UNIVERSITY OF CAPE COAST

A GRAPH-THEORETIC MODEL OF HAEMOGLOBIN DOMAIN

EDEM KOFI NETSEY

2020

UNIVERSITY OF CAPE COAST

A GRAPH-THEORETIC MODEL OF HAEMOGLOBIN DOMAIN

BY

EDEM KOFI NETSEY

Thesis submitted to the Department of Mathematics of the School of Physical Sciences, College of Agriculture and Natural Sciences, University of Cape Coast, in partial fulfilment of the requirements for the award of Masters of Philosophy degree in Mathematics.

JULY 2020

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature: .... Date: ....

Name: Edem Kofi Netsey

Supervisor's Declaration

I hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor's Signature: .... Date: ....

Name: Dr. Samuel Mindakifoe Naandam

ABSTRACT

In this work, we built a graph-theoretic model of the haemoglobin protein domain and subsequently examine the impact of single point mutations in the entire protein domain to ascertain how significantly the severe and mild mutations are from the wildtype mutation. We computed graph centralities measures and examined the impact of single point mutations in the β-chain of the haemoglobin by comparing the wildtype structure with six most common mutations in the domain. We visualized the impact of the virtual single point mutations by use of dendrogram. Our results showed that the severe mutation (E6V) is significantly different from the wild and mild mutations.

KEY WORDS

Change in Glutamic acid at position 6 to Valine (E6V)

Graph-theoretic modeling

Haemoglobin protein domain

Hierarchical Clustering

Sickle Cell Disease

Single point mutation

ACKNOWLEDGEMENTS

I am most grateful to the Almighty God for His gift of life in every circumstance. Secondly, I wish to extend my profound gratitude to my Supervisor and mentor Dr. Samuel M. Naandam for his continuous support and supervision in the course of writing this thesis. He really served as an inspiration to me. I also would love to extend my heart felt gratitude to Prof. Debrah Knisley and Dr. Samuel Kakraba whose contributions in diverse ways ensured a successful completion of this thesis. God richly bless you.

Additionally, I would love to say an endless thanks to my sisters Winifred Serwah, Justine Dziedzorm and Judith Makafui for their massive support and prayers. Finally, I am grateful to the entire Netsey family and my course-mates for their love and encouragement through this academic journey. God Bless you all.

DEDICATION

To my parent: Stephen Brown Netsey and Celestine Dedzidi Dede Ameku.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SCD	Sickle Cell Diseases
SSG	Sub-Sequence Graph
TLG	Top Level Graph
TLGMCM	Top Level Graph Molecular Centrality Measures

CHAPTER ONE

INTRODUCTION

Background to the Study

It is trivially viewed that nothing exist in isolation-the impression that everything in existence, either physically or chemically one way or the other are in association with other physical or chemical elements. In graph theory, links between different physical and chemical compositions (or nodes) are studied where properties such as immediacy, domination number, weights of distinct and compounding nodes among others are analysed to model any change in the natural existing states of molecules (Ni, Sugimoto, & Jiang, 2011). Like nodes in a graph that often estimates local effects within a graph network, edges on the other hand most likely estimates global effects of a graph. Proteins are made of network of non-covalently linked amino acid side chains (Steward, 2019). In a protein network, there are therefore carbon, nitrogen, hydrogen, as well as oxygen atoms present which interacts within its network or with other protein(s) base on their chemical composition and properties. Today, there are variable graph network analysis and visualization algorithms with distinctive properties for studying graph networks to access the impact of vital components.

Sickle Cell Disease is one of the most prevalent inherited disease causing a very high fatality rate especially among newborns every year. This disease is caused by a single point mutation in the amino acid sequence of the haemoglobin protein domain (Asare et al., 2018). The haemoglobin in the red blood cells is an oxygen transporter molecule in ensuring oxygen is transported to tissues and organs in the human body (Berg, Tymoczko, & Stryer, 2002). Only a handful of scholarly work exist on use of graph-theoretic approach in examining the effect of a point mutations on protein sequences. No literature exist on application of graph-theoretic modeling for examining the impact of single point mutations especially in the β-chain domain of the haemoglobin protein where sickle cell disease emanates.

We used two major network analysis tools in this work, namely Cytoscape and UCFS Chimera for visualysing and analysing the de-oxy hemoglobin protein. In a protein graph network, there are interaction between neighboring proteins to perform specific function within an organism. According to Brinda & Vishveshwara (2005) each amino acid in a protein structure is a node, and the strength of the non-covalent interactions between two amino acids is evaluated for edge determination.

Modeling in this thesis is done to determine the impact of likely changes in the states of a protein molecule due to a single point mutation of hemoglobin protein. According to Berg et al. (2002), hemoglobin (found in red blood cells) is a protein that functions as an oxygen transport molecule necessary for life processes in living organisms. A change in the amino acid sequence of the hemoglobin protein caused by a single point mutation in the amino acid sequence resulted in variety of hemoglobin disorder called Sickle Cell Diseases (SCD). In this research, we use graph-theoretic modeling to build a hierarchical graph for the human hemoglobin and used the nested graph to examine the impact of single point mutation that results in Sickle Cell Diseases.

In graph theory, we study and describe pairwise relationships between objects. Graph-theoretic modeling seeks to understand relationships in complex networks like system network biology, chemical interaction among others.

Chimera provided 3D coordinate data of amino acid residues as nodes for generating interactions at specific distance range in Cytoscape in order to create a network of the residues where interactions were based of adjacency and distance measure of at most 6 angstoms between residues and also between compounding residues or peptide. Sub-units of a specific sequence of the hemoglobin protein were compressed into a single unit to obtain a top level graph which forms the basis for our domain modeling (Kakraba & Knisley, 2016).

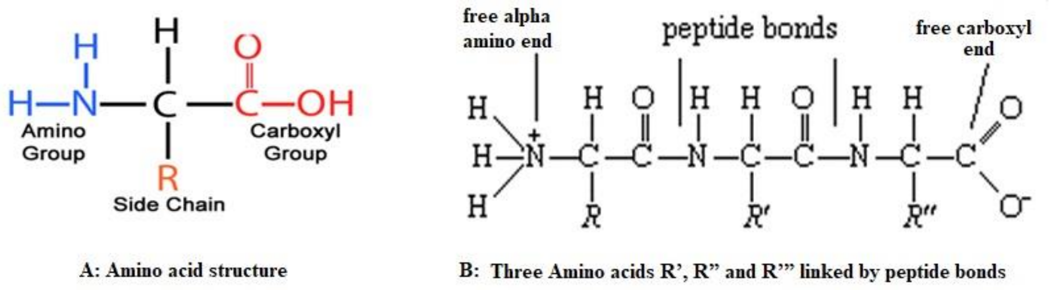
The definition of a graph in our context here differs completely from the functional way of defining a graph: we have a graph here as a set containing subsets called Vertex set (V) or nodes and edge set (E) or loops. The Vertexes or nodes are objects and the edges or loops interactions. Every day, different kinds of graphs are used for different purposes however, the purposes for employing graphs is the same across a wide variety of research disciplines; that is to establish relationships among different objects. We could decide to find the shortest distance moving from "town A"-to-"town B" (two-distinct places) which will require travelling through a number of other cities (nodes) from A through to B in order to either optimize profits or reduce stress by passing through a comfortable route. In that way we will be able to choose the best route base on the goal we intend to achieve.

Just as in graphs, proteins are complex molecules that exist in giant networks acting to build the structural elements of an organism thereby providing energy necessary for life processes. These complex proteins constitute highly sophisticated, fine-tuned molecular pumps that efficiently couple various sources of energy in the cell to transport a wide range of molecules across the membrane, often against their chemical gradient (Tajkhorshid, 2010).

Intentionally, graphing the hemoglobin protein domain will help us to identify the specific subdomain of the entire amino acid sequence where the single point mutation result in the SCD leading to severe health complications in some cases of the mutations.

Protein Structure

Amino acids are the building block of proteins (Steward, 2019). They carry the information necessary for protein synthesis. They can thus be considered as the 'blueprints' that contain the design of the living organism. According to Blanco & Blanco (2017) proteins are large size molecules (macro-molecules), polymers of structural units called amino acids. From organic Chemistry point of view, amino acids are organic compounds composing of amine ($-NH_2$) and Carboxyl (-COOH) functional groups along with a side chain (R group) that is specific for each amino acid; they differ from each other in their side chain R groups. We obtained the amino acid sequence of normal human hemoglobin which has the same sequence as 1A3N gene in the protein data bank (PDB); thus we generate our subdomain graphs with the 1A3N pdb data. The secondary structure of human hemoglobin revealed the existence helixes and coils which helped us in partitioning the full length of the normal hemoglobin of homo sapiens into subsequence. In partitioning the protein sequence according to Knisley et al. (2013) , each subsequence might contain one and only one type of secondary structure, either a beta strand, an alpha helix, or a loop. The loop regions may contain turns, a $\frac{3}{10}$-helix or an alpha helix with no more than 6 residues (Knisley et al., 2013).



A: Amino acid structure

B: Three Amino acids R', R" and R'" linked by peptide bonds

Figure 1: **Amino acid structure and 3 R-groups linked by peptide bonds**

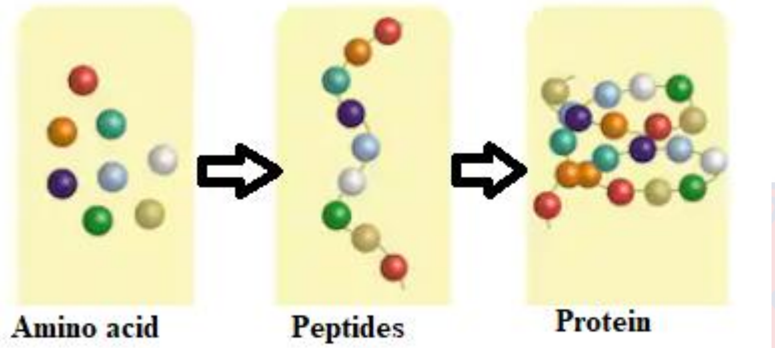


Figure 2: **Amino acids; Peptide; Protein**

Protein Structure Related to Functions

Several function of proteins exist. Proteins may function in various ways in transporting, storing other molecules like oxygen, providing mechanical support, movement, nerve impulse transmission, growth control and cell differentiation

(Berg et al., 2002). Different proteins perform specific functions based on how they fold in their secondary structure to prevent infections.

Illustratively, protein folding and maintenance is critical to protein functions: Misfolded uncorrected protein are seen in elevated levels in protein aggregates in Alzheimer's Parkinson's diseases among other new generative diseases (Kakraba et al., 2019). Distinct protein thus perform specific functions which in summary are characterized by how the protein folds in the secondary structure form. As expected, any defect in the functions of a protein will have some consequences in the organism. Although such defects are usually sufficient to cause harm or death in some cases to the organism, inadequate function of membrane transporters also constitutes the molecular basis of a wide range of serious human diseases, such as cystic fibrosis, familial intrahepatic cholestasis.

Amino acids in the primary structure are arranged sequentially in the polypeptide chain held together by peptide bonds during the process of biosynthesis. Here, the structure starts from the amino-terminal (N) to the carboxyl terminal (C). The primary structure of each protein is unique, due to both the different order and arrangement of the amino acids in the polypeptide and the total number of amino acids constituting the protein molecule (Kakraba, 2015). In the secondary structure, there exists highly regular local sub-structures on the actual polypeptide backbone chain defined by patterns of hydrogen bonds between the main-chain peptide groups. The α-helix and the β-strand (or β-sheets) are the two main types of secondary structure resulting from the folding of the polypeptide chain, which shows the character of the secondary structure of a specific protein.

