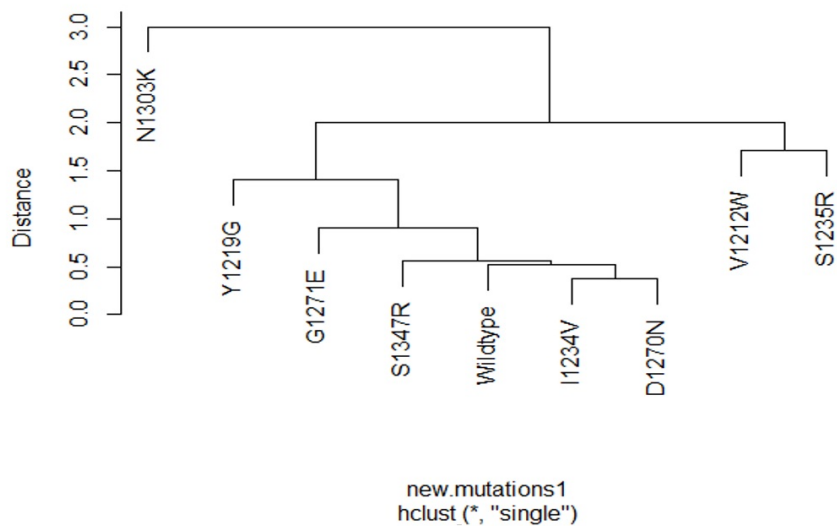


Figure 11: Clustering of Mutations, Output from R [4]

5.4 Discussion of Results

The *N1303K* mutation is one of the known mutations in NBD2 that causes cystic fibrosis. The evidence that the substitution of *N* with *K* at position 1303 leads to

variation in the arrangement of the molecule when folded in the lab has baffled researchers in their attempt to explain why the molecule does not fold appropriately in the cell. *N1303K* is said to be linked to pancreatic insufficiency cystic fibrosis[42, 8]. Our results (refer to Figure 11) show that the resulting structural effects of *N1303K* are expressively distinct from the wildtype. Also, it is obvious from our results that the difference between the wildtype and domain graphs caused by mutations like *I1234V*, *S1345R* and *D1270N* are less significant. More so, our results lead to a conclusion that *Y1219G*, *G1271E*, *V1212W* and *S1235R* are also considerably distinct from wildtype, even though they all belong to one bigger cluster. Our results call for the need for further investigations. For instance, thought provoking questions like, under what circumstance would *N1303K* match up to or mirror wildtype? In other words, what graph-theoretic or combinatorial descriptors of the graph containing *N1303K* would result in a graph that is very similar to wildtype or will cause the clustering of *N1303K* along the wildtype or other mild mutations? Answers to such questions are of paramount importance to us since it can lead us to gain a useful discernment into line of action for design of a molecule that can correct this specific mutation associated with cystic fibrosis.

6 CONCLUSION

In this chapter, we link our findings to existing body of knowledge on graph-theoretic models, main results are highlighted with appropriate recommendations made as to further researches. A summary of the entire work is also presented in this chapter.

6.1 Linking Findings to Existing Literature on Graph-Theoretic Models

The study was successful at building a graph-theoretic model for NBD2 and subsequently using the graph in examining the impact of single-point mutations on the NBD2 of CFTR. This work, though the first on a graph-theoretic model for NBD2 of CFTR, adds up to existing literature on graph-theoretic models for studying biological systems. Knowledge regarding the consequences of N1303K and other mutations is essential for drug design to treat cystic fibrosis. Like Knisley et al. [36], the results of this study point to the direction that graph-theoretic modeling holds a great potential as equipment in the search for appropriate design of drugs for the treatment of cystic fibrosis. Our findings indicate the existence of an obvious correlation between the molecular descriptors or combinatorial descriptors (invariants) of the proximity graphs of several respective clusters and their mutations. It can be argued that this is not a mere happening since functional similarities are evident from structural similarities.

6.2 Future Research Directions/ Open Problems

With the results of this study in view, the following questions can be asked:

- What is responsible for this observed relationship between these respective clusters?
- Why does N1303K cluster separately?
- Can similar graph-theoretic modeling approach be applied to study all other mutation-causing diseases and possibly suggest line of action for drug design for those diseases?

Questions such as these are worth considering as regards the problem at hand, and are relevant to ongoing research in computational biology. Similar graph-theoretic models can be built for all other mutation causing diseases to gain a meaningful insight into them and possibly suggest the line of drug design in treatment of those diseases.