Sometimes, the protein does not fold to show the existence of either α-helix or β-strand at a given iterated level and so gives a random coil structure in the secondary structural level. For the tertiary structures, proteins are visualized in three-dimensional structural view of monomeric and multimeric protein molecules by complete folding of the sheets and helices into a compact globular structure. The tertiary structure is held in position by hydrophobic and hydrophilic interactions. See Figure 3 for clarification.

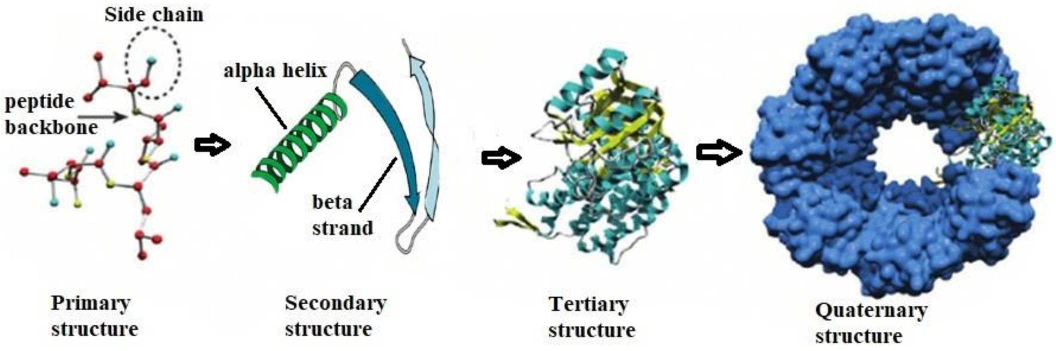


Figure 3: **Protein structural hierarchy**

Mutations: Causes and Effects

A single point mutation refers to a change of a single amino acid in the amino acid sequence of a protein resulting in the abnormal functioning of the protein. Single point mutations can result from addition, deletion, insertion of an amino acid.

Mutations might result in permanent damage to the genetic sequence or have a mild effect on the sequence (Kakraba, 2015). Kakraba and Knisley (2016) used a hierarchical graph modeling for examining the single point mutation effect on nucleotide binding domain 2 for cystic fibrosis transmembrane regulator. However, no literature exists on the use of nested graphs (in graph-theoretic modeling) in studying the effect of single point mutations on the hemoglobin domain, a motivation for this work. In this thesis, we present a mathematical model for haemoglobin of homo-sapiens, and use it to study how the domain of the red blood cell's defects causes sickle cell diseases as a result of mutation.

Sickle Cell Disease and Sickle Cell Anaemia

With the rising population in hemoglobin disorders among newborns globally, it is rarely possible for health practitioners today to have a perfect cure to satisfy the need of their patients. Sickle cell disease (SCD) is a hemoglobin disorder which occurs as a result of mutation in the DNA sequence of the hemoglobin protein. Cellular organisms use messenger RNA to convey genetic information where some RNA acts in catalyzing biological reactions, controlling gene expression, and communicating responses to cellular signals (Berg et al., 2002).

The process uses transfer RNA molecules to deliver amino acids to the ribosome, where ribosomal RNA then links amino acids together to form coded proteins. A rapid and timely transfer of genetic information and important elements is thus necessary in ensuring every cell and tissue gets their fair share of resources necessary in maintaining the growth and the normal functionality of the human body. The inability of the hemoglobin to function properly resulting in the sickle cell disease thus depends on how connected the subdomains of the amino residues are since a better connected domain will ensure a better transfer of functional and genetic information in the human body (Berg et al., 2002).

This mutation in the β-chain hemoglobin prevents the normal functioning of the hemoglobin as an oxygen transport molecule in the red blood cell. Several thousands of different proteins exist with each having its own particular amino acid sequence. The amino acid sequence of normal Haemoglobin “A” chain-B has 146 residues where a point mutation occurs each at a time resulting in the sickle cell disease with the most common and severe type in the β-chain (or chain-B) where Glutamic Acid (Glu or E) get substituted by a Valine (Val or V) at position 6 (E6V) producing Hemoglobin S in effect. The disease caused by this unique mutation as a result of replacement of glutamic acid with valine at position 6 (E6V) is called sickle cell anaemia.

In adults, normal hemoglobin structure (1A3N) has in total 4 chains. In the globin, mutation resulting in sickle cell disease occurs once the amino acid now available in the mutant protein no longer folds properly to enhance the functionality of the red blood cells. Thus, patients tend to have diverse health complications due to the abnormal protein resulting from the mutation. According Weatherall et al., (2006), the inherited hemoglobin disorders are characterized by an extremely diverse series of clinical syndromes of varying severity. Currently only a few people are able to afford a bone marrow transplant due to the treatment cost involved. Effective treatments however are available which can reduce symptoms and prolong life of patients. However, the diseases’ severity varies depending on the distinctive phenotype one has.

Hemoglobin Disorders Causing Sickle Cell Diseases

Hemoglobin is the pigment in the red blood cells that transfers oxygen to the tissues during human development (Weatherall et al., (2006). Hemoglobin disorders may be inherited in which case the structural hemoglobin variants and the thalassemias being the two groups of inherited hemoglobin disorders come from defective globin production following a recessive form of inheritance. By inference, if two carriers marry, there is a $\frac{1}{4}$ probability that any of their child will bear a defective genes from each parent, they are homozygous for the particular sickle cell disorder.

Among the several structural hemoglobin variants which may alter the function of the hemoglobin, only three (HbS, HbC, and HbE) are widespread. The homozygous state for the sickle cell gene results in sickle cell anemia, whereas the compound heterozygous state for the sickle cell and HbC genes results in HbSC disease (Weatherall et al., 2006). HbSC disease is milder but it also has important health implications.

Characterization of Protein Functions by Graph-Theoretic Modeling

A model is a scientific algorithm which is developed or under development purposely to study a problem for solution. Since proteins are made of amino acid residues linked together in a graph, there are interactions between these protein molecules based on their chemical composition and properties.

In modeling our binding domain, we partitioned the β-chain of the normal hemoglobin (particularly of adults) based on their secondary structures and determine their local effects by dominating sets (and thus dominating numbers) where possible. Unlike Kakraba & Knisley (2016) who used connectivity measures between residues at interacting distances of 8 angstom, we however considered a more interactive connectivity measure of at most 6 angstoms between residues as well as for hydrogen bonding. There is often likely interaction between neighbouring proteins to perform specific or related function within an organism. Graph-theoretic model thus gives enough understanding on protein structure and protein-protein interaction network. Other graph invariants for the purpose of our study were also computed and explained.

In this research, we used graph-theoretic modeling to build a hierarchical graph for the human hemoglobin binding domain to examine the global of virtual single point mutations causing sickle cell disease on the haemoglobin protein domain.

Statement of Problem

According to the World Health Organization (2006), 7% of the world's population is a carrier for haemoglobin disorders, and between 300,000 and 500,000 infants with the severe, heterozygous form of these diseases are born each year. In Ghana, approximately 15000 (2%) of Ghanaian newborns are diagnosed with SCD annually (Asare et al., 2018). The disease is caused by a single point mutation in the amino acid sequence of the haemoglobin protein. Kakraba & Knisley (2016) used graph-theoretic modeling to study the effect of single point mutations in the nucleotide binding domain 2 (NBD2) of the cystic fibrosis transmembrane conductance regulator. This follow-up work uses molecular descriptors from previous work by Kakraba & Knisley (2016) and analogous method to study the effect of the single point mutation in the haemoglobin domain. Like previous works, we build hierarchical graph for the haemoglobin protein domain and use graph invariants and graph centralities to examine the effect of virtual single point mutation on the protein domain.

Significance of the study

This research is significant in the sense that we obtain a deeper insight into how some popular phenotypes of the sickle cell diseases are ranked in order of severity by a dendrogram clustering of mutations in the β-chain haemoglobin protein domain. In this study, we build a nested graph for the haemoglobin protein domain and used it to study the virtual effects of single point mutations on the protein domain. This adds to the body of knowledge on application of graph-theoretic modeling to the study of diseases.

Research Objectives

This thesis seeks;

1. To develop a graph-theoretic model for haemoglobin protein domain that can adequately study sickle cell diseases.
2. To access the impact of single point mutations in the hemoglobin domain.
3. To determine the severity of distinctive phenotypes of the SCD mutation in the haemoglobin domain.

Delimitation

Different protein perform distinctive specific functions as their respective alignment of amino acid sequence in each protein varies. Knisley et al., (2013) built a nested graph and used it to study the effect of single point mutations in the nucleotide binding domain1 causing cystic fibrosis. Likewise, with improved molecular descriptors of 20 most essential amino acids in studying a single point mutation. Kakraba (2015), Kakraba & Knisley (2016) built the hierarchical graph for nucleotide binding domain2 and examined the effect of single point mutation in NBD2. In this research, we build a nested graph for normal haemoglobin protein (1A3N) and examine the consequence of single point mutations of the haemoglobin protein sequence in order to study the SCD adequately.

Definitions of Terms

Some of the terms and concepts used in this thesis are listed below.

Definition 1 (A graph)

A graph G(V, E) may be defined as a structure containing a set V of objects called vertices (or nodes) in which some pairs of the objects are ”related” by a set E called edges. Edges link objects in the structure of graph.

Definition 2 (A Clique)

A clique of a graph is a subset of vertices of an undirected graph such that every two distinct vertices in the clique are adjacent.

Definition 3 (Adjacency Matrix)

Any two vertexes are said to be adjacent if there exist an edge set linking the two such that the vertexes are said to interact. The matrix obtained for adjacent vertexes is called adjacency matrix.

Definition 4 (Protein Data Bank)

The Protein Data Bank is the main primary online repository or database for three-dimensional structural data of large biological micro-molecules such as proteins and nucleic acids which are determined by x-ray crystallography and nuclear magnetic resonance for the purposes of collecting of universal proteins, identification of protein families and domains, reconstruction of phylogenetic trees and for profiling of protein structures.

Definition 5 (Binding Domain)

A binding domain refers to a protein domain which binds to a specific atom or molecule. They are essential by helping to splice, assemble, and translate proteins for the function of many proteins.

Definition 6 (Mutation)

Mutation as used in this research refers to a change in the amino acid sequence of a protein. An instance is the replacement of glutamic acid at position 6 by valine which results in one of the severe type of Sickle Cell Disease.

Definition 7 (Nodes)

A node refers to the vertex where atoms or an object (protein) is located in a protein network.

Definition 8 (Amino Acid Sequence)

A sequential links or arrangement nucleic acids in the DNA of a specific protein.

Definition 9 (Peptides)

A network (or a link) of two or more amino acids lined by peptide bond(s).

Chapter Summary

We started the chapter by introducing the concept graphing in protein molecules, gave the background of the study and an introduction of some concepts of graph-theoretic modeling in Protein networks. We discussed the problem we intend to study and the significance of the studies. We also stated the objectives of our research and gave the definition of some terms. The chapter in all presents an entire structure of this thesis.

CHAPTER TWO

LITERATURE REVIEW

Introduction

In this chapter, we review some works of literature related to our study by first giving a background on graph-theoretic modeling with some major and relevant applications in the field of protein science as well as other diverse research disciplines with their applications in the manufacturing sectors. We also elaborate on the study of protein molecule networks by applying some important graph invariants in establishing major interactions within and between these molecules which formed the building blocks of living organisms.

Modeling of Proteins by Graph-Theoretic Approach

Graph-theoretic modeling seeks to develop an appropriate algorithm that will best define a pairwise relationship between objects based on their physical or chemical properties in a network.

The terms graph theory and network theory are often used interchangeably and the difference is perhaps more one of emphasis, with network theory describing the application of mathematical methods to real world systems, rather than the study of networks or graphs for their own sake (Hodges, 2019). In graph-theoretic modeling, we implement the knowledge obtained in graph theory, a sub field of discrete mathematics. Discrete mathematics is the branch of mathematics that deals with objects that are distinct and are often characterized by integers rather than continuous varying variables. Graph theory studies and describes pairwise relationships between these objects.

The history of graph theory emerges in the year 1736 when Leonhard Euler attempt to solve a problem relating to walks across all of the seven bridges by citizens of Konigsberg¨ crossing the islands only once, without ever repeating a single bridge as a person walk across all seven bridges exactly once during his walk. Konigsberg¨ at the time in Germany was a city built around a river with two large islands in the middle of the Pregel River, each connected to one another by seven bridges. Euler formulated mathematical approach in attempting to solve the problem.

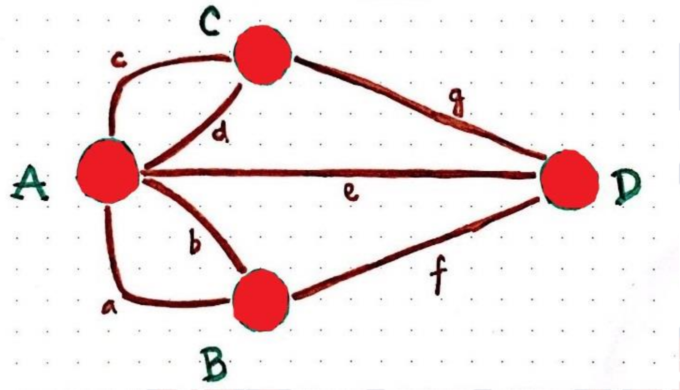


Figure 4: **Euler's Konigsberg's Problem**

Euler represented the landmarks as shown in Figure 4 with letters A, B, C, D (nodes) and decided to track a bridge "crossing" by the landmarks that one started at and ended at. To cross from A to B, the trip on the bridge (the edge here) would be referred to as AB. In crossing from position B to position D, the whole walk is seen as ABD. Hodges (2019) identified that using the central abstraction of networks, Euler did solve a real world problem by representing the system as a set of nodes (or vertices) that are joined together by a collection of edges (or links). Euler based his approach on adjacency of the nodes and the use of the central abstraction of networks by which he solves the problem by proving the problem of visiting all islands (the nodes) whilst crossing each bridge (the edges) only once had no solution. That is; in graph theory, two vertexes are adjacent if they have a common edge (that connects them). Also, one could classify vertices by their degrees. The degree of a vertex is the total number of edges that are adjacent to that vertex. It is not possible to solve the bridge problem if there are four vertices with an odd degree (Hodges, 2019). Euler, in his proof expected not more than 2 degrees for odd vertexes which was impossible.

Evidently, data structures within a graph according to Ni et al., (2011) relates object to each other where the interactions of these objects (or nodes) are very important as we attempt to get to another nodes reference to a formal position. In that case, there comes a need to consider the idea of traversal of a graph with usually different algorithms available for traversal, regarding the type of graph in question (whether directed or undirected).

Protein molecule(s) Interaction Networks

Proteins generally work as groups through a complex array of interactions performing a single biological function (Forero, 2017). It was shown that the complex interactions between different proteins were best viewed through graph theoretic modeling theory, in which a residue interaction graph is modelled regarding each residue as a vertex and each corresponding interactions as edges. Of course, amino acids are linked in sequential orders to form these functional proteins and thus weighted edges can be used in drawing interaction networks among them. The case is possible among proteins as well to be able to function and help in transports and metabolism in cells and tissues of an organism. Subgraphs or subdomains of a protein sequence have been thought to be the building blocks of networks which show important patterns in gene regulatory networks. We could therefore partition our sequences into sub sequences in order to obtain a more connected subgraphs base on a better interactive proximal distance between distinctive molecules. Subgraphs thus provide evidence that suggests there may be evolutionarily conserved characteristics across the protein-protein interaction (PPI) networks of different organisms just as in DNA traces. Thus for better understanding of the structure of protein-protein interaction networks, Forero (2017) described how the local structure of these networks was accounted for by the occurrence of small connected subgraphs, which he created. He tackled this problem of complexity of interactions in PPI network by proposing a method that is invariant to translations and rescaling of subgraph count distributions, and which detects similarities across networks with the different number of nodes or edges.

Further in modeling protein-protein interaction networks, a graph may be either directed or undirected. In a directed graph, we often differentiate between the in-degree and the out-degree, which refer to the total weights of edges into and out of a node. Hodges, (2019), further in his work, explained how a network of N nodes may be represented mathematically by an N × N adjacency matrix, A, where the entry A(ij) is equal to the weight of the edge between nodes i and j in a network containing N nodes. A symmetric adjacency matrix is one that is obtained if the network is undirected one. As the degrees of each of the nodes was compiled into an N × 1 vector d, a diagonal matrix of node degrees was defined: D = diag(A) by which the Laplacian matrix: L = D – A representation of the network was defined.

Just as graph topologies change with network structures appearing in their connected domains, it is necessary to consider graph isomorphism in protein interaction networks. Does the orientation of the haemoglobin affect its normal functioning? What about the amino acids in the sequence of this protein. Answers to this question are believed in the research context to provide a solution to a number of health complications associated with haemoglobin disorders as well as with the sickle cell disease. According to Jarman, (2017), social networks to computer networks, protein and transport networks, and neuronal networks of the mammalian brain, many of these networks share common structural properties.

Proteins, according to Blanco & Blanco (2017), are polymers of structural units called amino acids. Thus proteins have their foundation from amino acids composing of amine ($-NH_2$) and Carboxyl ($-COOH$) functional groups along with a side chain (R group) that is specific for each amino acid. They have a carboxyl group and amino group bonded covalently to α-carbon atom first. This peptide bonds linking the various amino groups can be weighted and their weights, when investigated based on physical and chemical compositions, tells us what elements are more pairwise-related in a connected domain. We also noted that in the secondary structure of the haemoglobin, there exist highly regular local sub structures on the actual polypeptide backbone chain defined by patterns of hydrogen bonds between the main chain peptide groups. The α-helix shows the characteristics of this specific protein with the tertiary structures helping us visualize in three-dimensional structural view of the haemoglobin as either monomeric and multimeric protein molecule.

Protein Binding Domain

Research shows that most mutations in DNA sequences do take place in the binding domains of a protein. DNA-binding domains are often part of a larger protein consisting of further protein domains with differing function. The majority of these mutations occur in the conserved central portion of the gene, but there has been little information about the function of this region (Pavletich, Chambers, & Pabo, 1993).

Mutations in CFTR are located in the nucleotide binding domain 1 and 2 (Knisley et al., 2013). Knisley et al. (2013), Kakraba (2015), Kakraba & Knisley, 2016) modelled NBD1 and NBD2 of the cystic fibrosis transmembrane regulator using a nested graph model. Unlike Knisley et al. (2013), Kakraba (2015), Kakraba & Knisley, 2016) in their work used atomic numbers in weighing the node/vertexes of amino acid molecule in the graph network which best describes the entire protein for that matter. According to Knisley et al. (2013), given an amino acid, the backbone and central carbon atom are represented by a single vertex and each of the atoms in the corresponding amino acid residue structure is represented by a vertex which was weighted by the mass of the corresponding atom. Edges however depicted molecular bonds with molecular bonding to hydrogen atoms being ignored. Kakraba (2015) in his thesis report did include hydrogen atoms to his molecular weight estimation which gave a true reflection of the various amino acid residues in the protein they worked with. Using the hydrogen suppressed models however, Knisley et al. (2013), obtained twenty corresponding vectors of descriptors based on the graph-theoretic measures of the twenty most common amino acids which were weighted domination, weighted diameter, circumference and weighted periphery. Measure of polarity and hydrophobicity were also used to compute graph descriptors to create a cluster in order to determine the impact of a single point mutation in the domain.

Later in a follow up work, Kakraba & Knisley (2016) did partition the sequence of CFTR corresponding to NBD2 domain into a number of sub-sequences on the basis of existence of alpha helixes or beta sheets or loop based in the secondary structure of the CFTR protein. They ensured they did not cut through a binding-site which is a necessary condition for drawing any protein domain. They obtain a number of sub-sequences Si: S1, S2, S3,...,Sn; where n is total number of subsequences.

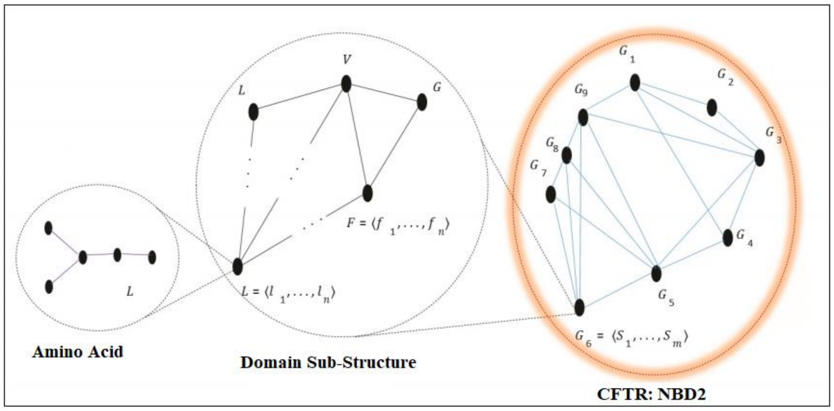


Figure 5: **Hierarchical graph for NBD2 (Kakraba & Knisley, 2016)**

Stability of Protein Molecule(s) Interaction Networks

It is seen that there exist some general underlying mechanisms for the emergence of certain complex network structures within and about a protein of which mutual relationship between structure and function in self-organising networks remained one major shared principle. To understand their emergence, Jarman (2017) decomposed the problem into two simpler problems and found their solutions by identifying; how a structure does effect dynamics and; also how the dynamics do shape the structure. The directed chain and the directed cycle connectivity configurations were considered in Jarman’s work which was distinguished by a single edge with stability analysis revealing radical changes in the patterns of dynamics. According to Jarman (2017) stability successfully reduces the highly complex problem of complex network emergence to patterns of connectivity through the understanding of the self-organisation of complex network structures across a broad range of contexts. As seen, Forero (2017) and Hodges (2019) were concerned with the application of graph theory in their research work (network modeling) whereas Jarman was into the stability of complex network systems. No research, however, took into accounts the various or possible subdomains of a protein. Knisley et al. (2013), Kakraba (2015), Kakraba & Knisley (2016) considered drawing out the subdomains of nucleotide-binding domain 1 and 2 to examine the effect of single point mutation in cystic fibrosis transmembrane conductance regulator.

Applications of Graph-Theoretic Modeling

Graph theory over the years have reasonably contributed to diverse research disciplines through knowledge advancement and the applications of graph-theoretic modeling to fields such as engineering, physical, biological and material sciences. One major area of the applications of graph-theoretic modeling systems is protein networks where there is always a number of pairwise relationships between molecules as well as atoms (Ni et al., 2011). Graph-theoretic modeling approach has been successfully used to study the dynamics of protein network domains. Another important area that we can apply graph theory is the area of drug design. The protein function is very essential in the discovery and design of drugs and since nested graph can be used to understand the protein function, it can thus be used as a powerful tool in drug discovery. Graph-theoretic modeling is also applied to solve issues relating to crime, scheduling of flight, transportation, etc. Figure 6 below showed a guilt by association graph of a company's email a few weeks before the company went bankrupt in which case vital information relating to sales was leaked by some members of the company which led to other people selling their stocks in time before the market went down (Jensen, 2020). The vertices in this network are people, of which some were not mentioned; the edges were emails sent across.