6.3 Summary

One of the most prevalent inherited diseases among Caucasians is cystic fibrosis. This disease is caused by a mutation in a membrane protein, the cystic fibrosis transmembrane conductance regulator (CFTR). CFTR is known to function as a chloride channel that regulates the viscosity of mucus that lines the ducts of a number of organs. Generally, most of the prevalent mutations of CFTR are located in one of two nucleotide binding domains, namely, the nucleotide binding domain 1 (NBD1). However, some mutations in nucleotide binding domain 2 (NBD2) can equally cause cystic fibrosis. In view of the fact that currently, there exists no literature on building a graph-theoretic model for NBD2 and using the using graph-theoretic model in

studying the effect of single point mutation on the NBD2, this work becomes the first in this direction. A mathematical model using graph theory to exam the impact of a single point known in NBD2 to cause cystic fibrosis is presented in this research work. In this work, a model for NBD2 is built using a hierarchical graph. With the use of this model for NBD2, we examine the impact or consequence of single point mutations on NBD2. For each atom in the structure of an amino acid residue, we symbolize it by a vertex in the lowest level of the graph. As regards the residues, we represent them by vertices in the midlevel grpah. The subdomain vertices are each represented by a vertex in the toplevel graph of NBD2. Using this model for NBD2, we examine the impact or consequence of single point mutations on NBD2. We collate the wild type structure with eight of the most prevalent mutations and observe how the NBD2 is affected by each of these mutations. A meaningful insight into the profound structural effect of a single point mutation on the NBD2 is gained using the nested graph for NBD2.

BIBLIOGRAPHY

- [1] Amino acid disorder screening. <http://medical-dictionary.thefreedictionary.com/Amino+Acid+Disorders+Screening>, 2008. Retrieved on 3/20/2015.
- [2] Secondary structure of proteins. <http://encyclopedia2.thefreedictionary.com/Secondary+structure+of+proteins>, Retrieved on 1/25/2015.
- [3] An introduction to protein molecules: The building blocks of life. <http://www.brighthub.com/science/medical/articles/6050.aspx>, Retrieved on 2/15/2015.
- [4] R: A language and environment for statistical computing. r foundation for statistical computing. <http://www.R-project.org/>, Retrieved on 3/12/2015.
- [5] Functions of proteins. <http://www.123helpme.com/preview.asp?id=141063>, Retrieved on 3/15/2015.
- [6] Three dimensional structures of protein. <http://www.ukessays.com/essays/biology/the-three-dimensional-structures-of-proteins-biology-essay.php>, Retrieved on 3/16/2015.
- [7] About cystic fibrosis. <http://www.cff.org/AboutCF>, Retrieved on 3/20/2015.
- [8] Cftr2. http://cftr2.org/mutation.php?view=scientific&mutation_id=4, Retrieved on 3/20/2015.

- [9] Counting point mutations. <http://rosalind.info/problems/hamm/>, Retrieved on 3/20/2015.
- [10] Mutations. <http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/M/Mutations.html>, Retrieved on 3/20/2015.
- [11] What are the principal roles of protien? <http://healthyeating.sfgate.com/principal-roles-protein-body-3678.html>, Retrieved on 3/20/2015.
- [12] <https://www.flickr.com/photos/ajc1/464066753/>, Retrieved on 3/21/2015.
- [13] Role of proteins in the body. <http://sciencelearn.org.nz/Contexts/Uniquely-Me/Science-Ideas-and-Concepts/Role-of-proteins-in-the-body>, Retrieved on 3/3/2015.
- [14] Crystal structure of human nbd2 complexed with n6-phenylethyl-atp (p-atp). <http://www.rcsb.org/pdb/explore/explore.do?structureId=3gd7>, Retrieved on 4/6/2014.
- [15] The protein data bank. <http://www.pdb.org>, Retrieved on 4/6/2014.
- [16] G. Agnarsson and R. Greenlaw. *Graph Theory-Modeling, Application and Algorithms*, volume 1. Pearson Education, Inc., 2007.
- [17] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walters. *Molecular Biology of the Cell*, volume 2nd Edition. New York and London: Garland Science, 2002. The Shape and Structure of Protiens.