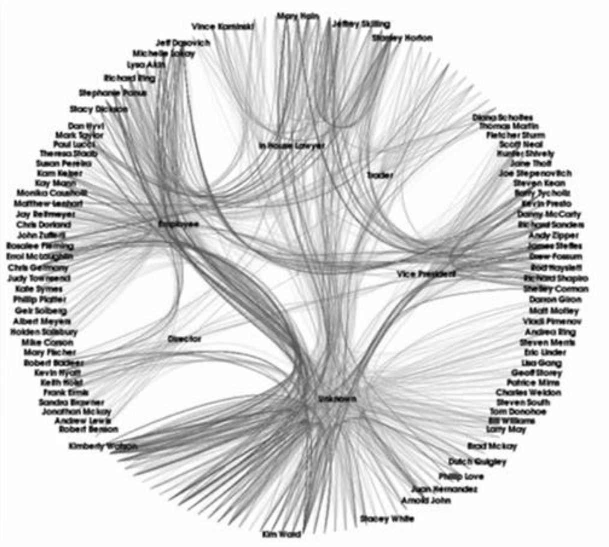


Figure 6: **A Screenshot of a guilt by association graph (Jensen, 2020)**

Also in the field of transportation, graph-theoretic model help to optimise profits in some cases and as well helps to reduce a number dangerous risks in experimentations. Mail delivery systems use graph theory to find the optimal node, thereby saving fuel and resources.

Chapter Summary

In this chapter we did discussed the background of graph theory and its modeling approaches with some related literatures been reviewed. We looked at literature on the main source of knowledge on graph theory. We went ahead to review literature on some protein-protein interaction networks methods used in the study of proteins and provided some applications in DNA sequencing and drug design. Graph-theoretic modeling simply involves application of graph theory to model/study real life problems. Although the study of graph theory have been in existence for long time, its applications has recently found space in several research and application fields.

CHAPTER THREE

RESEARCH METHODS

Graph Invariants as Measure of Molecular Properties

The method applied in this work is analogous to that used by Kakraba (2015), Kakraba & Knisley (2016) in examining the impact of a single point mutation in the cystic fibrosis transmembrane conductance regulator. We used some of the molecular database for the 20 most essential amino acids computed by Kakraba & Knisley (2016) and applied similar graph-theoretic modeling approach to study the effect of single point mutations on the haemoglobin domain of the sickle cell disease. We also used absolute difference in molar masses for node descriptors which we transcended to the edges as edge weights based on molar masses of compounding residues generated for each subdomain at the top level graph. We also used the weights in our weighted graph to generate further new descriptors for the top level graph each assigned to specific subdomain.

We first compared the Wildtype Haemoglobin and known mutation phenotypes in obtaining the molecular sequences and structures of the normal human haemoglobin "A" (with gene name 1A3N) in order to locate the position of the mutation in the structure. We used the 1A3N pdb protein id for our study. We obtained the full sequence length of 1A3N and used the secondary structure to guide our partitioning into sub-sequence as shown in Figure 7.

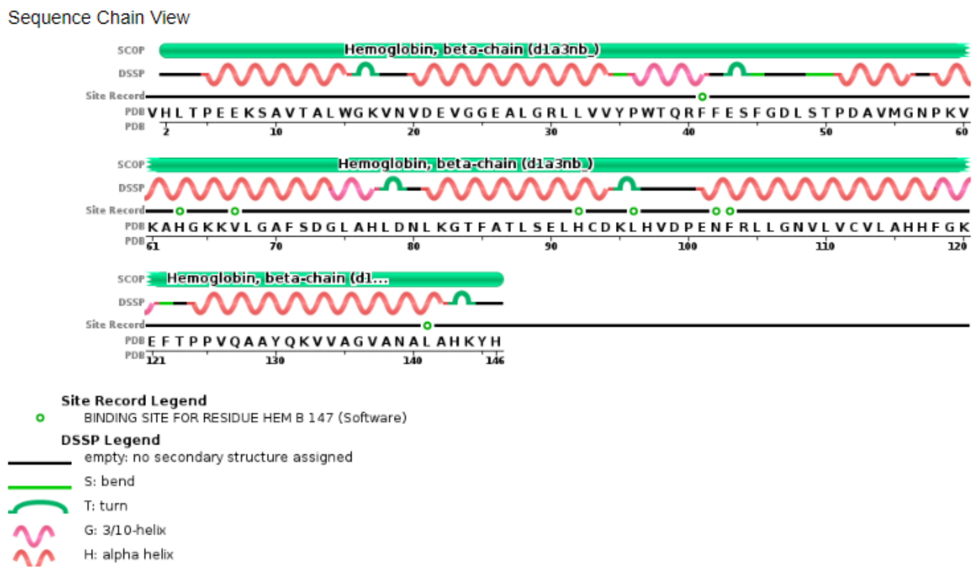

Sequence Chain View
SCOP
Hemoglobin, beta-chain (d1a3nb_)
DSSP
Site Record
PDB VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
PDB 2 10 20 30 40 50 60
SCOP
Hemoglobin, beta-chain (d1a3nb_)
DSSP
Site Record
PDB KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
PDB 61 70 80 90 100 110 120
SCOP
Hemoglobin, beta-chain (d1...
DSSP
Site Record
PDB EFTPPVQAAYQKVVAGVANALAHKYH
PDB 121 130 140 146
Site Record Legend
BINDING SITE FOR RESIDUE HEM B 147 (Software)
DSSP Legend
empty: no secondary structure assigned
S: bend
T: turn
G: 3/10-helix
H: alpha helix

Figure 7: **Sequence view of 1A3N in pdb (Tame & Vallone, 1998)**

Structural Visualization

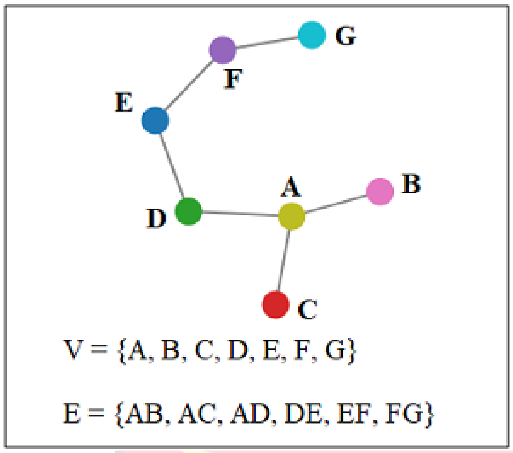
UCSF Chimera program is a protein molecular visualization and modeling tool (Pettersen et al., 2004). Chimera program (Pettersen et al., 2004), as launched in cytoscape, also aided in visualizing our protein's molecular structure in which we would distinguish our binding domain into subsequence for analysis and later generate our residue network in cytoscape application. According to Shannon et al. (2003), cytoscape is an open source software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles, among and others. Using cytoscape program, we created our sub-domain graphs corresponding to the subsequences.

Interactions and Residue Network Creation

A graph is a set G = (V, E), where V = {1, 2, 3, …, n} is its node or vertex set (a non-empty countable set of elements) and $E \subset V \times V$ is its edge set. A set of nodes interacts by edge set if and only if they fall within a distance of 6Å. Degree of a vertex is the number of edges falling on it. It tells us how many other vertices are adjacent to that vertex. Weighted degree of a vertex then is the weight we assigned to an edge falling on the vertex. All weighted measures were determined at a maximum diameter of 6A° between any two vertexes. Descriptors considered were order of graph, the degree of the graph, eccentricity, domination number. We also repeated the procedure by considering the weights of vertexes rather than edges to computer molecular descriptors of each subdomain graph.

Weighted Graph

Basically, a weighted graph is a graph that has weights assigned to its edges or nodes. We first of all considered molecular weights of the various amino acid residues especially or vertexes in the case of dominated sets within the network in which we determined both un-weighted and weighted domination numbers. Weighted degrees as well were computed from subdomain graphs as well as eigenvectors whose results or descriptors based on molecular weights formed a vital basis in estimating the impart of a single point mutation in the wildtype hemoglobin domain of homo sapiens (humans).



V = {A, B, C, D, E, F, G}

E = {AB, AC, AD, DE, EF, FG}

Figure 8: **A Sample Graph, G = (E, V)**

Consider Graph G in Figure 8 above. The idea of dominating set can be demonstrated supposing the nodes of the graph represents seven cities which needed a policing service within their township. The question we then ask ourselves is that- “in which of these towns can we build a police station in order to minimize the wastage of resources such that each community gets an immediate access to policing services?” Obviously, we have town “A” and town “F”; thus the dominating set graph G = (V, E) is given as a subset U of V such that every vertex not in U is adjacent to at least one member of U.

e.i. U = {A, F}; The domination number $\gamma(G)$ is also given as the number of vertex(s) in U, thus; $\gamma(G) = 2$

The weighted (G) is also given as the degree of each dominating vertex, thus;

Total weighted (G) = deg(A) + deg(F) = 3 + 2 = 5

Adjacency Matrix

For a graph G = (V, E); where V = {1, 2, … , n} . We considered an interaction within a maximum of 6Å between any two vertexes for residues to interact. Functionally, two equivalent graphs are isomorphic, but the converse is not true. Thus the adjacency matrix has a binary matrix which have entries zero(s) and ones in which case each diagonal entry is zero (no self-loop).

For an N number of nodes, there is an "N × N" adjacency base on closeness measure within a 6Å radius distance from any given vertex given by;

$A_{(\mathrm{NxN})} = [\mathbf{x}_{ij}]$, such that the entries;

$$\boldsymbol{A_{ij}} = \begin{cases} \mathbf{1} & ;\boldsymbol{ij} \in \boldsymbol{E} \\ \mathbf{0} & ;\boldsymbol{ij} \notin \boldsymbol{E} \end{cases} \qquad (1)$$

Suppose the graph G = (V, E) in figure 8 above is a residue interaction network; where V = {A, B, C, D, E F, G} is the vertex set consisting of 7 residues at defined positions along respective residue sequence of our haemoglobin protein with edge set E = {AB, AC, AD, DE, EF, FG}.

The adjacency matrix is thus given as;

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Centrality Measures

1. Degree Centrality:

The degree centrality $C_{(Di)}$ of a vertex $V_i \subset V$ in a graph G(V, E) is defined as the number of vertex in V adjacent vertex V_i. This is given below in equation (2):

$$\mathbf{C_{(Di)}} = \sum_j \boldsymbol{A_{ij}} \qquad (2)$$

2. Closeness Centrality $\mathbf{C(c_u)}$:

This measures the shortest paths from vertex *i* to all other vertexes V_j in the graph.

$$\mathbf{C(c_u)} = \frac{\mathbf{n-1}}{\sum_v \boldsymbol{d(u,v)}} \qquad (3)$$

where $u \neq v$; and d(u; v) is the shortest path between node u and any v in the graph. "n" is the number of vertex.

3. Eigenvector Centrality $\mathbf{x_i}$:

The eigenvector score x_i shows how influential or connected a vertex is relative to other vertexes in the graph. We used the eigenvector components corresponding to real largest eigenvalue obtained from the characteristic equation $\det(A - I) = 0$; of adjacency matrix A.

$$\mathbf{x}_i = \frac{\mathbf{1}}{\boldsymbol{\lambda}} \sum_j \mathbf{A}_{ij}\mathbf{x}_j \qquad (4)$$

where x_i is the relative centrality score of vertex i : $\boldsymbol{A_{ij}}$ is the $ij^{(th)}$ element of the adjacency matrix and λ is the greatest eigenvalue in A.

4. Betweenness Centrality $\mathbf{C_B(n_i)}$:

This is the ratio of all the shortest paths passing through a node and the total number of shortest paths in the network.

$$\mathbf{C_B(n_i)} = \sum_{j<k} \frac{g_{jk}(\mathbf{n_i})}{g_{jk}} \qquad (5)$$

g_{jk} is the number of geodesics(short paths) connecting jk and; $g_{jk}(n_i)$ is the number that node *i* is on.

In comparing isomorphic graphs which are symmetric (but the inverse is not), the greater the size-to-order ratio of one graph is to another, then we say the former has more interaction or connectivity than the later. This is often true and may often be considered in instances of determining the functionality of proteins since proteins with similar functionality often do interacts. Betweenness centrality is based on communication flow and vertexes with a high betweenness centrality score do lie on communication paths and can control information flow. Nodes in the graph that have many “shortest paths” going through them, analogous to major bridges and tunnels on a highway map (Yu, Kim, Sprecher, Trifonov, & Gerstein, 2007). Eigenvector centrality (EC) measures a vertex’s influence on a network with highly connected vertex in the graph having high EC scores. EC serves as a measure of the connectivity against a fixed scale when normalized, so it can be used to reliably compare different networks (Negre et al., 2018). In this work, some molecular descriptors assigned to the amino acids in the subdomain graphs were taken from previous work by Kakraba & Knisley (2016).

Nested Graph-Theoretic Model of Haemoglobin Protein Domain

I used cytoscape mainly for our network analysis. However, UCSF chimera (as launched via program launch pathway in cytoscape) is a program for the interactive visualization and analysis of molecular related data and structures. Chimera takes a modelled protein structure in PDB format as input and gives a 3D structural view of the peptides/ amino residues for our protein analysis analysis.

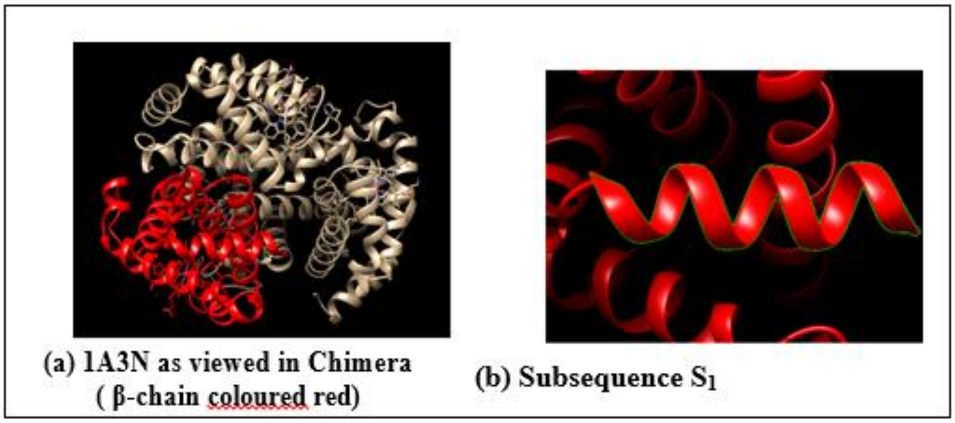
In chimera, we distinguished and selected our favourite sequence in the wildtype haemoglobin sequence chain B to generate a residue structure for analysis. Java program earlier installed in our browser system was a necessity to convert our python codes in chimera program which is in readable in cytoscape.

Structural visualization in Chimera also assisted in partition our binding-domain into subdomains.

I created a 3D interaction graph of our favourite domain sequence at interactive distance of 6 angstoms in Cytoscape using structureViz application also in Cytoscape. We determined weighted domination of the graph based on molecular descriptor such as the size and order of the graph were computationally obtained using a number of network algorithm in cytoscape based on graph theory literatures.

Sequence Partition

I obtained our sequence data from GeneBank as a primary sequence data and run the sequence in the protein data bank upon which we had a 100% corresponding protein structure which is de-oxy haemoglobin protein with identity 1A3N. We identified the chain "B" and upon a magnified visualization in chimera, we partitioned the sequence into sub-sequences on the basis of existence of α-helixes, β-strands and loops. We avoided cutting through the binding sites, alpha helix, beta strands as these contained important biological information needed to be preserved.


(a) 1A3N as viewed in Chimera (β-chain coloured red)

(b) Subsequence S_1

Figure 9: **1A3N Sequences Visualized in Chimera**

The Figure 9(a) shows a visualized result of the chain-B being distinguished by action colouring in Chimera. Figure 9(b) shows subsequence S_1 by demonstrating proceedings by method of spinning about axis for the sub-divisions of our β-binding domain of the human hemoglobin "A". We repeated the procedure for each of the sub-sequences. We also used the plain sequence view in Figure 7 to locate binding sites.

Table 1: Subsequence Partition of 1A3N

Sub-domain	Sequence	Reasons
S_1	1...17: VHLTPEEKSAVTALWGK	Coil, alpha helix, turn
S_2	18...35: VNVDEVGGEALGRLLVVY	Coil, Alpha helix, bend
S_3	36...48: PWTQRFFESFGDL	3/10 helix, coil, binding site
S_4	49...57: STPDAVMGN	Bend, alpha helix, coil
S_5	58. . . ..73: PKVKAHGKKVLGAFSD	Alpha helix,
S_6	74...79: GLAHLD	bending site
S_7	80...94: NLKGTFATLSELHCD	Alpha helix, binding site
S_8	95…117: KLHVDPENFRLLGNVLVCVLAHH	Turn, coil, alpha helix, site
S_9	118…123: FGKEFT	3/10 helix, bend
S_{10}	124…146: PPVQAAYQKVVAGVANALAHKYH	alpha helix, turn, coil

Source: Tame & Vallone (1998)

The full human haemoglobin “A” (β-chain) obtained from the GeneBank is:

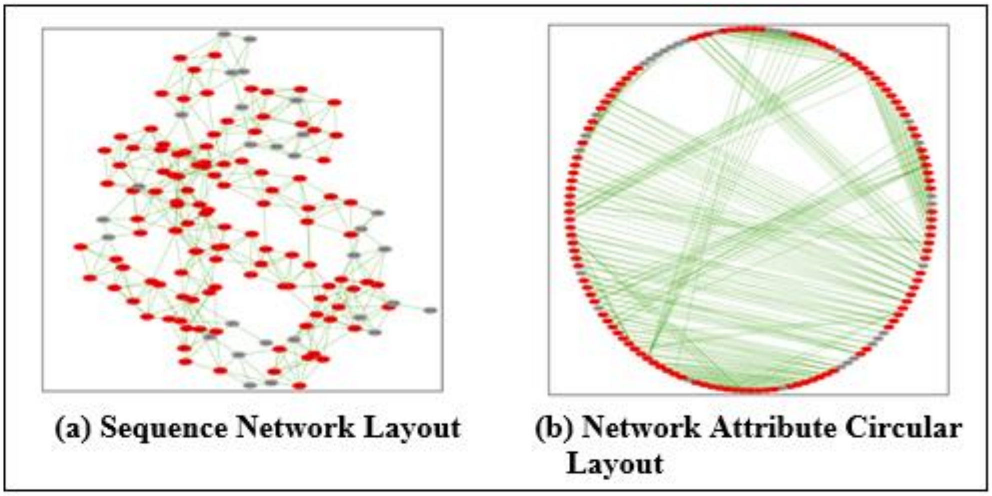
1-40: VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQR

41-80: FFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN

81-120: LKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK

121-146: EFTPPVQAAYQKVVAGVANALAHKYH

A residue interaction network of the entire sequence is presented in the Figure 10 below.

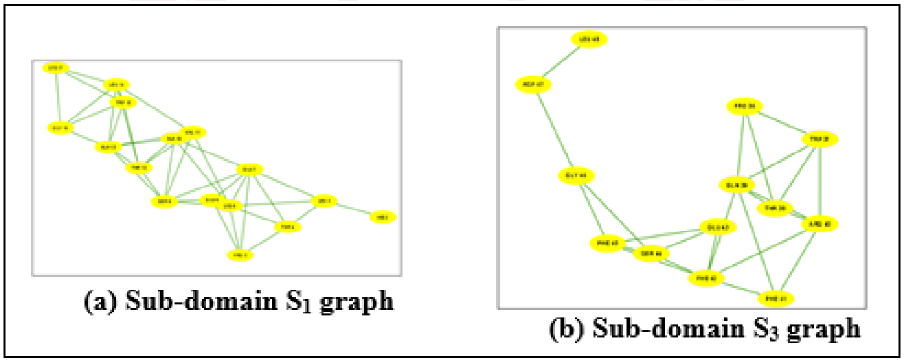


(a) Sequence Network Layout

(b) Network Attribute Circular Layout

Figure 10: **β-Chain Sequence Network Layout of 1A3N in Cytoscape**

Also, the graph in Figure 11 below demonstrates original orientations of two of our residues and created corresponding sub-graphs in Cytoscape. We magnified the orientations for each domain graph and visualized highly functional connected residues to determine a number of graph invariants (e.g. eccentricity and weighted domination) for the analysis of local effects within the residue network.



(a) Sub-domain S_1 graph

(b) Sub-domain S_3 graph

Figure 11: **Original orientation of two sample sub-graphs in Cytoscape**

Residue Interaction Graphs In Cytoscape

For the purpose of this thesis, regarding all the subsequence graphs, we excluded contacts, clashes, hydrogen bonds and backbone connectivity in interaction networks. We however maintained an interacting distance of 6Å between adjacent residues and included interacting distances between CA atoms. Each subdomain residue network was compressed into a single node or vertex and their respective centroid determined as interaction point to other compressed vertexes. We then used edge embedded spring layout to link each of the sub-graphs of the compressed β-chain of 1A3N protein structure to create a graph of 10 vertex based on whose interactions we computed descriptors for the top level graph.

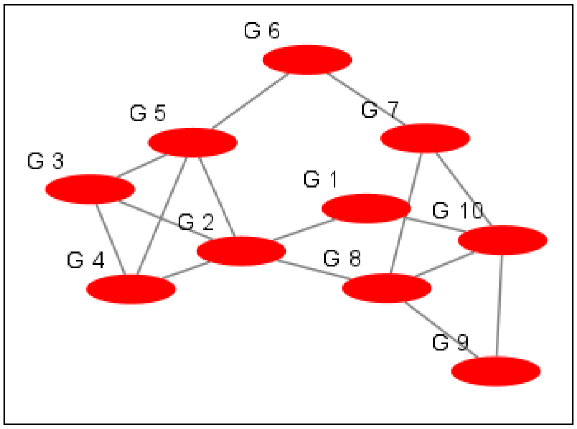


Figure 12: **Compressed β-Chain Haemoglobin Binding Domain**

I used the reaction forces along the edges to compute new molecular centrality descriptors using absolute difference in molar masses as weight along interacting edges (on assumption that acceleration of molecular substances between residues is constant or equal for any two interacting residues) in the top level graph as shown in Figure 12.

Weighted Edge Model

A weighted node within a graph gives a local measure/effect within a graph since it does less in considering the whole graph whereas we obtains a global measure/effect from edge weight.

A weighted undirected graph G = (V, E, w) is an undirected graph G = (V, E) with a function w: E∈R$^+$, where R$^+$ is set of positive Real numbers. The adjacency matrix of a weighted graph G will be denoted $A_{(N\times N)}$, and is given by;

$$A_{(N\times N)}w(i,j)=\begin{cases} w(i,j) & ;if \quad (i,j)\in E \\ 0 & ;if \quad (i,j)\notin E \end{cases}$$

The degree of a weighted graph G denoted by D$_i$, is the sum of the upper or lower diagonal entries of the adjacency matrix A such that:

$$D_{i=}\sum_j A_{ij} \text{ ... ... } (3.5)$$

Where A$_{ij}$ is the ijth element of A.

As adjacent residues within the haemoglobin interacts, there exist forces usually by contact or attraction by which they obtain energy in executing their functions as oxygen transport molecule as well as the disseminating necessary information for the maintenance of the biological system of humans.