- [18] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanel, I. Venger, and S. Pietrokovski. Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology*, 344(4):1135 – 1146, 2004.
- [19] C. B. Anfinsen. The formation and stabilization of protein structure. *Biochemical Journal*, pages 737–749, 2013.
- [20] S. Atwell, S. Antonysamy, W.B. Guggino, K. Connors, S. Emtage, T. Gheyi, J.F. Hunt, H.A. Lewis, F. Lu, J.M. Sauder, P.C. Weber, D. Wetmore, and X. Zhao. Crystal structure of human nbd2 complexed with n6-phenylethyl-atp (p-atp). <http://www.rcsb.org/pdb/explore/literature.do?structureId=3GD7&bionumber=1>, Retrieved on 4/2/2015.
- [21] G. Böhm and R. Jaenicke. Correlation functions as a tool for protein modeling and structure analysis. 1992.
- [22] K.V. Brinda and S. Vishveshwara. A network representation of protein structures: Implications for protein stability. *Biophysical Journal*, 89(6):4159 – 4170, 2005.
- [23] J.M. Chen, C. Cutler, C. Jacques, G. Bœuf, E. Denamur, G. Lecointre, B. Mercier, G. Cramb, and C. Férec. A combined analysis of the cystic fibrosis transmembrane conductance regulator: Implications for structure and disease models. *Molecular Biology and Evolution*, 18(9):1771–1788, 2001.
- [24] E. Coutsiias, A. Seok, D. Chaok, and Ken A. Using quaternions to calculate rmsd. *Journal of Computational Chemistry*, 25, 2004.

- [25] G. R. Cutting, L. M. Kasch, B. J. Rosenstein, J. Zielenski, L.C. Tsui, S. E. Antonarakis, and H. H. Kazazian Jr. A cluster of cystic fibrosis mutations in the first nucleotide-binding fold of the cystic fibrosis conductance regulator protein. 1990.
- [26] N. Deo and F. George. *Graph Theory with Application to Engineering and Computer Science*, volume 2nd edition. Prentice Hall of India Private Limited, New Delhi, 1984.
- [27] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. OUP Oxford, 2013.
- [28] J. Drake. Comparative rates of spontaneous mutation. 1996.
- [29] United Nations Food and Agriculture Organisation. *Human Nutrition in the Developing World*. FAO Corporate Document Repository. Agriculture and Consumer Protection, 1997.
- [30] R. C. Burghardt G. A. Johnson S. W. Kim X. L. Li M. C. Satterfield G. Wu, F. W. Bazer and T. E. Spencer. Impacts of amino acid nutrition on pregnancy outcome in pigs: mechanisms and implications for swine production. *Journal of Animal Science*, 88(13):E195–E204, 2010.
- [31] T.W. Haynes, D. Knisley, E. Seier, and Y. Zou. A quantitative analysis of secondary rna structure using domination based parameters on trees. *BMC Bioinformatics*, 7(1):108, 2006.

- [32] S.J. Hubbard, F Eisenmenger, and J.M. Thornton. Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Protein Science*, 3(5):757–768, 1994.
- [33] A. Jamerson. What are the principal roles of protein in the body? <http://healthyeating.sfgate.com/principal-roles-protein-body-3678.html>, Retrieved on 3/7/2015.
- [34] D. Knisley and J. Knisley. Predicting protein–protein interactions using graph invariants and a neural network. *Computational Biology and Chemistry*, 35(2):108 – 113, 2011.
- [35] D. Knisley and J. Knisley. Vertex-weighted graphs and their applications. *Util. Math.*, 94:237–249, 2014.
- [36] D. Knisley, J. Knisley, and A.C. Herron. Graph-theoretic models of mutations in the nucleotide binding domain 1 of the cystic fibrosis transmembrane conductance regulator. *Computational Biology Journal*, page 9, 2013.
- [37] J. Kyte and R. F. Doolittle. A simple method for displaying the hydrophobic character of a protein. *Journal of Molecular Biology*, 157(1):105 – 132, 1982.
- [38] H.A. Lewis, C. Wang, X. Zhao, Y. Hamuro, K. Connors, M.C. Kearins, F. Lu, J.M. Sauder, K.S. Molnar, S.J. Coales, P.C. Maloney, W.B. Guggino, D.R. Wetmore, P.C. Weber, and J.F. Hunt. Structure and dynamics of {NBD1} from {CFTR} characterized using crystallography and hydrogen/deuterium exchange mass spectrometry. *Journal of Molecular Biology*, 396(2):406 – 430, 2010.