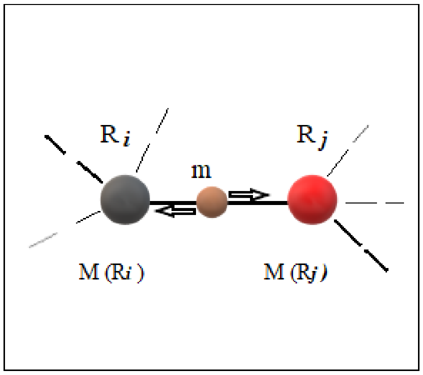


Figure 13: **Displaced Weight based on momentum along edge R$_i$–R$_j$**

Suppose there is equal acceleration for an interaction force between residue R$_i$ and R$_j$ in Figure 13, then the interacting force is proportional to the displaced mass along the edge which enhances the flow of information or functionality between this two residues. From the figure 13, M(R$_i$) and M(R$_j$) are respective molar masses of residue Ri and Rj.

The change in molar mass per average edge degree along the edge between R$_i$ and R$_j$ given by;

$$\Delta M = \left|\frac{M(R_i) - M(R_j)}{\bar{d}}\right| \times gmol^{-1} \qquad (6)$$

Where $\bar{d}$ is the average weighted degree (also defined as the average adjacent degree per node of the graph).

The unit of molar mass of residues is gram per mole ($gmol^{-1}$); thus weight between adjacent residues is the mass in kg per one mole of substance displaced along their edge is,

$$m = \left|\frac{M(R_i) - M(R_j)}{\bar{d}}\right| \times 10^{-3} kg \qquad (7)$$

This is the weight we assigned to any interacting edge and computed graph centralities based on edge weight to obtain the top level descriptors. Thus mass in kg per one amount of substance (in mole) displaced between any two adjacent residues is the weight which is equal to the absolute difference in molar mass per average weighted degree based on edge.

Nested Graph-Theoretic Model of Haemoglobin Protein Domain

The graph in Figure 14 below shows a hierarchy of our top level graph made of compounding residues being compressed into a single node. Basically, we created a graph of a healthy haemoglobin (full wildtype) protein sequence of the β-chain which we used as a basis to perform mutation at the top level. We achieved this by regrouping and compressing each subdomain into a single vertex and linked these graph at their centroid point to any adjacent grouped centroid node of another subdomain within the stipulated 6Å interacting range. The result yielded a cluster of two communities of interacting main compounding residues with a number of cliques shown in Figure 14. We then proceeded and show the nested graph of the β-chain binding domain in Figure 14 below which depicts the main composition of the domains at each graph domain level.

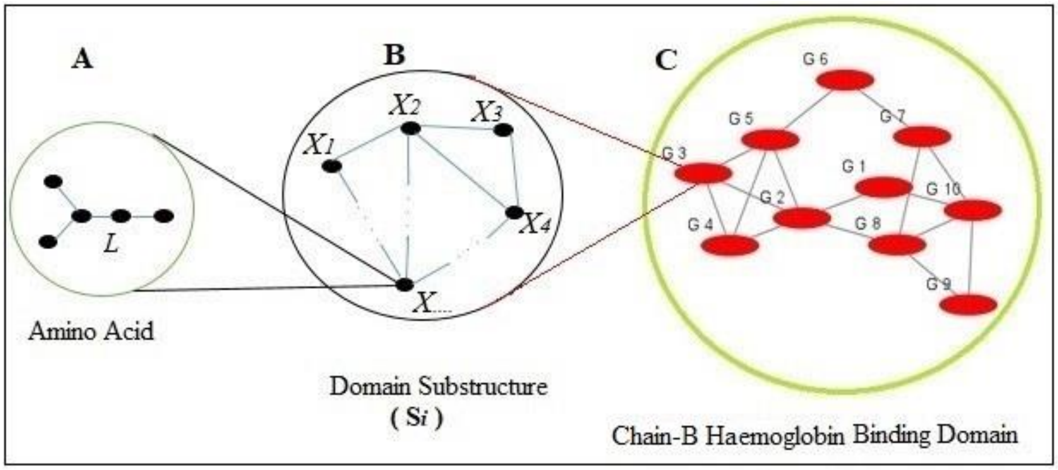


Figure 14: **A Hierarchical Graph of β-Chain Haemoglobin Binding Domain**

The top level domain graph C of is made of compressed subdomain residue network with each having a unique secondary structure. The subdomain graphs are made of different amino acids at specific location point along the β-chain sequence.

CHAPTER FOUR

RESULTS AND DISCUSSION

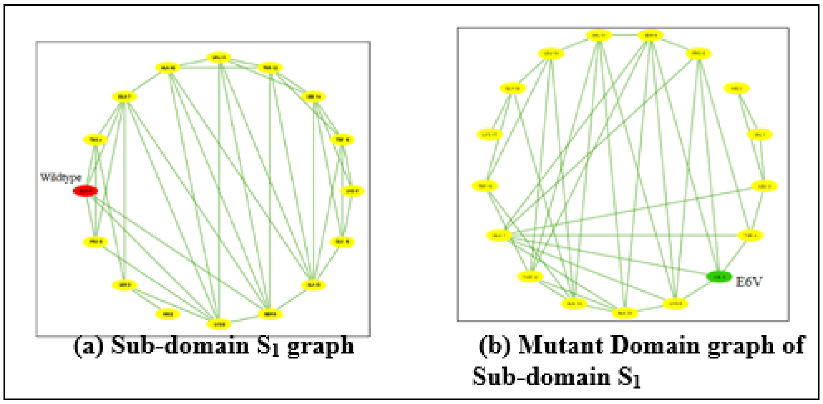
Visualizing Effect of Single Point Mutations on the Haemoglobin Domain

In this chapter, we discuss the entire results regarding the impact of a single point mutation in the normal haemoglobin protein chain “B” resulting in sickle cell disease (a defective haemoglobin sequence). Consequent to this mutation is the health-related implications in patients as the haemoglobin in the red blood cells fails to function effectively as oxygen transport molecule in the body due to a different molecular orientation of the new residue structure with it the new distinct molecular properties.

Some of the resulting mutation in this regard gives a severe health complications which often leads to death while others had mild health complications where carriers of this gene type live normal lives.

Impact of Displaced Mass on Mutation along Interacting Edges

We first computed descriptors for the various subdomains including those of mutant domains with respective specific positions along the protein sequence. We resubmitted our mutant sub-sequence of corresponding subdomain to I-Tasser to obtain edge interactions between adjacent nodes. We chose a predicted model with higher C-scores having a visual similarity with the non-mutated domain. C-score is an I-Tasser prediction score in the range [-5, 2]; the higher the score, the better the prediction. We re-computed our molecular descriptors. For each mutant domain we computed corresponding top level descriptors for both cases of un-weighted and weighted graphs.



(a) Sub-domain S_1 graph

(b) Mutant Domain graph of Sub-domain S_1

Figure 15: **Subdomain S_1 and its corresponding mutant domain E6V graph**

The displaced mass along each edge based on the molar masses of adjacent residues per 1 mole of substance was computed for all wildtype subdomains and the mutant domain for respective subsequence and their graph centralities computed using the top level graph and we present the results in Table 2 below. We then proceeded to cluster these results to determine the impact of a single point mutation in the -chain domain of the normal human haemoglobin “A”.

Table 2: Vertex Composition and Molecular Weights of Top Level Graph

Top level vertex compositions	Weight / $10^{-3}kg$
G1: VHLTPEEKSAVTALWGK	2154.3
G2: VNVDEVGGEALGRLLVVY	2208.3
G3: PWTQRFFESFGDL	1846
G4: STPDAVMGN	1035
G5: PKVKAHGKKVLGAFSD	1952.3
G6: GLAHLD	714.9
G7: NLKGTFATLSELHCD	1901.2
G8: KLHVDPENFRLLGNVLVCVLAHH	3020.5
G9: FGKEFT	817.9
G10: PPVQAAYQKVVAGVANALAHKYH	2829.2
E6V: VHLTPVEKSAVTALWGK	2124.3
E6K: VHLTPKEKSAVTALWGK	2153.4
V23I: VNVDEIGGEALGRLLVVY	2222.4
E26K: VNVDEVGGKALGRLLVVY	2207.4
K82N: NLNGTFATLSELHCD	1887.1
K95E: ELHVDPENFRLLGNVLVCVLAHH	3021.4

Source: Tame & Vallone (1998)

In Table 2, we computed new molecular descriptors based on molecular descriptors of residues in the compressed nodes/vertexes using the top level graph C. We located the mutant domains of the top level graph which is G1, G2, G7 and G8 and computed new displaced mass weights along the edges with and the results presented in Table 3.

For each mutant top level domain, we calculated their centralities and present the results in Table 4 all based on residue molecular weight.

Table 3: Edge Weights in kg per average degree of interaction in Top Level Graph

Interactions	Mutation Phenotypes						
	WildType	E6V	E6K	V23I	E26K	K82N	K95E
G3:G5	0.035433	0.035433	0.035433	0.035433	0.035433	0.035433	0.035433
G9:G10	0.670433	0.670433	0.670433	0.670433	0.670433	0.670433	0.670433
G8:G10	0.063767	0.063767	0.063767	0.063767	0.063767	0.063767	**0.064067**
G8:G9	0.7342	0.7342	0.7342	0.7342	0.7342	0.7342	**0.7345**
G1:G10	0.224967	**0.234967**	**0.225267**	0.224967	0.224967	0.224967	0.224967
G4:G5	0.305767	0.305767	0.305767	0.305767	0.305767	0.305767	0.305767
G3:G4	0.270333	0.270333	0.270333	0.270333	0.270333	0.270333	0.270333
G10:G7	0.309333	0.309333	0.309333	0.309333	0.309333	**0.314033**	0.309333
G7:G8	0.3731	0.3731	0.3731	0.3731	0.3731	**0.3778**	**0.3734**
G5:G6	0.412467	0.412467	0.412467	0.412467	0.412467	0.412467	0.412467
G6:G7	0.395433	0.395433	0.395433	0.395433	0.395433	**0.390733**	0.395433
G2:G5	0.085333	0.085333	0.085333	**0.090033**	**0.085033**	0.085333	0.085333
G2:G4	0.3911	0.3911	0.3911	**0.3958**	**0.3908**	0.3911	0.3911
G2:G3	0.120767	0.120767	0.120767	**0.125467**	**0.120467**	0.120767	0.120767
G8:**G2**	0.270733	0.270733	0.270733	**0.266033**	**0.271033**	0.270733	**0.271033**
G1:G2	0.018	**0.028**	**0.0183**	**0.0227**	**0.0177**	0.018	0.018

Source: Generated from Analysis (2020)

The coloured weight results in Table 3 represent change in weights as a result of a single point mutation for each specific domain.

Table 4: Weighted Molecular Centrality Measures of Mutation Top Level Graph

TLGMCM	WildType	E6V	E6K	V23I	E26K	K82N	K95E
WMaxDeg	1.4418	1.4418	1.4418	1.4371	1.4421	1.4465	1.443
WMeanDeg	0.936233	0.940233	0.93635	0.93905	0.93605	0.93717	0.93647
WMinDeg	0.2429667	0.262967	0.2436	0.247667	0.242667	0.242967	0.242967
WMaxEig	0.582758	0.582855	0.582742	0.583486	0.582763	0.581722	0.582701
WMeanEig	0.24784	0.24751	0.24785	0.24725	0.24782	0.24765	0.24783
WMinEig	0.036708	0.035591	0.036708	0.036166	0.0366554	0.036391	0.036696
WMaxClos	0.1917701	0.192546	0.1918	0.19122	0.191804	0.19193	0.191854
WMeanClos	0.15243	0.15311	0.15245	0.15217	0.15244	0.15252	0.15247
WMinClos	0.10002	0.10071	0.10001	0.09987	0.10002	0.10011	0.10003

Source: Generated from Analysis (2020)

Weighted molecular centrality measures as computed at the top level in Table 4 were based on centrality measures in equation 3.1, 3.2 and 3.3 for the top level graph C in figure14 with the weights in Table 3 being assigned to edges as weights.

WMaxDeg = Weighted maximum degree:

WMeanDeg = Weighted mean degree:

WMinDeg = Weighted minimum degree

WMaxEig = Weighted maximum eigenvector:

WMeanEig = Weighted mean eigenvector:

WMinEig = Weighted minimum eigenvector:

WMaxClos = Weighted maximum closeness:

WMeanClos = Weighted mean closeness:

WMinClos= Weighted mean closeness

Hierarchical Clustering and Impact of a Single Point Mutation in the β-Chain of Haemoglobin "A"

To visualize the impact of single point mutations in the β-chain haemoglobin protein domain, we used R statistical software to create a hierarchical cluster with the single-linkage function. The single linkage function was used for our hierarchical clustering because it gives a less bias estimate in the spread. The Figure 16 depicts the clustering of our haemoglobin protein domain.

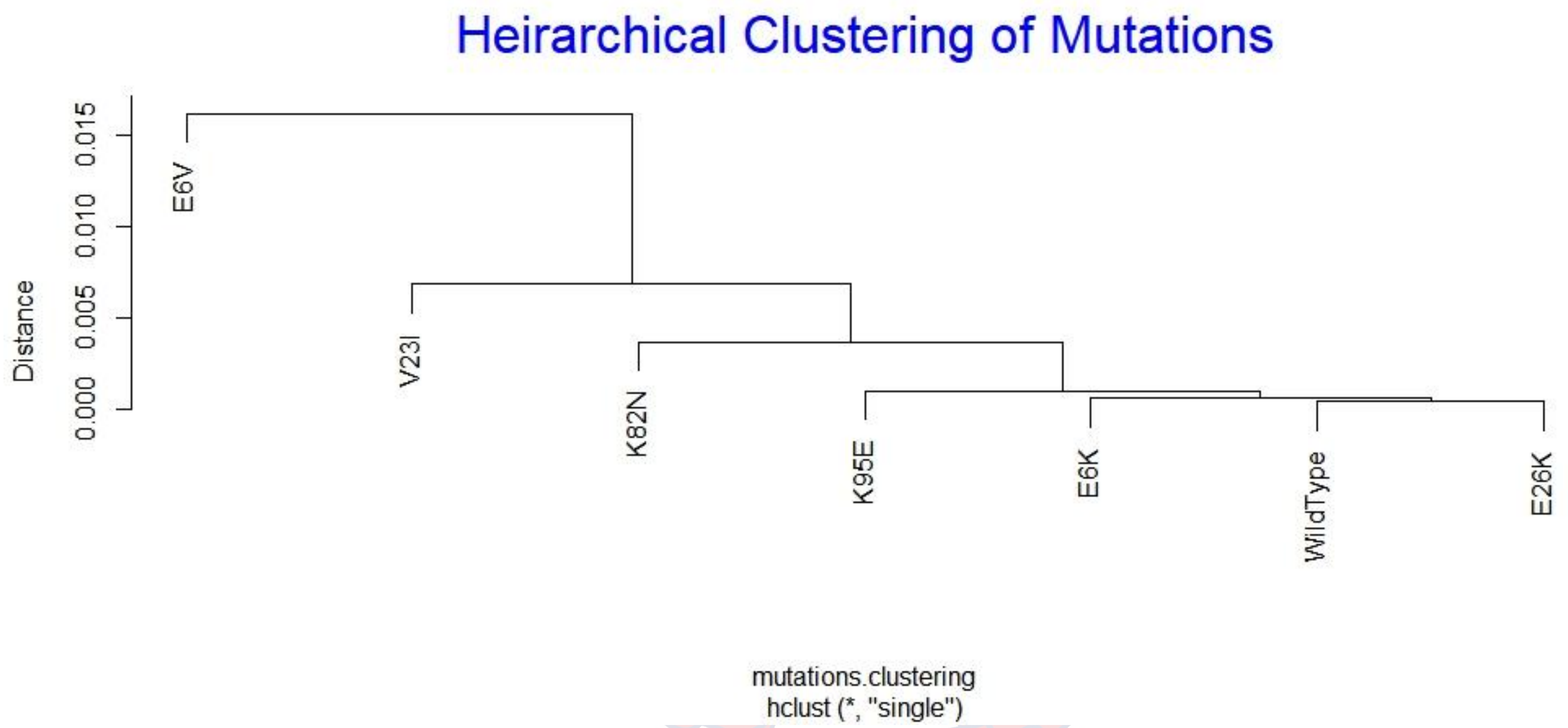


Figure 16: **Single point mutations in the haemoglobin protein**

The dendrogram clustering of our molecular descriptors results of Table 4 in Figure 16 evidently shows a maximum distinct spread from the Wildtype resulting from the replacement of glutamic acid at position 6 by valine (E6V).

As obvious, E6K, E26K and K95E mutations showed a much less significant effect in spread. Additionally, K82N and V23I are as well considerably distinct from wildtype, even though they all belong to one bigger cluster. The impact of the severe mutation type is at 0.015 Euclidean distance away from the wild type mutation.

A further research might question "why does the mutation E6V cluster entirely different in a different cluster?" Also, "under what graph-theoretic circumstance, based on our computation of molecular descriptors, can the spread between E6V and the wildtype becomes minimal as possible?"

A solution to this very question is clinically important as it will in effect produce a knowledge-base leading to the design of a molecule most likely to correct this specific mutation in the haemoglobin domain of humans.

Also, the severity of the mutations relative to the Wildtype are ranked in the order;

K82N<V23I<E6V

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMEDATIONS

Overview

In this chapter, we summarize major results and provided conclusions based on set objectives which reflects relevant literatures forming the basis of this research. We proceed to make recommendations to enhance further research to find favourable effective solutions to health complications regarding sickle cell diseases.

Summary

The dendrogram clustering of our molecular descriptors evidently shows a maximum distinct spread from the Wildtype resulting from the replacement of glutamic acid at position 6 by valine (E6V).

Also, E6K, E26K and K95E mutations showed a much less significant effect in spread whereas K82N and V23I were being considerably distinct from the wildtype, even though they all belong to one bigger cluster. The impact of the severe mutation type is at 0.015 Euclidean distance away from the wild type mutation.

Conclusions

We used graph-theoretic methods to study the effect of single point virtual mutations on the haemoglobin protein domain for mutations that are associated with sickle cell anaemia

We showed that E6V causes significantly huge impact on the domain, consistent with literature that E6V is a highly severe type of sickle cell disease. Clinically, mutation E6V is associated with death in most situations. Evidently, three mutations relative to the wildtype as presented in our work were ranked in the order of severity as K82N<V23I<E6V based on our model.

This work adds to knowledge on graph-theoretic modeling for examining the effect of single point mutations on a protein domain. Also, future work might explore what corrector molecule will cause mutation E6V to cluster along the mild or the wild type and this answers might hold a therapeutic intervention.

Recommendations

There are several diseases today that arise as a result of single point mutations that have not yet been studied. Future works may examine the effects of how a single point mutations leads to some of these other diseases.

Also, since we have shown through graph-heoretic model that we can visualize the impact of a single point mutation on an entire protein domain, suppose we can find a corrector molecule (say an amino acid or a protein) that can attach itself to the haemoglobin domain and then we recomputed descriptors of the top level graph, re-cluster our descriptors to obtain a result that makes E6V cluster closer enough to the Wildtype, then that corrector amino acid or protein molecule can be considered for treatment of sickle cell anaemia disease.

REFERENCES

Asare, E. V., Wilson, I., Benneh-Akwasi Kuma, A. A., Dei-Adomakoh, Y., Sey, F., & Olayemi, E. (2018). Burden of Sickle Cell Disease in Ghana: The Korle-Bu Experience. *Advances in Hematology*, *2018*, 1–5. https://doi.org/10.1155/2018/6161270 (Original work published)

Berg, J. M., Tymoczko, J. ., & Stryer, L. (2002). Hemoglobin Transports Oxygen Efficiently by Binding Oxygen Cooperatively. *Biochemistry. 5th Edition. New York: W H Freeman*. (Original work published)

Blanco, A., & Blanco, G. (2017). Medical Biochemistry. In *Medical Biochemistry*. https://doi.org/10.1038/199943a0 (Original work published)

Brinda, K. V., & Vishveshwara, S. (2005). A network representation of protein structures: Implications for protein stability. *Biophysical Journal*. https://doi.org/10.1529/biophysj.105.064485 (Original work published)

Forero, L. E. O. (2017). Modelling protein-protein interaction networks [PhD thesis]. *University of Oxford*. (Original work published)

Hodges, M. (2019). Approaches for studying allostery using network theory (Imperial College of Science, Technology and Medicine). Imperial College of Science, Technology and Medicine. https://doi.org/https://doi.org/10.25560/70800 (Original work published)

Jarman, N. (2017). *Multistability, synchronization, and self-organization in networks of nonlinear systems with changing graph topologies*. Retrieved from https://lra.le.ac.uk/handle/2381/40913 (Original work published)

Jensen, L. J. (2020). Network Biology: Introduction to STRING and Cytoscape. Retrieved from Youtube website: https://www.youtube.com/watch?v=lH75WJgLeoo&t=1460s (Original work published)

Kakraba, S. (2015). *A Hierarchical Graph for Nucleotide Binding Domain 2*. Retrieved from https://dc.etsu.edu/etd (Original work published)

Kakraba, S., Ayyadevara, S., Penthala, N. R., Balasubramaniam, M., Ganne, A., Liu, L., … Shmookler Reis, R. J. (2019). A Novel Microtubule-Binding Drug Attenuates and Reverses Protein Aggregation in Animal Models of Alzheimer's Disease. *Frontiers in Molecular Neuroscience*. https://doi.org/10.3389/fnmol.2019.00310 (Original work published)

Kakraba, S., & Knisley, D. (2016). A graph-theoretic model of single point mutations in the cystic fibrosis transmembrane conductance regulator. *Journal of Advances in Biotechnology*, *6*(1), 780–786. https://doi.org/10.24297/jbt.v6i1.4013 (Original work published)

Knisley, D. J., Knisley, J. R., & Herron, A. C. (2013). Graph-Theoretic Models of Mutations in the Nucleotide Binding Domain 1 of the Cystic Fibrosis Transmembrane Conductance Regulator. *Computational Biology Journal*, *2013*, 1–9. https://doi.org/10.1155/2013/938169 (Original work published)

Negre, C. F. A., Morzan, U. N., Hendrickson, H. P., Pal, R., Lisi, G. P., Loria, J. P., … Batista, V. S. (2018). Eigenvector centrality for characterization of protein allosteric pathways. *Proceedings of the National Academy of Sciences*, *115*(52), E12201–E12208. https://doi.org/10.1073/pnas.1810452115 (Original work published)

Ni, C., Sugimoto, C. R., & Jiang, J. (2011). Degree, Closeness, and Betweenness: Application of group centrality measurements to explore macro-disciplinary evolution diachronically. *Proceedings of ISSI 2011 - 13th Conference of the International Society for Scientometrics and Informetrics*. (Original work published)

Pavletich, N. P., Chambers, K. A., & Pabo, C. O. (1993). The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots. *Genes and Development*. https://doi.org/10.1101/gad.7.12b.2556 (Original work published)

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *Computational Chememistry*. (Original work published)

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., WangJonathan, T., Ramage, D., … Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303 (Original work published)

Steward, K. (2019). Essential Amino Acids: Chart, Abbreviations and Structure | Technology Networks. (Original work published)