- [39] A.R. Mashaghi, A. Ramezanzpour, and V. Karimipour. Investigation of a protein complex network. *The European Physical Journal B - Condensed Matter and Complex Systems*, 41(1):113–121.
- [40] L. Osborne, R. Knight, G. Santis, and M. Hodson. A mutation in the second nucleotide binding fold of the cystic fibrosis gene. 48(PMC1682979), 1991/03/.
- [41] H. Abdel Rahman, A. Abdul Wahab, M. O. Abdel Rahman, and Ossama Abdel Rahman Mostafa. Faecal elastase-1 concentration in cystic fibrosis patients with cftr i1234v mutation. *Acta Pædiatrica*, 95(9):1066–1069, 2006.
- [42] D. Rapino, I. Sabirzhanova, M. Lopes-Pacheco, R. Grover, W.B. Guggino, and L. Cebotaru. Rescue of nbd2 mutants n1303k and s1235r of cftr by small-molecule correctors and transcomplementation. March 23, 2015.
- [43] K.E. Roberts, P.R. Cushing, P. Boisguerin, D.R. Madden, and B.R. Donald. 2012.
- [44] R. Samudrala and J. Moult. A graph-theoretic algorithm for comparative modeling of protein structure1. *Journal of Molecular Biology*, 279(1):287 – 302, 1998.
- [45] R. J. Trudeau. Introduction to graph theory. 1993.
- [46] S. Vishveshwara, K. V. Brinda, and N. Kannan. Protein structure: Insights from graph theory. *Journal of Theoretical and Computational Chemistry*, 01(01):187–211, 2002.
- [47] I. Wagner and H. Musso. New naturally occurring amino acids. *Angew. Chem. Int. Ed. Engl*, 22:816—828, 1983.

- [48] M. Watford. Glutamine metabolism and function in relation to proline synthesis and the safety of glutamine and proline supplementation. *The Journal of Nutrition*, 138(10):2003S–2007S, 2008.

VITA

SAMUEL KAKRABA

- Education: B.Ed. (Mathematics), University of Cape Coast,
Cape Coast, Ghana, 2007-2011
M.S. (Mathematical Sciences),
East Tennessee State University, Johnson City,
Tennessee, USA, 2013-2015
- Awards: Faculty Award: Outstanding Graduate Student,
2014-2015 Department of Mathematics & Statistics,
ETSU
Graduate Assistantship, ETSU, 2013-2015
- Teaching Experience: Spring 2014, Jan.-May.,
one section of Probability & Statistics,
Department of Mathematics, ETSU
Summer 2014, June-Aug.,
One section of Probability & Statistics,
Department of Mathematics, ETSU
Fall 2014, Aug.-Dec.,
Two sections of Probability & Statistics,
Department of Mathematics, ETSU
Spring 2015, Jan.-May.,
Two sections of Probability & Statistics,
Department of Mathematics, ETSU
2010-2013 Wesley Girls' High School,
Department of Mathematics, Cape Coast, Ghana,
Oct. 2006 - June 2007
Montessori Primary School, Cape Coast, Ghana
Sept. 2004 - Sep. 2006
Cherish International School, Cape Coast, Ghana

Affiliations: American Mathematics Society, 2014-2015
Midsouth Computational Biology & Bioinformatics Society (MCBIOS), 2015
Mathematics and Statistics Club, ETSU 2013-2015
Abstract Algebra club, ETSU 2013-2015

Conferences: KME Tennessee Beta Chapter, 2015
MCBIOS 2015 Conference, Little Rock, Arkansas, “A Graph-theoretic Model for Nucleotide Binding Domain 2 of Cystic Fibrosis Transmembrane Conductance Regulator” .
2015 Appalachian Research Forum at ETSU, 04/08/2015, “Associations of Alcohol Consumption and Skin Allergy with Non-Melanoma Skin Cancer: Findings from the 2012 National Health Interview Survey.”
2015 Appalachian Research Forum at ETSU, 04/09/2015, “A Graph-theoretic Model for Nucleotide Binding Domain 2 of Cystic Fibrosis Transmembrane Conductance Regulator” .

Pending Publications: “A Graph-theoretic Model for Nucleotide Binding Domain 2 of Cystic Fibrosis Transmembrane Conductance Regulator” .
by Samuel Kakraba and Debra Knisley.
“Associations of alcohol consumption and skin allergy with non-melanoma skin cancer: findings from the 2012 National Health Interview Survey” .
by Samuel Kakraba and Kensheng Wang.

Research Interest: Computational Biology, Bioinformatics, Epidemiology, Biostatistics, Medical Research, Statistical Genetics, Multivariate Statistical Analysis, Mathematics Education