Tajkhorshid, E. (2010). Dynamical View of Energy Coupling Mechanisms in Active Membrane Transporters. *NBCR Computational Science Seminar Series*. California. Retrieved from https://nbcr2u.wordpress.com/ (Original work published)

Tame, J., & Vallone, B. (1998). DEOXY HUMAN HEMOGLOBIN. *RCSB Protein Data Bank*, *2*. https://doi.org/10.2210/pdb1a3n/pdb (Original work published)

Weatherall, D., Akinyanju, O., Fucharoen, S., Olivieri, N., & Musgrove, P. (2006). Chapter 34. Inherited Disorders of Hemoglobin. In D. T. Jamison, J. G. Breman, A. R. Measham, G. Alleyne, M. Claeson, D. B. Evans, … P. Musgrove (Eds.), *Disease Control Priorities in Developing Countries (2nd Edition)* (pp. 663–680). World Bank Publications. https://doi.org/10.1596/978-0-8213-6179-5/Chpt-34 (Original work published)

WHO. (2006). *Joint WHO-March of Dimes Meeting on Management of Birth Defects and Haemoglobin Disorders* (Vol. 2). Geneva, Switzerland. (Original work published)

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., & Gerstein, M. (2007). The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLoS Computational Biology*, *3*(4), e59. https://doi.org/10.1371/journal.pcbi.0030059 (Original work published)

APPENDIX A

Table 5: Weighted Molecular Descriptors of Subsequence Graphs

Subsequence	**e1**	**e2**	**e3**	**e4**	**e5**	**e6**	**e7**	**e8**	**e9**	**e10**	**e11**
S1	17	43	84	15	24	5	4.94118	0.25521	18	0.0315	0.2294
S2	18	52	104	15	36	6	5.77778	0.23519	22.3333	0.02632	0.22116
S3	13	24	48	10	42	6	3.69231	0.22179	17.8462	0.03527	0.24101
S4	9	21	42	7	7	3	4.66667	0.40741	4.22222	0.08499	0.31253
S5	16	45	90	11	7	5	5.625	0.25521	17.375	0.03193	0.23723
S6	6	10	20	5	15	2	3.33333	0.58333	1.66667	0.15337	0.39865
S7	15	40	80	13	24	5	5.33333	0.26444	16	0.03463	0.24401
S8	23	59	118	19	49	9	5.13043	0.1525	50.6087	0.01453	0.17825
S9	6	11	22	5	12	2	3.66667	0.66667	1.33333	0.1629	0.39954
S10	23	62	124	19	48	8	5.3913	0.16982	43.8261	0.01584	0.18762

Source: Generated from Analysis (2020)

Table 5 annotations are such that;

e1 = number of nodes:

e2 = graph size (number of interactions/edges);

e3 = total vertex weighted degree of the graph;

e4 = edge weighted domination number;

e5 = node weighted domination base on atomic number; e6 = Diameter;

e7 = Degree mean value; e8 = Eccentricity mean value;

e9 = Betweenness mean value; e10 = Closeness mean value;

e11 = Eigenvector mean value:

Domination, Betweenness and Eccentricity measures were used in assessing the impact of specific nodes in the graphs to determine how their influences within the graph networks.

Descriptor e1, e2, e3, e4, e6, e7, e8, e9, e10, e11 are computed based our subdomain graph in appendix C.

Descriptor e5 was based on molecular descriptors d13 from Kakraba & Knisley (2016).

APPENDIX B

Table 6: Centrality Measures of Un-Weighted TLG

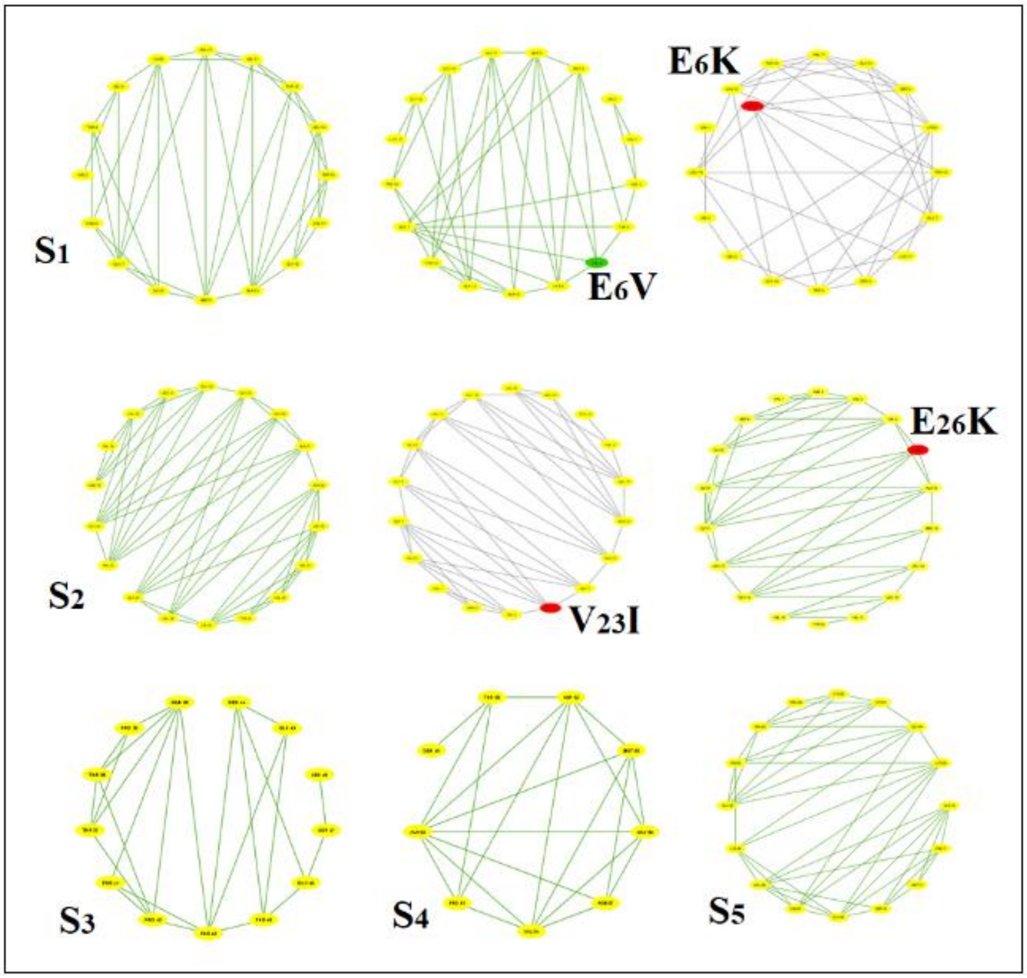
Top Level Mutation	e6	e7	e9	MaxEcc	MeanEcc	MinEcc	MeanClos	MaxEig Vec	MeanEig Vec	MinEig Vec
Wildtype	3	3.2	7.8	0.5	0.36667	0.33333	0.06040	0.48123	0.30149	0.17768
E6V	6	4.70588	23.4118	0.33333	0.21961	0.16667	0.02654	0.38370	0.21120	0.01986
E6K	6	4.82353	23.0588	0.33333	0.22353	0.16667	0.02685	0.36870	0.21709	0.00349
V23I	6	5.05882	22.4706	0.33333	0.22745	0.16667	0.02697	0.33410	0.22421	0.05131
E26K	7	5	26.4444	0.25	0.19974	0.14286	0.02424	0.39217	0.20037	0.00283
K82N	5	5.06667	16.5333	0.33333	0.25556	0.2	0.03390	0.34738	0.24224	0.05851
K95E	5	5.04348	36.1739	0.25	0.21957	0.2	0.01741	0.38911	0.17462	0.01085

Source: Generated from Analysis (2020)

MaxEcc = Maximum Eccentricity; MeanEcc = Mean Eccentricity; MinEcc = Minimum Eccentricity; MaxEigVec = Maximum Eigenvector: MeanEigVec = Mean Eigenvector: MinEigVec = Minimum Eigenvector; MinClos= Mean Closenes

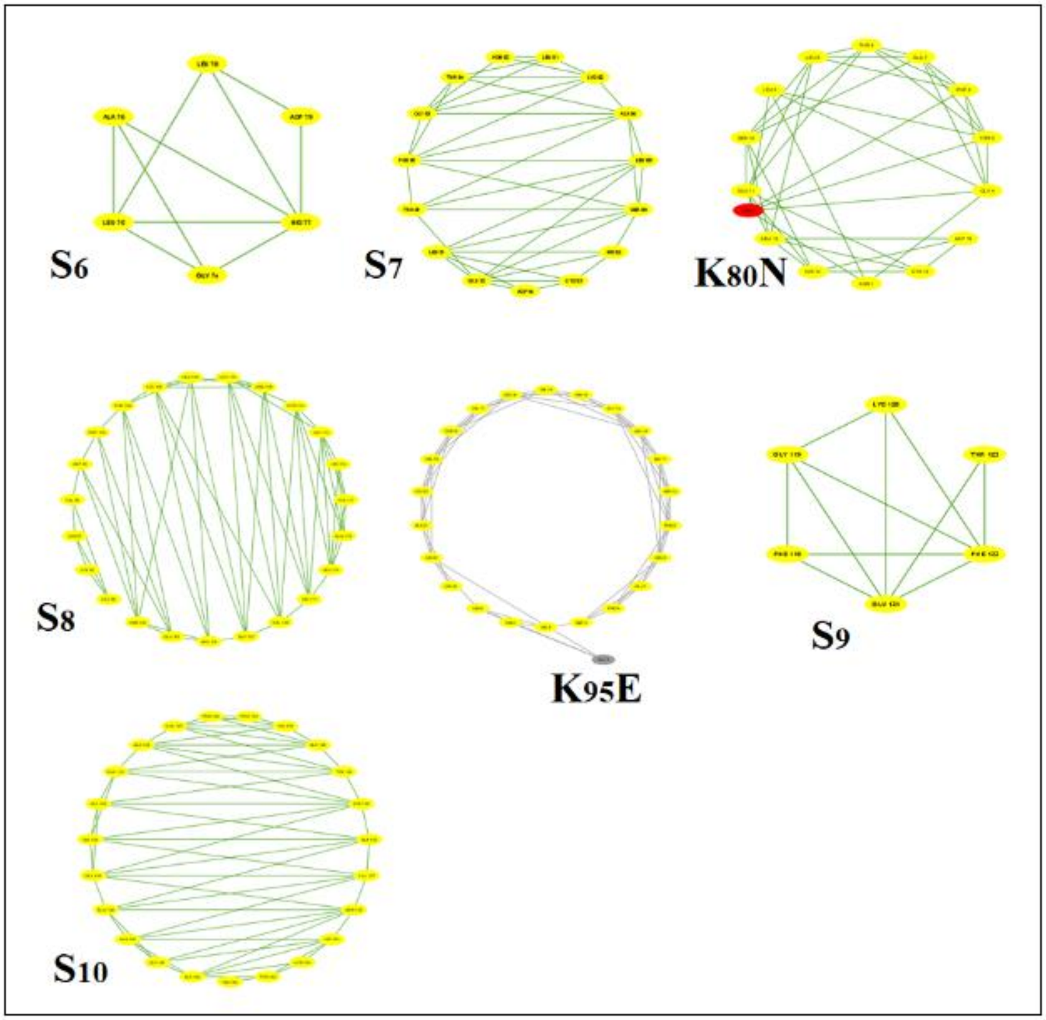
APPENDIX C

SUB-SEQUENCE AND MUTANT-DOMAIN GRAPHS



APPENDIX D

SUBDOMAIN AND MUTANT-DOMAIN GRAPHS


S6
S7
K80N
S8
K95E
S9
S10